INSTITUTO
UNIVERSITÁRIO
DE LISBOA

A Data-Driven Approach to Improve Customer Churn Prediction Based on Telecom Customer Segmentation

Tianyuan Zhang

PhD in Applied Business Management

Jury president:

Professor Elizabeth Reis, Full Professor

Jury Members:

Dr. Isabel Pedrosa, Assistant Professor

Dr. Martinha Piteira, Assistant Professor

Dr. Álvaro Rosa, Associate Professor with Habilitation

Supervisors:

Dr. Sérgio Moro, Associate Professor with Habilitation

Dr. Ricardo F. Ramos, Assistant Professor

October, 2022

iscte

BUSINESS
SCHOOL

A Data-Driven Approach to Improve Customer Churn Prediction Based on Telecom Customer Segmentation

Tianyuan Zhang

PhD in Applied Business Management

Jury president:

Professor Elizabeth Reis, Full Professor

Jury Members:

Dr. Isabel Pedrosa, Assistant Professor

Dr. Martinha Piteira, Assistant Professor

Dr. Álvaro Rosa, Associate Professor with Habilitation

Supervisors:

Dr. Sérgio Moro, Associate Professor with Habilitation

Dr. Ricardo F. Ramos, Assistant Professor

October, 2022

BUSINESS
SCHOOL

A Data-Driven Approach to Improve Customer Churn Prediction Based on Telecom Customer Segmentation

Tianyuan Zhang

PhD in Applied Business Management

Jury president:

Professor Elizabeth Reis, Full Professor

Jury Members:

Dr. Isabel Pedrosa, Assistant Professor

Dr. Martinha Piteira, Assistant Professor

Dr. Álvaro Rosa, Associate Professor with Habilitation

Supervisors:

Dr. Sérgio Moro, Associate Professor with Habilitation

Dr. Ricardo F. Ramos, Assistant Professor

October, 2022

iscte INSTITUTO UNIVERSITÁRIO DE LISBOA

A Data-Driven Approach to Improve Customer Churn Prediction Based on Telecom Customer Segmentation

Tianyua

Statement of honor

Submission of doctoral thesis

I the undersigned state on my honor that:

- The work submitted herewith is original and of my exclusive authorship and that I have indicated all the sources used.
- I give my permission for my work to be put through Safe Assign plagiarism detection tool.
- I am familiar with the ISCTE-IUL Student Disciplinary Regulations and the ISCTE-IUL Code of Academic Conduct.
- I am aware that plagiarism, self-plagiarism or copying constitutes an academic violation.

Full name Tianyuan Zhang

Course Doctor of Business Administration

Student number 75400

Email address zhangty0612@gmail.com

Personal email address 1226711820@QQ.com

Telephone number +351 910031793

ISCTE-IUL, 21/08/2022

Signed

张天源

Abstract

Numerous valuable clients can be lost to competitors in the telecommunication industry, leading to profit loss. Thus, understanding the reasons for client churn is vital for telecommunication companies. The process, combined with the massive data accumulation in the telecom industry and the increasingly mature data mining technology, motivates the development and application of a customer churn model to predict customer behavior. Therefore, the telecom company can effectively predict the churn of customers and avoid customer churn. Facing this challenge, this research aims to improve customer targeting using customer segmentation approaches based on data science. To achieve this aim, the research was divided into two stages. A literature review on customer churn prediction was conducted in the first stage. Results showed that the most widely used data mining techniques are decision tree (DT), support vector machines (SVM), and Logistic Regression (LR). Following the results, in a second stage, data were collected from three major Chinese telecom companies to create a churn prediction model to predict telecom client churn through customer segmentation using Fisher discriminant equations and LR analysis. Results showed that the telecom customer churn model constructed by regression analysis had higher prediction accuracy (93.94%) and better results.

This research contributes to academia by revealing research gaps, providing evidence on current trends, and helping to understand how to develop accurate and efficient Marketing strategies. Moreover, this study will help telecom companies efficiently predict customer churn and take measures to prevent it, thereby increasing their profits.

Keywords: Customer churn, consumer behavior, telecommunications; customer segmentation; data mining; targeted marketing

JEL：C55

Resumo

Inúmeros clientes valiosos podem ser perdidos para a concorrência no setor das telecomunicações, levando à perda de lucros. Assim, entender os motivos da rotatividade de clientes é vital para as empresas de telecomunicações. O processo combinado com o acumular massivo de dados no setor das telecomunicações e a tecnologia de mineração de dados cada vez mais madura, motiva o desenvolvimento e aplicação do modelo de rotatividade de clientes para prever o comportamento do cliente. Portanto, uma empresa de telecomunicações pode prever efetivamente a rotatividade de clientes e evita-la. Face a este desafio, este estudo visa melhorar o direcionamento de clientes usando abordagens de segmentação de clientes baseadas em ciência de dados. Para atingir este objetivo, a pesquisa foi dividida em duas etapas. Na primeira etapa, foi realizada uma revisão da literatura sobre previsão de rotatividade de clientes. Os resultados mostraram que as técnicas de mineração de dados mais utilizadas são Decision Tree (DT), Suuport Vector Machine (SVM) e Regressão Logística (LR). De acordo com os resultados, numa segunda etapa, foram recolhidos dados de três grandes empresas de telecomunicações chinesas para criar um modelo para prever a rotatividadeatravés da segmentação de clientes, usando equações discriminantes de Fisher e LR. Os resultados sugerem que o modelo construído através de LR apresentou um maior nível de previsão (93,94%) e melhores resultados.

Este estudo contribui para a academia, revelando lacunas de pesquisa, fornecendo evidências sobre as tendências atuais e ajudando a entender como desenvolver estratégias de Marketing precisas e eficientes. Além disso, este estudo ajudará as empresas de telecomunicações a prever com maior eficácia a rotatividade de clientes e tomar medidas direcionadas para evitar a sua perda, aumentando assim seus lucros.

Palavras-chave: Rotatividade de clientes; comportamento do consumidor; telecomunicações; segmentação de clientes; mineração de dados; marketing direcionado

JEL：C55

Acknowledgements

First of all, I sincerely appreciate my two supervisors: Professor Sérgio Moro and Professor Ricardo F. Ramos, for all their guidance on the thesis. The great help and support during my thesis writing encouraged me to overcome all the difficulties, improve my academic ability and keep growing in my career. Without their guidance and care, my thesis could never have been successfully completed. Spending time with them and following their professional suggestion, I have learned much knowledge and broadened my horizons.

Secondly, I sincerely thank my family and my wife, Hong Yu, who have given me fully understanding during my research. Furthermore, they have been a huge motivation and support for me to finish my studies along the way. I appreciate their unconditional love and help.

Thirdly, I would like to thank the professors and teachers in the ISCTE and Southern Medical University, where I worked with. Without their encouragement and help, I would never have had the opportunity to attend the project and finish my studies. They have given me selfless help and a lot of valuable suggestions, which helped me better adapt to the study at ISCTE and find a balance between my study, work, and life.

Table of contents

Table of contents for tables

Table of contents for figures

Glossary

Items	Meaning of expression
ANFIS	Adaptive Neuro Fuzzy Inference System
ANN	Artificial Neural Network
AUC	Area Under the Curve
BBN	Bayesian Belief Network
BPNN	Back Propagation Neural Network
BI	Business Intelligence
CA	Covering Algorithm
CCCP	Cross-Company Churn Prediction
CCPBI-TAMO	Customer Churn Prediction Business Intelligence Using Text Analytics With Metaheuristic Optimization
CCP	Customer Churn Prediction
CPIO-FS	Chaotic Pigeon-Inspired Optimization-Based Feature Selection
CRBT	Color Ring Back Tone
CRM	Customer Relationship Management
DL	Deep Learning
DM	Data Mining
DMEL	Data Mining Algorithm
DT	Decision Tree
EA	Exhaustive Algorithm
EMPC	anticipated maximum profit measure for customer churn
ERNN	Elman Recurrent Neural Network
ESN	Echo State Network
FCM	Fuzzy C-means
Fee with China Mobile	The cost generated by customers using China Mobile service
Fee with fixed line	The cost generated by customers using fixed line
FIS	Fuzzy inference systems
Fixed monthly cost	The customers’ monthly fixed fee
GA	Genetic Algorithm
GBM	Gradient Boosted Machine
GP	Genetic Programming
JRNN	Jordan Recurrent Neural Network

KA	Knowledge Acquisition
KBS	Knowledge-Based System
KMO	Kaiser–Meyer–Olkin
KNN	K-Nearest Neighbors
LA	LEM2 algorithm
LLM	The Logit Leaf Model
Local fee	The cost generated by customers using local call service
LR	Logistic Regression
LSTM	Long Short-Term Memory
ML	Machine Learning
MLP	Multilayer Perceptron
MMS	Multimedia Messaging Service
MOU	Minutes Of Usage
NCL	Negative Correlation Learning
NN	Neural Network
PCA	Principal Component Analysis Algorithm
PLS	Partial Least Square
PSO	Particle swarm optimization
Roaming fee	The cost generated by customers using the mobile phone roaming service
RDR	Ripple Down Rule
RF	Random Forest
ROC	Receiver Operating Characteristic
SAE	Stacked Autoencoder
SE	Simulated Expert
SFO	Sunflower optimization
SMS	Short Message Service
SNA	Social Network Analysis
SVM	Support Vector Machines
Total long-distance MOU	The total minutes of customers making long-distance call
Total monthly called MOU	The total minutes of customers are called monthly
Total monthly caller MOU	The total minutes of customers calling someone else
Unicom's network fee	The cost generated by customers using China Unicom intranet

	service
XGBOOST	Extreme Gradient Boosting

Chapter 1 Introduction

1.1 Introduction

With the rapid development of computer and Internet technologies, people's lives have undergone earth-shaking changes. Changes in the form of communication have prompted the telecommunications industry to flourish (Sun, 2017). In the "Big Data era" of information explosion, as one of the leading industries in the information age, the development of the telecom industry depends not only on communication technology but also on the resource optimization and configuration capabilities of enterprises and the management of enormous information and data resources becomes an enterprise. Massive data accumulation in the telecommunication (telecom) industry and the widespread application of data warehouse technology make it possible to gain insight into customer behavior characteristics and potential needs through systematic customer historical data records. It also provides prerequisites for targeted Marketing in the telecom industry (Wang et al., 2018).

Telecom operators have accumulated a large amount of customer information and consumption data during their development. These data truly and objectively reflect the behavior of consumers. Combining data mining technology with the rich data resources of the telecom industry can help telecommunications companies predict customer churn and develop more accurate, efficient, and effective marketing strategies.

Client churn is a significant problem for telecommunication companies, resulting in decreased profit (Bach et al., 2021). Moreover, this is particularly relevant since telecommunication companies operate in a saturated global market, meaning it is increasingly challenging to retain customers. Although such companies make considerable marketing investments to acquire new users, retaining a customer is usually less expensive than acquiring a new one (Kim et al., 2020). For these reasons, avoiding customer churn has become a significant concern for telecommunications companies.

Customer churn refers to the loss of a customer in favor of a competitor (Xie et al., 2009), reflecting the end of the relationship. Customer churn prediction allows one to identify the reasons for the end of the relationship and assemble a strategy that will minimize the churn rate, increasing profits. Thus, anticipating a customer’s intention to end a relationship is instrumental for telecommunication companies and is considered a competitive advantage. Therefore, it is essential to evaluate and analyze the customers' satisfaction and to conduct customer churn prediction activity for the telecom industry to make strategic decisions and relevant plans.

Previous studies have attempted to understand customer churn. For instance, (Bach et al., 2021) suggested a clustering and classification framework for churn management. (Fathian et al., 2016) proposed a new combined model based on ensemble and clustering classifiers. (Holtrop et al., 2017) aimed to anticipate customer churn using the principles of data anonymization. Although multiple studies have aimed to explain and predict customer churn, no study has tried to predict telecom client churn through discriminant analysis and LR.

Facing this identified gap in the literature, this study aims to improve customer targeting using customer segmentation approaches based on data science. To achieve this aim, this research was divided into two stages. In the first stage, a literature review on customer churn prediction was conducted, which highlighted the most widely used data mining techniques: Decision Tree (DT), SVM, and Logistic Regression (LR). After that, in a second stage, data were collected from three major Chinese telecom companies to create a churn prediction model to predict telecom client churn through customer segmentation using Fisher discriminant equations and LR analysis. Our study extends the previous work of Zhang (2018) by innovatively showing how LR analysis can be applied to build a telecom customer churn prediction model. It is expected that the results of this study will help telecommunications managers to identify the customer churn profile and create strategies to retain customers, avoiding customer churn by taking measures such as reducing monthly fixed fees. Additionally, by better predicting churn, the evolution of data mining technology makes it possible for the telecom industry to implement targeted Marketing, which motivates to effectively control the income decrease and increase the revenue of the telecommunication companies.

1.2 Thesis structure

Considering the objective of this research, this thesis comprehends five stages. In stage one (Chapter two), a background introduction is presented. Then, in stage two (Chapter three), a literature review addresses customer churn prediction, Customer Relationship Management (CRM), and the latest data mining technological progress. And study one is presented. Study one comprehends a systematic literature review that analyzed 40 articles. In stage three (Chapter four), hypotheses are proposed to investigate the factors that have impacts on Chinese telecom customer churn. In Stage four (Chapter five), factor analysis to characterize expense, call, and SMS attributes, discriminant telecom customer loss model, and LR model of telecom customer churn prediction will be conducted. Finally, a general conclusion is presented in the fifth stage (Chapter six).

1.3 Research questions

Despite the availability of various data mining techniques for customer churn prediction, the telecom industry continues to face challenges in accurately predicting customer churn. Therefore, this study aims to investigate the effectiveness of data mining techniques for customer churn prediction in the telecom industry, specifically in Chinese telecom companies. Furthermore, the study aims to examine the impact of customer segmentation on churn prediction accuracy and identify the key factors that influence customer churn in Chinese telecom companies.

This study seeks to answer the following research questions:

1. What are the most effective data mining techniques for customer churn prediction in the telecom industry?
2. How can customer segmentation using Fisher discriminant equations and logistic regression be used to predict customer churn in the telecom industry?
3. What are the key factors that impact customer churn in Chinese telecom companies?
4. How accurate and effective is the telecom customer churn model constructed by Fisher discriminant equations and logistic regression analysis in predicting customer churn?
5. What are the implications of the research for telecom companies in terms of customer targeting and marketing strategies?

1.4 Methodology

To achieve the research objectives, the following methodology will be adopted:

1. Conduct a systematic literature review to identify the most widely used data mining techniques for customer churn prediction in the telecom industry. This review will be conducted by searching electronic databases such as Scopus, Web of Science, and IEEE. The search terms will include "customer churn prediction," "telecom industry," "data mining techniques," and related keywords. The inclusion and exclusion criteria will be developed to select the most relevant studies.

2. Collect data from three major Chinese telecom companies to create a customer churn prediction model through customer segmentation using Fisher discriminant equations and logistic regression analysis.

3. Develop hypotheses to investigate the key factors that impact customer churn in Chinese telecom companies and test these hypotheses using statistical analysis.

4. Evaluate the accuracy and effectiveness of the customer churn prediction model constructed by regression analysis by comparing it to other commonly used techniques.

5. Draw conclusions and provide recommendations for telecom companies to improve customer targeting and develop effective marketing strategies based on the research findings.

Chapter 2 Background

2.1 Customer churn prevention

Customer churn is a growing issue in today's competitive and rapidly growing telecom industry. The focus of the telecom industry has shifted from acquiring new customers to retaining existing customers owing to the associated high cost (Hadden et al., 2007). The telecom industry can save the marketing cost and increase sales by retaining existing customers. Therefore, it is essential to evaluate and analyze the customers' satisfaction and to conduct customer churn prediction activity for the telecom industry to make strategic decisions and relevant plans.

Customer churn problems could be solved from two different angles. One is improving customer churn prediction models and boosting predictive performance (Verbeke et al., 2012). Another is understanding the most critical factors that drive customer churn, such as customer satisfaction. Customer churn prediction is considered a managerial problem is driven by individual choice. Therefore, many researchers mention the managerial value of customer segmentation (Hansen et al., 2013). Customer churn prediction models need to create actionable insights and have a good predictive performance by considering the two research angles.

Customer churn prediction is part of CRM since retaining and satisfying the existing customers is more profitable than attracting new customers for the following reasons:

(1) Profitable companies usually keep long-term and good relationships with their existing customers to focus on their customer needs rather than searching for new and not very profitable customers with a higher churn rate (Reinartz & Kumar, 2003);

(2) the lost customers can influence others to do the same using social media (Nitzan & Libai, 2011);

(3) long-term customers have both profit and cost advantages. On the profit dimension, long-term customers tend to buy more and can recommend people to the company using positive words. On the cost dimension, they have less service cost since a company already masters information about them and understands their customer needs (Ganesh et al., 2000);

(4) Competitive marketing actions have a more negligible effect on long-term customers (Colgate et al., 1996);

(5) Customer churn increases the demand and the cost to draw new customers and decreases the potential profits by the lost sales and opportunities. These effects lead to the fact that retaining an existing customer costs much less than drawing a new customer (Torkzadeh et al., 2006). Therefore, customer churn prediction is essential in a customer retention strategy.

Currently, CRM is valued by many companies since customer retention, which concentrates on developing and keeping long-term, loyal, and profitable customer relationships, is an essential factor for the company to win investment. Developing effective retention methods is critical for businesses, especially for telecom operators, since they lose 20% to 40% of customers annually (Orozco et al., 2015). Retaining existing customers does not have the cost of advertising, educating, or creating new accounts as attracting new customers. Consequently, compared with attracting new customers, retaining an existing customer is five times cheaper (McIlroy & Barnett, 2000). Decreasing the customer churn rate from 20% to 10% can save about £25 million annually for the mobile operator Orange (Aydin & Özer, 2005).

The value that customers provide to businesses will decline as a result of customer turnover. The company will lose its edge in the market if its customers keep defecting in droves. When a company's requirements for expansion outstrip its ability to attract new clients, existence becomes a precarious balancing act. It is well known that when running a business, it is more important to maintain the satisfaction of existing customers than to seek new ones. This may be accomplished by identifying at-risk clients and then reaching out to them with retention offers that are specifically designed to keep them as customers. Predictive algorithms that can pinpoint which consumers are most likely to defect soon are essential for this kind of strategy. Especially in the telecommunications industry, businesses are becoming more interested in the ability to estimate the likelihood that a customer would leave. A number of churn prediction models have been developed, the most of which mainly rely on data mining concepts employing machine learning and metaheuristic algorithms (Keramati & Marandi, 2015). This study aims to explore some of the most cutting-edge churn prediction techniques developed in recent years (Ahmed & Linen, 2017). The second goal is to find the consumers on the verge of leaving and estimate how long they will stay. Ahmed and Linen (2017) examines the many methods used to anticipate customer turnover to establish the factors contributing to this phenomenon. This article gives a comprehensive summary of churn prediction approaches to assist organizations in understanding customer churn, indicating that hybrid models, as opposed to separate algorithms, provide the most accurate churn prediction (Ahmed & Linen, 2017). The telecom industry can then use this knowledge to meet the requirements of high-risk customers better and reverse a churn decision (Ahmed & Linen, 2017).

Digital customer relationship management (CRM) solutions are becoming more popular as the digital infrastructure supporting them matures. Companies in the telecommunications sector are more digitalized, and this tendency is becoming more noticeable in that sector. This work examines the practicality of predicting customer turnover. In addition, the study reveals that boosting may be used to enhance churn prediction models (Lu et al., 2012). A single logistic regression model is utilized to assess the result. Boosting provides a sufficient separation of churn data, according to the findings of the studies. Therefore it is advised for churn prediction analysis (Lu et al., 2012).

Customer churn is an issue that sits towards the top of the list for larger corporations (Van et al., 2003). Due of the direct effect on revenue, companies are striving to develop methodologies for predicting future client attrition. Determine the elements that cause customers to leave in order to decrease customer turnover. Massive amounts of SyriaTel's raw data were utilized to construct and test the model in the Spark environment, where they were turned into a substantial dataset (Ahmad & Aljoumaa, 2019). The dataset SyriaTel utilized to evaluate the model includes all customer data from the prior nine months. In this model, a number of decision-making techniques were evaluated, including DT, RF and Extreme Gradient Boosting (XGBOOST). As a part of the churn prediction model, this approach was used for categorisation (Ahmad & Aljoumaa, 2019).

Customer turnover has become a major concern in the telecommunications business, making churn prediction key for retaining customers and reducing losses (Ullah et al., 2019). Two more important markers of a classification model are high prediction accuracy and easily understandable results. Many businesses in the telecommunications industry depend on being able to compete on the market. Telecommunications companies have a major challenge in dealing with customer churn as competition between businesses intensifies. To effectively reduce customer churn, businesses concentrate on keeping their current clientele rather than acquiring new ones (Jahromi et al., 2014). Prior studies focused on determining which telecom customers are more likely to transfer providers. Using the Bayesian Belief Network, a model was created to forecast the behavior of consumers who are likely to leave (Bhattacharyya & Dash, 2022). This is based on information supplied by a Turkish telecommunications firm. In the Bayesian Belief Network, the average number of minutes spent on calls, the average amount billed, the frequency with which customers are approached by different service providers, and the kind of tariff were the most significant factors in explaining customer turnover (Keramati et al., 2014).

In most industries where switching costs are high, customer attrition or churn dominates the landscape of operations (Customers want to move or change their providers for different reasons) (Edward et al., 2010). This problem is prevalent and fast developing in the telecommunications business. Companies must take proactive efforts to reduce customer churn since the industry is very competitive and the number of prepaid clients is expanding (Olle & Cai., 2014). Olle and Cai (2014) developed a mixed-learning approach for forecasting mobile communication network churn. To test the efficacy of the model, experiments were conducted using the machine learning program WEKA and a real dataset from an Asian mobile operator. It is evident, based on the data, that the new hybrid technique is more accurate than independent procedures (Olle & Cai., 2014).

2.2 Data mining current research state and application in customer churn prevention

Predicting customer churn has been a subject of data mining. Compared with traditional surveys, data mining is better at investigating customer churn (Huang et al., 2012). Traditional surveys suffer from high costs and limited access to the customer. However, data mining overcomes this kind of problem, which provides a conclusion based on the analysis of historical data. Therefore, data mining has become the most common method in customer retention to predict if a customer will churn or not and identify patterns using customers' historical data (Liu & Fan, 2014).

Predicting customer churn is difficult since customer behaviors are heterogeneous (Amin et al., 2019). In the past, companies have tended to investigate customer churn using traditional methods such as surveys. However, the data mining approach has been proven to be an efficient and better solution (Huang et al., 2012). Specifically, a customer churn prediction model could be established to understand the factors that lead to customer churn and to predict customer loss. The model could be optimized through data mining to improve its prediction accuracy (Verbeke et al., 2012). Moreover, customer segmentation is often combined with customer churn prediction for greater management effectiveness (Hansen et al., 2013).

By comparing the accuracy of telecom customer churn prediction models constructed using different data mining methods, this research can measure which data mining method is best (Ahmad et al., 2019). In addition to accuracy, there are other metrics for measuring the performance of customer churn prediction models, such as the understandability and intuitiveness of the model (Bock & Poel, 2012). Idris et al. (2019) established a telecom customer churn prediction model with good understandability and intuitiveness using the GP-AdaBoost method.

Data mining has emerged as a critical computer science topic with expanding industry importance in recent years. There's little question that data mining research will continue to grow in the following decades. The original concept of "data mining" serves as a starting point for a discussion of what individuals feel will be the field's defining difficulties in the coming years (Kriegel et al., 2007). Increasing need for technologies that aid in uncovering and analyzing information buried in large volumes of data has increased in popularity. Sources of abundant data, such as databases and warehouses, are now available. In addition to database systems, intelligent information systems, statistics, machine learning, and expert systems, data mining research encompasses a vast array of disciplines (Han et al., 2022). Large-scale datasets from the real world provide a huge theoretical and practical challenge for data mining, which has become an important and active field of research (Liu et al., 2012). Many different elements of data mining have been studied in depth in various related domains. However, the issue is so unique that these studies must be expanded to encompass the nature of the contents of real-world databases. They're obligatory. This chapter will explain the fundamentals of data mining theory and practice, as well as a few examples in the real world (Deogun et al., 1997). Rough set mining is a significant emphasis of this book. Hence a chunk of the book is dedicated to outlining the current status of rough sets in relation to real-world databases. Furthermore, this research demonstrates that the notion of rough sets provides a solid foundation for data mining applications (Deogun et al., 1997).

Although data mining and customer relationship management (CRM) have risen in relevance in recent years, few in-depth research and categorization systems examine their various aspects in detail (Rygielski et al., 2002). Data mining and customer relationship management (CRM) trends are analyzed using a bibliometric technique that evaluates the SSCI database for data mining and CRM research topics from 1989 to 2009 (Tsai, 2011). Using a bibliometric analysis technique, data mining and CRM were examined in SSCI articles from 1989 to 2009. Tsai (2011) found 1181 articles utilizing data mining and 1145 publications using

CRM. Using the following eight categories, articles about data mining and CRM were grouped. (1) publication year, (2) citation, (3) country/territories, (4) document type, (5) institute name, (6) language, (7) source title, and (8) subject area for various distribution statuses, to examine the differences and how data mining and CRM technologies have developed during this time period, as well as trends in data mining and CRM technology based on the above results. Thus, it is feasible to identify the characteristics of authors with high, moderate, and low publishing activity (Parry & Urwin, 2011). This information may also be used to evaluate scientific research trends and identify the scope of CRM and data mining research by analyzing the growth of article authors. Governments and businesses may use the aforementioned data to predict future trends and demand for data mining and CRM researchers and then design suitable training initiatives and regulations. Researchers in data mining and CRM will benefit from this study since it serves as a road map for future work, abstracts technological trends, and makes it easier to accumulate information. (Tsai, 2011).

Human actions have been influenced by knowledge from the beginning of time. There are a number of methods for data mining, but the most prevalent is to seek for patterns and trends in the massive volumes of data contained in databases (Han et al., 2022). Rather than covering all potential future directions for data mining, this article will concentrate on those that seem to have the most significant promise and potential for use in actual data mining projects (Paidi., 2012). A key component of operational research (OR) is extracting intriguing patterns from data, which is why data mining is so important. Data mining is a prevalent activity in many sectors, including credit risk assessment, marketing, fraud detection, and counterterrorism (Phua et al., 2010). All of these decision-making processes are increasingly reliant on data mining (Tufféry, 2011). However, there are still a number of challenges that need to be resolved, such as data quality concerns (Cai & Zhu, 2015). Here, the future of data mining and its significance in operational research are discussed (OR) (Baesens et al., 2009).

Due to the expansion of data warehouses that aggregate operational, customer, supplier, and market information, data has exploded (Mohanty et al., 2013). To remain competitive, you must evaluate data fast and thoroughly (Kleissner, 1998). The gap between consumers' ability to perceive and act on the information they possess has widened as data warehouse technologies have progressed. With the aid of technology and services for data mining, the gap may now be filled (Hung et al., 2010). Data mining is a complimentary technology to existing decision support technologies that provides business analysts and marketing professionals with a novel method to company analysis (Delen, 2014). Data mining allows the automatic recognition of previously unknown patterns and the automated prediction of trends and behaviors (Hormozi & Giles, 2004). An introduction of the data mining lifecycle and knowledge discovery lifecycle is followed by an evaluation of data mining challenges and demands within a business (Schmidt & Sun, 2018). This research finishes with a look at recent advancements in the workplace use of data mining tools and procedures (Kleissner, 1998).

It is impossible to exaggerate the significance of data and information to human activity. Due to the relevance of getting knowledge/information from enormous data warehouses, data mining has become a vital aspect of human life (Saroja & Appa, 2013). There has been an explosion in the number of data mining applications that may be used in all areas of human life, from business to health care to education, and these applications have improved many people lives. Amado et al. (2018) propose a semi-automated technique to spot the important trends in this field. According to the conclusions of the research, Big Data, Marketing, the geographical location of the authors' affiliation (country/continent), Products, and Sectors are the most relevant terms and topics associated with the five dimensions (Faisal et al., 2021). Between 2010 and 2015, a total of 1560 articles were analyzed. According to the findings, there is a dichotomous character to research, with Big Data publications not clearly linking innovative methodology with Marketing benefits. Also identified were few publications by writers from other continents (Subotnik et al., 2011). As a consequence, research in Big Data applications to Marketing is still in its infancy, necessitating that Big Data thrive in the Marketing sector by making more direct business-oriented initiatives (Amado et al., 2018).

The creation of data in the construction business is expanding rapidly. Data mining (DM) has developed as an effective method for knowledge discovery in the construction sector (Cios & Kurgan, 2005). An in-depth analysis of DM's use in the construction sector is still scarce, despite the industry's rapid expansion in this area. The goal in writing this study is to give a complete overview, focused on construction-related topics, of the DM application literature published between 2001 and 2019 (Yan et al., 2020). DM apps are becoming more popular in the construction business, particularly since 2016, with a large number coming from China. Data sources, DM functions, and widely used DM approaches in the construction sector will all be thoroughly examined and described in this paper." The core study foci in nine major application areas are energy, safety, building occupancy, and occupant behavior. From the results, four significant issues and future research paths have been identified. Researchers get a thorough understanding of the present state of DM applications and the heuristic implications for future study (Yan et al., 2020).

Data mining has become a term since it is so trendy right now. Using data mining methods, it is feasible to extract useful information from massive datasets (Shaw, 2001). Misuse of the tool might lead to inaccurate and pointless data being gathered. To design a successful data mining project, you need to know what you're doing and what you're looking for. Integrate and clean or alter the data sources, mine the data, go over and prune the mining findings, and then report on the final results are all processes in a data mining project (Thuraisingham, 2000).

Data mining's success in high-profile disciplines like e-commerce, marketing, and retail has prompted its expansion into various fields and businesses (Simon & Shaffer, 2001). The healthcare industry is one of those that is just now making inroads. It's still an "information-rich" yet "information-poor" environment in the healthcare sector. Information is abundantly available to healthcare systems. On the other hand, there is an absence of effective data analysis tools for discovering hidden patterns (Aljumah et al., 2013). Using data mining techniques, Soni et al. (2011) want to provide an overview of current strategies for uncovering new information in databases, with a particular emphasis on the prediction of cardiac disease. The results demonstrate that the Decision Tree outperforms Bayesian classification, which is occasionally as precise as the Decision Tree. However, other predicting approaches, such as KNN, Neural Networks, and Classification based on clustering, perform poorly (Fan et al., 2011). Genetic algorithms may be used to minimize data size to find the ideal subset of attributes that can be used to predict heart disease more accurately than a Decision Tree or Bayesian Classification (Soni et al., 2011).

Measuring over time is a standard procedure in almost all scientific fields (Honaker & King, 2010). Time-series data mining aims to discover as much useful information as possible by analyzing the data's structure. Even if humans have a built-in ability to do these jobs, computers nevertheless face a difficult challenge. In this post, Esling and Agon (2012) will take a look at a variety of time-series data mining methods. It begins with an overview of the tasks most drawn scholars' attention. This is because, in most situations, time-series tasks need similar components to be implemented. Thus this research splits the literature into sections based on these common elements (Esling & Agon, 2012). Each aspect's results have been divided into different categories after a comprehensive examination of the relevant literature (Ghose & Ipeirotis, 2010). The formalization of four robustness classes and the subsequent classification of any distance would be the next logical step. The study concludes by presenting several potential future research directions and trends (Esling & Agon, 2012).

As computer databases are increasingly used to store and retrieve data, a new and emerging technique called data mining has emerged to extract new, implicit, and practical knowledge from massive datasets (Lakshmi & Raghunandhan, 2011). Knowledge Discovery in Datasets (KDD) permits data exploration, analysis, and visualization of large databases without a predefined hypothesis, which is why it is also known as KDD. Modeling is used to understand how data mining works and create predictions. This research gives an overview of DM technology, including its definition, motivation, methodology, and architecture, as well as the types of data mined, features, and classifications used in data mining (Lakshmi & Raghunandhan, 2011).

As a result of its widespread usage in high-profile disciplines like e-commerce, marketing, and retail, knowledge discovery in databases (KDD) has grown in popularity across a wide range of businesses. The areas of medicine and public health are two of the most recent to uncover the benefits of big data analytics (Canlas, 2009).

An overview of current KDD approaches for healthcare, and public health is presented in this research study. Data mining and healthcare, in general, are major topics of discussion as well (Canlas, 2009).

According to the findings, data mining is increasingly being used for health care policy purposes, disease outbreak identification, unnecessary hospital fatalities, and fraud detection (Canlas, 2009).

Given the current fascination with AI and ML, predictive maintenance is one of the most discussed novel ideas in event prediction. It has a wide variety of helpful resources for making accurate predictions (Zeng, 2015). Lack of service is a prime example of a significant challenge for the mobile communication industry. The current methods for addressing service unavailability are purely reactive. However, the time needed to resolve the issue is normally very long. The user experience and revenue loss may be greatly reduced if we could identify the actions that cause disturbances in our network to foresee them by utilizing AI algorithms. One of the primary AI methods for predicting events is data mining. This research aims to show how powerful data mining approaches may be used to solve this long-standing problem (Kamel et al., 2021). For Rare Association Rule Mining, the study suggests a novel method that uses concurrency for event prediction (RAR). The proposed approach predicts potential outcomes and recommends a course of action to avoid them (Kamel et al., 2021).

An important part of data warehousing and database management is called "data mining." This may be accomplished using a wide range of methods and tools. It is possible to use data mining in various fields, from business and medicine to engineering. Data mining for student profiles and categorization is the topic of this article. As one of the most prevalent ways to detect co-relations among a group of things known as "mining associations," the Apriori algorithm is used in the student profile (Parack et al., 2012).

Additionally, K-means clusters may be utilized to divide many students into smaller groups. Data mining may be quite helpful in the academic world when identifying relevant information about students based on their educational records. Parack et al. (2012) use the Apriori algorithm to analyze the academic records of different students and attempt to identify correlations between various criteria, including test scores, term grades, attendance, and practical exams, to create a profile of each student. To further organize the pupils, this research uses K-means clustering on the same data set. Algorithms implemented in educational systems can be used to accurately profile students (Parack et al., 2012).

From 1989 through 2009, the SSCI database was used to locate the subject heading "data mining," which was then analyzed using a bibliometric technique. Using bibliometric analysis, data mining was detected in 1181 publications published in SSCI journals between 1989 and 2009 (Hung & Zhang, 2012). A technique known as the K-S test is also used in this work to ensure that the results obey Lotka's rule (Huang & Yang, 2012). In addition, the study examines the historical literature to discover data mining technology diffusions. Abstracting technological trends and projections, and facilitating knowledge accumulation, the document offers a roadmap for future study and saves time for data mining experts. Higher-quality publications are more likely to have a common phenomenon known as "Success Breeds Success" (Tsai, 2012).

Data mining techniques are discussed by Petre (2012). There are several ways to mine data for meaningful information. Data mining methods are now routinely incorporated into everyday tasks. Businesses may save expenses and increase efficiency by using data mining tools, and constantly inundating customers with personalized advertising (Petre, 2012).

The paradigm of cloud computing requires the usage of data mining apps and techniques. Cloud computing may be used to undertake data mining techniques that allow customers to get relevant information from a virtual data warehouse while minimizing storage and infrastructure costs (Petre, 2012).

Only a small number of research have looked at the elements that influence the data sources of online search queries. Google Trends and Baidu Index were compared in a study to evaluate this problem (Morris et al., 2010). These two business use data from Google and Baidu, two of the world's major search engines. To begin with, Vaughan and Chen (2015) examined the two systems' features and functionalities by reviewing documentation and doing comprehensive testing. In order to test this, Vaughan and Chen (2015) conducted empirical research that gathered data from both sources. Using data from both sources, Vaughan and Chen (2015) could forecast the quality of Chinese colleges and firms. The search volume statistics from both services were strongly associated despite the variations in technology, such as different ways of language processing. Combining the two data sources did not increase the prediction capacity. In terms of data availability, however, there was a massive disparity between the two. Compared to Google Trends, Baidu Index delivered a more significant amount of search traffic. According to the findings, Google Trends is disadvantaged because of the country's smaller population. This discovery has implications far beyond China. Google's user bases in many nations may be lower than those in China, which might lead to the same dilemma as that of China (Vaughan & Chen, 2015).

As a result of its widespread use in high-profile domains like e-commerce, data mining has found a home in a variety of other sectors (Kumar et al., 2020). In the medical area, there is still a dearth of knowledge despite the quantity of information. Medical systems include an abundance of data (Lalazaryan & Zare, 2014). A paucity of efficient data analysis tools stops us from uncovering the underlying relationships and patterns in the data. The term "heart disease" embraces a broad variety of heart-related disorders (Mohapatra et al., 2021). These medical terminologies describe unanticipated conditions that affect the heart and its components. Using DM methods like as clustering in medical sector, it is possible to analyze heart-related problems (Karimi et al., 2015). A key challenge in data mining is categorizing the collected data (Banu & Gomathy, 2013). A classifier provides a short and unambiguous explanation for each class that can be used to categorize subsequent entries (Wood & Salzberg, 2014). Decision trees are used to produce class models by various prominent classifiers. The decision tree uses the C4.5 algorithm as its training method to display the severity of a heart attack (Banu & Gomathy, 2013).

For analyzing enormous amounts of information, data mining has emerged as a helpful tool (Devi, 2014). Most academics rely on scholarly publications and patent information to uncover and follow emerging technological trends (Foote et al., 2010). It is common for the Data Mining technology to employ a data integration approach to create a Data warehouse to aggregate all data into a single location and then run an algorithm against that data to extract relevant Module Prediction and knowledge assessment. However, it has not been shown that a single data-mining approach is suitable for every area and data collection (Malik et al., 2012).

Because system changes might impact the overall system performance, data mining methods used in such a complex context must be very dynamic (Wu et al., 2013). The necessity to mine data from dispersed sources led to distributed data mining. Many algorithmic techniques are available in the area of Distributed Data Mining (DDM) to deal with these issues in analyzing distributed data in a fundamentally distributed way that pays particular attention to the resource limits of the data set. This article offers an overview of Distributed Data Mining techniques, methodologies, and trends to uncover information from distributed data efficiently and effectively (Raghupathi & Raghupathi, 2014).

Kanevski (2004) will look at several cutting-edge methods for analyzing spatial environmental data. Sequential simulations may be used to measure uncertainty and regional variation (Matsen et al., 2010). Using a case study from the tragedy at Chernobyl, the usefulness of the recommended technique is shown. This research demonstrates that probability mapping may be used successfully in decision-making by combining ML data-driven and geostatistical model-based techniques (Kanevski, 2004).

There is an urgent need for sophisticated technologies that can extract hidden but relevant information about building performance improvement from massive data sets because of the continually rising and enormous body of data in the building industry (Dash et al., 2019). Recent proposals for relevant knowledge discovery have focused on data mining as an emerging discipline of computer science. This study aims to highlight recent achievements in building data mining by providing an overview of the research that focuses on both predictive and descriptive tasks. Based on this overview, this research will investigate some of the most significant issues facing science now and in the future (Yu et al., 2016).

In recent years, the application of DM methods has increased in the domains of chemistry, materials science, and engineering. This article will cover a wide range of subjects, including process optimization, design, and evaluation in chemistry, materials science, and engineering (Harding et al., 2006). Although the research and application goals are diverse, there are still many essential similar elements in their data mining. Data-mining philosophy is then applied to processes' design and optimization goals in scientific and industrial ones (Li et al., 2019).

Many companies have systematically recorded large amounts of data on their operations, goods, and customers during the last decade. At the same time, researchers and engineers in many areas have been collecting ever-larger quantities of experimental data, such as the terabytes of MRI data used to study human brain activity. Data mining is concerned with figuring out the best way to use this previously collected information in order to find patterns and make better judgments (Romero & Ventura, 2013).

In response to the confluence of multiple recent developments, there has been an increase in interest in data mining, which utilizes previous data to uncover patterns and better future judgments (Moro et al., 2019). In the discipline of data mining, sometimes referred to as "knowledge discovery from databases," "advanced data analysis," and "machine learning," several practical applications have already been established (Shanahan, 2012). These include medical outcome analysis, credit card fraud detection, customer purchase behavior prediction, personal interest prediction, and manufacturing process optimization. Several exciting scientific problems have arisen as a result, including how computers may learn from their prior experiences (Lee & Ma, 2012).

Delhi's rapid growth and urbanization have contributed to a rise in air pollution over the last several years. As a result, academics have begun looking into the topic. Analyzing and forecasting future changes in Delhi's air pollution via data mining has been the method of choice (Khanna & Sharma, 2020). Linear regression and multilayer perceptron are two of the data mining methods used. There have been trends in several air pollutants, such as Sulfur dioxide (SO2), nitrogen dioxide (NO2). This research has seen a 45.9 percent rise in PM 10 concentrations over the next several years using the above-mentioned methods (Taneja et al., 2016). CO and NO2 levels may rise as the number of two-wheeled vehicles on the road grows. Using non-sulfur gasoline and strict pollution control methods may help reduce other pollutants like SO 2 (Taneja et al., 2016).

E-learning has become an increasingly significant part of the teaching and learning process, and new procedures are needed to assess its success. E-learning is the background for this review, focusing on data mining research in the assessment context, which is seen as a latent issue in this environment. An educational perspective on EDM research into the teaching and learning process is what this research wants to convey in this study (Douglas et al., 2021). To get a more comprehensive evaluation, Using the search terms "data mining" and "education," Rodrigues et al. (2018) obtained 525 hits from the bibliographic database. Non-enhancing the teaching and learning process-focused articles were eliminated, leaving 72 articles in the final set of findings. Examining viewpoints on the evolution of teaching and learning activities allowed for the generation of ideas, the identification of trends, and the identification of future study routes (Juhaňák et al., 2019).

In today's world, KM and DM are becoming essential. Yet, there are still few in-depth studies and classification methods to explain the features of each. The following eight criteria were used to classify KM and data management papers, according to the findings of this study: publication year, citations from other sources (such as academic journals) (Pagani et al., 2015). In addition, through a comparison of the author's ascent to prominence, these results also assist to gauge the level of increase in DM and KM (Helm et al., 2014). Governments and corporations may make accurate predictions about future trends and requirements in DM research based on the facts offered here (Dwivedi et al., 2020). Future study may concentrate on important topics as a result of this analysis, which serves as a road map for future work, abstracts information on technology trends, and makes it simpler to acquire fresh information (Oztemel & Gursev, 2020). Higher-quality publications are more likely to have a common phenomenon known as "Success Breeds Success" (Tsai, 2013).

Educational data mining is a discipline currently undergoing development. It involves the creation of tools for analyzing different types of data gathered from the educational sector (Baker et al., 2016). Data mining plays an important role in education, especially in an online learning environment where behavior is being monitored. As a result of data mining's ability to analyze and uncover the hidden information within the data itself, a task that would be very complex and time-consuming if carried out manually. When it comes to data mining in educational research, this review aims to look at how it has been done before and how it's been done recently and see how it may be used in the future. The existing study's shortcomings are analyzed, and new research plans are suggested (Salloum et al., 2020).

An expanding field of study, data mining aims to discover intriguing patterns and laws in datasets. Large amounts of ordinary business data that may be inferred at a high level, such as consumer purchasing habits, supermarket shelving criteria, and stock trends, might be useful to organizations (Shekhar et al., 2011). Several different methods have suggested data mining of association rules. Sequential algorithms have been the focus of most of the study. Parallel methods for association rule data mining are presented and studied in this work on the parallelism, synchronization, and data location difficulties. Additional improvements for sequential and parallel algorithms are also shown. The suggested improvements have been demonstrated in experiments to enhance performance significantly. A good speed-up has been obtained for the parallel approach. Still this research believes that other similar I/O techniques are needed to increase performance (Zaki et al., 1996).

This research describes a comprehensive literature analysis and categorization method for data mining in academic libraries. Forty-one contributions from 1998 to 2014 were selected and assessed for their direct relevance in achieving this objective (Subhash & Cudney, 2018). There were four categories for the articles: services, quality, collection, and user behavior. These categories were determined via the four most important data mining operations: clustering, association, categorization, and regression (Guzman et al., 2015). The majority of research on website and online service usability and collection development has focused on collecting and using behavioral evaluations, according to the findings (Al-Debei et al., 2015). Classification and regression models are the most often utilized library data mining capabilities (Guzman et al., 2015).

Scientific and academic research now relies on the efficient processing of large datasets enabled by modern information technology (Fahad et al., 2014). It includes anything from weather forecasting and genetics to complicated physics simulations and biological and environmental studies. "Big Data" refers to data streams with a greater rate of change and a more comprehensive range of characteristics. When gathering data and conducting short, basic queries, the infrastructure needed to enable Big Data acquisition must have low, predictable latency. Allowing for huge transaction volumes and dynamic data structures in a distributed context. There is a lot more involved in processing data than merely obtaining and recognizing it, comprehending it, and documenting it (Cai et al., 2016). For large-scale analysis to be successful, all of this must be done automatically. There must be an ability to represent data structure and semantics discrepancies so that computers can comprehend and resolve "robotically." There is still much work to be done before automated error-free difference resolution can be achieved. Based on current data mining research, this study presents a paradigm for using big data in data mining (Sowmya & Suneetha, 2017).

Modern process monitoring methods aid chemical plant operators and engineers in interpreting current data trends by analyzing vast historical databases of sensor readings from sensors in the past (Udugama et al., 2020). In contrast, many of the greatest techniques for monitoring and training requirements that information be grouped into groups. In reality, such an organization does not exist, and the amount of time it takes to produce classified training data impedes to the employment of sophisticated process monitoring techniques. An engineer's expertise in the process is not required to uncover fault states in historical databases using data mining, or knowledge discovery approaches borrowed from computer science literature. Clustering and feature extraction strategies in industrial chemical process data are evaluated in this research (Wang et al., 2020). When comparing the efficacy of various dimensionality reduction and data clustering algorithms, supervised clustering metrics use the proper labels in the data to compare outcomes (Thomas et al., 2018).

Thanks to a new machine learning technique, total burnt areas for specific wildfire outbreaks may now be accurately predicted. It is used to analyze data from the Montesinho Natural Park in Portugal, which has a long history of studying forest fires (Jain et al., 2020). The model is clear and does not have any hidden layers or regressions. This enhances its ability to mine more complex data sets. Traditional machine learning methods cannot match this dataset's top burned-area prediction accuracy. Predicting burnt areas in a two-stage process offers valuable information on the effect of the input factors on the indicated burned areas. It is possible to mine each total burnt area incident's data more thoroughly when the goal functions MAE and RMSE are optimized separately (Cerna et al., 2020). A better knowledge of the factors that drive each fire occurrence might positively impact agriculture, ecology, environment, and forestry. This level of precision and understanding in making predictions instills trust in the process through which they are made. It offers the information needed to respond effectively and prevent further damage in a particular burn occurrence. By helping to prevent specific kinds of burn accidents from reoccurring or spreading, such well-informed actions should have both short- and long-term positive effects (Wood, 2021).

Data mining is one of the most rapidly expanding fields in the 21st century (Aldowah et al., 2019). Because so much data is being created, this research must find the most intriguing patterns and trends (Mäntylä et al., 2018). Various methods and algorithms have been developed and used over the previous several decades to uncover such tendencies. Using data mining techniques, Nazir et al. (2019) aim to extract the essence from massive text corpora. As a means of understanding current trends, Nazir et al. (2019) looked at data from 2014 to 2018. A new dataset, including the abstracts and information of 5,843 publications, has been acquired for this purpose. ScienceDirect indexes all of the journal articles, including data mining results. Methodologies like noun phrases and TF-IDF mining are used to uncover the development of hot data mining trends through time using various techniques. The year-by-year and overall progression of these hot trends is shown in the form of a word cloud for easier visualization. Data mining stands on the shoulders of giants, machine learning algorithms (Chu, 2022).

For the time being, dashboards are the most popular method of monitoring corporate performance. Key Performance Indicators (KPIs) serve a critical function in swiftly giving accurate information by evaluating current performance against a goal necessary to achieve corporate objectives in a dashboard (Badawy et al., 2016). The problem is that KPIs aren't well recognized, and it might be difficult to choose an acceptable KPI for each company goal. When it comes to trend predicting and data correlation visualization, Data Mining methods are often used. Data Mining approaches may be driven by integrating these two factors to generate particular KPIs for business goals in a semi-automated manner. The key advantage of the method is that firms may examine their behavior using current data without having to depend on existing KPI lists or test KPIs throughout a cycle. KPIs may be identified by applying the proposed methodology to the disciplines of MOOCs (Peral et al., 2017).

Educational Data Mining (EDM) is the tool for data exploration in academic fields (Calders & Pechenizkiy, 2012). EDM employs computational methods to analyse scholastic records in order to investigate educational issues (Romero & Ventura, 2013). Using the most recent and significant works in this topic, this report provides a complete review of the current literature. Academic performance and institutional efficiency may be improved by developing models based on educational data analysis methodologies, according to this research (Anoopkumar & Rahman, 2016). It gathers and relegates relevant research to educators and professional organisations, and determines the most significant work. Anoopkumar and Rahman (2016) look for research that provides well-supported advice for educating and energizing the institution's most impuissant pupils. The findings of these investigations provide light on approaches for improving the educational process, predicting student performance, and comparing the accuracy of data mining algorithms (Papamitsiou & Economides, 2014).

For the study of geographical data, Spatial Data Mining (SDM) technology has emerged as a new discipline (Perumal et al., 2015). A Geographic Information System (GIS) utilizes data obtained from diverse sources and stored in a variety of formats to characterize geographic attributes in terms of latitude and longitudinal coordinates (Arbenina, 2021). The quantity of data created by geodatabases from satellite pictures, which contain orbital characteristics, and other sources, such as water bodies, forest covers, soil quality monitoring, etc., is increasing rapidly. There has been an increase in the use of GIS in traffic research, tourism, health care management, and the conservation of biodiversity (Cetin & Sevik, 2016). The importance of using computer tools to retrieve information from geodatabases has increased. This brief review touches on GIS data formats, representation models, data sources, data mining techniques, and SDM tools. On the basis of research into different literatures, this study discusses the concerns of GIS data and proposes an architecture to solve the challenges of GIS data (Perumal et al., 2015).

For sensitive information to be exchanged in the context of data analysis, validation, and publication, privacy must be maintained. Internet phishing poses a serious danger to the broad dissemination of sensitive information via the internet. Data sharing is frequently rejected or misinformation is shared because of the skepticism and disagreements among different information providers about how to ensure that their data is protected from unauthorized exposure (Price & Cohen, 2019). Through rigorous categorization of a list of published literature, this article presents a panoramic picture of fresh viewpoint and methodical interpretation. Existing techniques to data mining that protect privacy, as well as their benefits and drawbacks, are examined in depth (Tufféry, 2011). Due diligence shows gaps and shortcomings in current research, as well as historical progress and current research problems. For the security and maintenance of one's private information, more important improvements must be made (Aldeen et al., 2015).

Data mining has been around for quite some time. Large amounts of data with qualities like speed, volume, and diversity are classified as "Big Data." (Wu et al., 2013). As the amount of data generated by businesses continues to expand at an exponential rate, big data mining is becoming more important. The term "big data mining" refers to the process of sifting through enormous amounts of information in order to derive actionable insights (Sriramoju, 2017).

As a result, businesses may use Big Data mining to make informed choices. But security is a major consideration as well. Big Data mining offers both potential and risks, according to the research cited. There are two crucial findings that emerged from an examination of the relevant literature (Tien, 2013). Starting with the premise that "big data mining gives additional chances for corporate development and change," Second, "Big data mining may have ramifications for security." In experiments, UCI machine learning repository benchmark datasets are used. Using Aproiri, modified Association rule mining is performed (Sriramoju, 2017). Differential privacy is a data security strategy that requires computations to be insensitive to data alterations of any database entry (Zhang et al., 2011). This method protects the mapper against both internal and external attacks. The creation of a proof-of-concept application prototype (Wang & Sawhney, 2014). Two sets of tests are run in the presence of an attacker and in the absence of an attacker (Addetia et al., 2020). Even in the face of an adversary, the proposed application may accomplish the desired results (Koscher et al., 2010).

Crime analysis is a rigorous strategy to finding and evaluating patterns and trends in criminal activity. ' As the use of computers in law enforcement grows, analysts of criminal data will be able to assist them in solving crimes more quickly. As a result, Chauhan and Sehgal (2017) may discover previously undiscovered, relevant information from an unstructured data set. The term "predictive policing" refers to the use of analytical and predictive approaches to help detect criminals. It is getting more difficult to evaluate large amounts of crime data manually, and today's criminals are growing more digitally savvy, therefore it is necessary to employ advanced technology in order to keep the police on top of them (David & Suruliandi, 2017).

Baek and Doleck (2021) did a literature review of empirical research published in both areas to compare and contrast two closely related but distinct topics, Educational Data Mining (EDM) and Learning Analytics (LA). 492 LA articles and 194 EDM pieces from 2015–2019 were combined. Data analytic methods, common terms, theories, and definitions were used to examine the similarities and contrasts between the two areas of study. Both sectors' investigations lacked a defined theoretical foundation, as this investigation observed. In these domains, ideas of self-regulated learning are most often applied (Greene et al., 2011). Through keyword research, this investigation discovered a strong connection between the two sectors, with "learning analytics" being the most often used term in EDM and vice versa in LA. EDM studies, on the other hand, tended to identify terms pertaining to technical procedures, while LA studies focused on social issues. Finally, various writers characterize the distinction between the two disciplines in different ways. Some demarcate the distinctions, while others treat them as if they were interchangeable (Kendig, 2016).

Rather of producing precise findings for use in the future, those working in the educational area choose to focus on prediction. A frequent examination of educational databases is required in order to monitor changes in curricular trends (Chatti et al., 2012). Predictive data mining models based on classification-based algorithms are used in this work to identify and show pupils who are slow learners. Data from a high school is filtered using WEKA, an open source tool, to find the variables of interest. WEKA is designed to evaluate and apply many classification algorithms to the dataset of student academic records, such as Multilayer Perception, Nave Bayes, SMO, J48, and REPTree (Kaur et al., 2015). As a consequence, data are created based on all five classifiers, and a comparison of their accuracy is performed to determine which classifier is the best overall. All five classifiers are demonstrated in a knowledge flow model in this study (Kaur et al., 2015).

Research centers and universities all have a role in shaping long-term trends in the refractory industry. Due to its complexity, it is impossible to predict who the most important players are and what the future trends will be in this field of study (Huang et al., 2015). In this research, a detailed map of international cooperation amongst the main nations in the refractory industry was created (Moreira et al., 2017). These nations all have one thing in common: a strong sense of teamwork, which enables them to deal with complicated issues. The research of a specific phrase during a period deemed to reflect a lack of technical advancement in the last 10 years yielded a wealth of information about current technology trends. In addition, a significant rise in the use of computer simulations to aid in the advancement of refractory technology was seen (Kato, 2010).

Servers are overflowing with data as the Internet becomes more global. The amount of data generated during the past two years is equal to the total amount generated throughout the previous decade. The rapid growth of data is owing to the widespread availability of Internet of Things-based devices (Uddin & Rahman, 2012). This data has become a useful tool for forecasting the future. Analysts are anticipating the future trend based on their own domain's data because of the wide variety of computer devices being used. Data analysis technologies has gotten sluggish over time. The fundamental reason for this is because data is being generated at a pace that is outpacing the technology available for its access. There are a variety of ways to mine for meaningful data. Amin and Garg (2019) examine in great depth how different data mining algorithms utilize and interpret data. Artificial Neural Networks, Linear Discriminant Analysis Methods, Nave Bayes, and Support Vector Machines are among the algorithms examined. Big data sets are utilized as input for these algorithms. This study focuses mostly on how current data algorithms interact with large datasets. Twitter comments have been studied as part of the investigation (Salloum et al., 2017).

Educational data mining has been the subject of a growing amount of theoretical and practical study in recent years. To improve the teaching-learning process, the discipline of "learning analytics" utilizes approaches, strategies, and algorithms to assist users in discovering and extracting patterns from previously collected educational data (Calvet & Juan, 2015). Despite this, several criteria for the application of new technology in teaching and learning processes remain unaddressed by these analyses. A literature study revealed no indication of a reexamination of the application of learning analytics in technical education (Williams et al., 2012). According to this paper's conclusions, researchers have a clearer understanding of how far the discipline has progressed and what remains to be done. In order to do this, a systematic mapping study focusing on the classification of publications by kind of research and type of contribution has been conducted (Neto et al., 2011). The major emphasis of case study research is on software and computer science engineering. Learning analytics may also be applied in other educational contexts, although this study focuses on its application in engineering education (Buenaño et al., 2019).

Chemical exposure risk assessments need an understanding of environmental pollutants' transformation products (TPs). It is, however, very difficult to recognize TPs in complicated environments. Advanced data mining methods have dramatically sped the discovery and identification of TPs (Escher & Fenner, 2011). High-resolution mass spectrometry (HRMS) data gathering techniques and methods for enriching TPs from diverse sample matrices are discussed in this paper. TPs are also considered in terms of knowledge gaps and potential future developments (Li et al., 2021).

Traditional data mining approaches, such as statistical calculations, machine learning, artificial intelligence, and database technologies, lack a conceptual or semantic understanding of the data. This implies that the relevance and implications of a user are not considered (Dangare & Apte, 2012). In the recent decade, semantic data mining approaches have been proposed that use ontologies as background knowledge to enable and improve data mining performance (Ristoski & Paulheim, 2016). The unique technique of arranging the surveyed articles and the critical analysis and summary of the surveyed articles are among the most significant contributions of this literature review, which intends to enlighten researchers in this expanding area about their creative methodologies and approaches (Adadi & Berrada, 2018). An extensive review of domain ontologies utilized in preprocessing, modeling, and postprocessing activities in semantic data mining is presented in this survey study. Using a framework for describing data resources, this research looked at the function of semantic data mining in data science and the procedures and techniques that go into it (Padhy et al., 2012).

Using standard data mining process models might be beneficial for data mining project managers (Mariscal et al., 2010). Utilizing cross-industry models such as CRISP-DM for data mining, is less costly and requires less time. Additionally, standard models facilitate the sharing and reuse of best practices and limit the quantity of new knowledge that must be taught (Mariscal et al., 2010). Concept mining is a technology that leverages unstructured text data to generate novel and meaningful insights (Gandomi & Haider, 2015). Using a design science technique, the CRISP-IM was conceived and created. If you're looking to find patterns in academic literature, patents, or any other textual collection, you may use the customized CRISP-IM to assist the process. The CRISP-post-implementation IM's assessment will be studied in the future (Ayele, 2020).

The use of big data in academic research has become a common practice. An examination of over 36,000 academic papers published between 2012 and 2017 across all academic fields using topic modeling and word cooccurrence analysis is presented in this article (Boyd & Crawford, 2012). Data storage, collecting, and analysis were among the subjects that emerged from the research, which was published mostly in computational domains. In reality, these subjects have become more common in recent years. According to these findings, the basic areas of big data research have matured, and there are exciting new research directions using big data in the social sciences, health, and medicine. Using the information in this article, scientists and policymakers may better understand how big data is being used across many academic fields (Mohammadi & Karami, 2022).

Some strategies are needed in order to forecast or make a choice based on the information gathered in a research project. Large data sets need data transformation and preparation using data mining methods once they have been collected. There are now a variety of open-source and commercial data mining tools accessible today (Kim & Ko, 2012). A broad variety of software solutions are included in this collection, ranging from easy-to-use data mining suites to enterprise data warehouses with built-in data mining capabilities to early research prototypes for novel methodologies that are still under development This kind of software is essential for researchers who want to analyze their data. There is a slew of tools available, like WEKA, orange, Rapid Miner, and Tanagra, to name just a few (Patel & Desai, 2015).

Multimedia mining is a burgeoning field of study because to the Internet's abundance of multimedia data. It is possible to mine data via the use of multimedia. To find trends and make predictions, data is segmented using algorithms. Data mining is still a difficult undertaking, despite its numerous triumphs (Mammeri et al., 2015). In the past, the findings of multimedia mining were typically disappointing. Searches for similarities, identification of relationships, entity resolution, and categorization may all be performed using multimedia data mining. There have been new methods created as mining processes have progressed. Recent research on multimedia data mining has included deep learning techniques (Wlodarczak et al., 2015).

Because of the link to health effects, air quality prediction is a hot issue in study. The forecast informs the local populace ahead of time about the level of pollution in the region, allowing them to take preventative actions to safeguard their health (Che et al., 2020). Predicted air pollution trends in Maharashtra's capital city of Mumbai are analyzed using open source software called Prophet Algorithm. In order to anticipate and predict time series data, the Prophet uses machine learning algorithms. It is based on an additive model that incorporates seasonality into non-linear patterns. Use of this method produces graphs showing the trending pattern of contaminants in Mumbai's air (Sadhasivam et al., 2021).

Predicting how much rain will fall is one of the more difficult aspects of doing weather forecasts. Using accurate and timely rainfall forecasts may be quite beneficial in preparing for: current building projects, transportation activities, agricultural jobs, aviation operations, and the flood scenario, among other things (Lemos & Rood, 2010). Data mining algorithms may accurately forecast rainfall by identifying hidden trends in historical meteorological data. As a result of this study, new data mining approaches for rainfall forecasting have been critically examined. Papers published between 2013 and 2017 from well-known online search libraries are included in the study of Aftab et al. (2018). Researchers may use this study to assess the most recent advances in rainfall prediction using data mining approaches, as well as to provide a baseline against which they can make comparisons in the future (Henderson et al., 2018).

Educational data mining is used to identify and resolve educational issues in the context of teaching and learning (Romero & Ventura, 2013). In this study, a comprehensive literature review on educational data mining in mathematics and science education is conducted. According to the study subjects and data mining methods, 64 articles were analyzed (Shin & Shim, 2021). Data mining in mathematics and science education has been widely utilized to comprehend students' behavior and thinking process, uncover variables impacting student progress, and give automated evaluation of students' written work. Research in recent years has focused on developing learning systems that enhance instructors' training and students' education by using data mining methods like text mining (Romero & Ventura, 2013). More studies in the area of scientific education than mathematics education have used data mining. On the basis of past reviews and EDM research done in the context of scientific and mathematics education, this research compares the key findings of the study. Lastly, It discusses the significance of the findings for science and math education, as well as possible research avenues (Krapp & Prenzel, 2011).

2.3 Methods for predicting churn

In the current information era, new advances in BI are required for businesses to retain a competitive edge and broad appeal (Lee & Shin, 2018). Similar similarities exist between consumers who churn and those who do not. This paper provides a dynamic CCP strategy for BI by combining TA with a metaheuristic optimization technique (CCPBI-TAMO) (Pustokhina et al., 2021). In addition, the computational complexity is reduced by using the chaotic pigeon-inspired optimization-based feature selection (CPIO-FS) technique for the feature selection operation (Pustokhina et al., 2021). The next stage in optimizing the CCP's potential is a hyperparameter tweaking method known as sunflower optimization (SFO). After conducting a comprehensive simulation analysis on the standard customer churn prediction dataset, the experimental values revealed that the proposed model outperformed the compared methods, with maximum accuracy of 95.56 percent, 93.4 percent, and 92.7 percent, respectively, on the first three datasets used in the experiment (Pustokhina et al., 2021).

In the domain of customer churn prediction, there are two common algorithms with strong predictive performance and comprehensibility: DT and LR (Verbeke et al., 2012). However, these two methods have the following limitations: DT tends to struggle with linear connections between variables, while LR struggles with interaction effects between variables (Greer & Van, 2010). Consequently, the logit leaf model (LLM) is offered as a novel approach that might categorize data more accurately (Caigny et al., 2018). LLM prefers to generate distinct models on subsets of data (rather than the complete dataset), which may improve prediction accuracy while maintaining model readability (Breed, 2019). The LLM consists of two phases: the segmentation phase and the prediction phase. In the first phase, customer groups are identified, and in the second, a model is developed for each tree leaf. After the test and case study, this research identified some significant benefits of LLM over DT or LR (Caigny et al., 2018).

In order to estimate customer turnover in telecom firms, many methodologies were employed (Adwan et al., 2014). The majority of these techniques include non-linear learning techniques, such as partial least squares regression (PLS) (Hong et al., 2013). The majority of similar publications employed a single data mining technique to gather information, while the rest attempted to compare many approaches to forecast churn (Ahmad & Aljoumaa, 2019). Brandusoiu et al. (2016) suggested an up-to-date data mining strategy to predict prepaid customers' churn using a dataset of 3,333 consumers. Some functions include customer message and voicemail count information. Predicting churn factors are Tree ML methods, such as Bayes Networks and Neural Networks (NN) (Lee & Jo, 2010). Area Under the Curve (AUC) is used to evaluate the algorithm's performance (Kumar & Indrayan, 2011). 99.10% is the AUC value for Bayes Networks, whereas 99.55% is the AUC value for NN. This research used a tiny dataset and has no missing values (Nanni et al., 2012).

Makhtar et al. (2017) provided a methodology for predicting telecom customer attrition using rough set theory. In comparison to other algorithms such as DT and LR, the rough set classification method has superior prediction accuracy, according to the authors (Talasila et al., 2020). However, the majority of techniques solely concentrate on forecasting client attrition with more precision (Verbeke et al., 2011). Significantly few techniques examined the intuitiveness and comprehension of a churn prediction system in order to identify the customer churn cause (Bock & Poel, 2012). However, Idris et al. (2019) developed an enhanced churn prediction technique based on genetic programming's (GP) powerful searching ability supplemented by AdaBoost, which can identify the aspects that contribute to telecom customers' churn behavior. This work intends to utilize the searching and learning capabilities of the GP-AdaBoost algorithm to the development of an intuitive and effective churn prediction system for telecom customers (Caigny et al., 2018).

There are two well-known data mining techniques with exceptional forecast precision and readability (Martínez et al., 2015). One is the DT technique, while the other is the LR method. Nonetheless, both approaches have limitations: it is difficult for DT to deal with the linear relationships between variables, and it is difficult for LR to deal with the interaction effects between variables. Thus, the LLM technique classifies data more well. LLM's performance and readability are superior to those of DT and LR (Caigny et al., 2018).

ML and deep learning (DL) are suitable for predicting consumer loss. ISMOTE-OWELM, an improved synthetic minority oversampling approach, was utilized to increase the accuracy of customer churn prediction (Pustokhina et al., 2021).

Predicting client attrition is a problematic issue in the telecommunications sector (CCP). Business analysts are tasked with deducing the root causes of customer churn and patterns in customer behavior from existing churn data (Sudharsan & Ganesh, 2022). This work uses PSO to pick features and four sophisticated machine learning methods to predict customer turnover. Accuracy and precision are used as performance metrics in an experiment. In the first stage of the proposed technique, customer churn data is classified using classification algorithms; the accuracy of the Gradient Boosted Tree, DT, k-NN, and Naive Bayes is 93%, 90%, 89%, and 89%, respectively (Kanwal et al., 2021).

The proliferation of robust, competing service providers has created complex issues for the telecommunications industry, which has evolved rapidly with the proliferation of communication technologies. The constant loss of customers is the biggest problem facing the global telecommunications industry (Vu, 2013). The fundamental objective of this project is to create algorithms for spotting churning consumers, combining customers with similar consumption patterns, and extracting the significant patterns concealed within the gathered data (Fernández et al., 2018). Consumer data was used to construct a projected churn model, which was then applied to compute the customer churn rate for five distinct telecom providers (Ahmed & Linen, 2017). For model creation, the Pearson chi-square test, cluster analysis, and association rule mining were used to classify the significant variables. This information will guide future marketing and public relations activities (Sarker, 2021). Association rule mining with the FPGrowth component was designed to discover patterns and trends in the obtained data that have a substantial influence on the growth and revenue of telecommunications companies. Finally, the best model is found by comparison analysis, and the model's accuracy, consistency, and reliability in testing are established (Alwis et al., 2018).

In recent decades, prediction models based on data and using machine learning techniques have gained popularity (Mosavi et al., 2018). There have been several successful implementations of similar models in various fields, including medicine, crime prediction, and movie rating. The telecommunications sector has followed this pattern, using prediction algorithms to identify unsatisfied consumers (Kourou et al., 2015). Telecom companies around the world have analyzed customer churn by looking at a variety of factors with a variety of learners, including DT, SVM, NN, etc., because of the enormous financial cost of customer churn. Ahmed et al. (2017) present a comprehensive literature review covering the years 2000–2015, presenting churn prediction methods, datasets, impactful characteristics, and classifiers. The purpose of this study is to show how methods have progressed from using basic features/learners to using more advanced feature engineering/sampling methods. Ahmed et al. (2017) also discuss the problems recently encountered in churn prediction and provide some ideas for how they could be fixed. It is hoped that this study will help academics, data analysts, and telecom operators choose the most appropriate methods and characteristics for developing churn prediction models (Wassouf et al., 2020).

Since losing customers may significantly impact revenue, preventing customer churn is a top priority for any business that relies on repeat customers. In order to maximize the service provider's profit, it is crucial to building reliable churn prediction models that can be used by customer relationship management to inform the creation of successful retention programs (Lemmens & Gupta, 2020). Rodan and Faris (2015) propose training an Echo State Network (ESN) to forecast customer churn using a SVM method, both of which are often used in the telecommunications industry. Both a widely used publicly available dataset and a smaller, locally-obtained dataset are used to train and evaluate the suggested method. Based on experimental data, ESN with SVM readout is superior to other well-known machine learning models (Rodan & Faris, 2015).

Predicting customer attrition or churn is a common use case for banking technology. It considered a situation in which consumers of a multinational central bank decided to stop using the bank's services. The bank decided to look into the possible causes of such a high percentage of client churn (Alkhatib & Abualigah, 2020). Experimenting with a dataset of 10k records, Veningston et al. (2022) look for prospective consumers who are more likely to stop using the bank's value adds soon. In order to forecast future clients and spot potential churn, this article employs supervised classification models trained using various cutting-edge machine learning methods. These models were trained on the aforementioned massive amount of historical banking data. There are 13 characteristics and a class label in the dataset. It discovered that the Nave Bayes model had the highest accuracy, at 86.29 percent. Effectively using churn prediction techniques could benefit businesses in the telecommunications industry by revealing which customers are likely to switch to a competing network shortly, as well as in the human resources department by revealing which employees are likely to churn, allowing for more forethought in the selection of replacement workers (Minor et al., 2011).

With so much rivalry in this industry, it's crucial to put in the effort to understand client motivations and estimate their likelihood to leave. Many older algorithms have been used to anticipate churn and, from that, to develop a wide variety of client retention strategies. However, with the arrival of deep learning paradigms, this research has seen algorithms that provide this job with a fresh perspective (Kian & Yusoff, 2012). Because of deep learning, multilayered models may represent data at several levels of abstraction. Because it generates high-quality features automatically, it also considerably simplifies feature engineering. The results reveal that the self-learning, multilayered ANN model with tokenized data input outperforms traditional classification techniques (Fiore et al., 2013).

E-commerce has a high client attrition rate, and the statistics on consumer churn are severely skewed. Wu and Meng (2016) introduce an enhanced SMOTE and AdaBoost-based e-commerce customer churn prediction model with the dual goals of better predicting which customers would churn and making it easier to distinguish between churning and non-churning ones. To begin, the churn data is processed using an enhanced version of SMOTE that includes oversampling and undersampling techniques to deal with the imbalance issue and then uses the AdaBoost algorithm to make predictions. An empirical research conducted on a business-to-consumer electronic commerce platform provides a comparison of this model's efficiency and accuracy with those of more established customer churn prediction algorithms (Erdmann & Ponzoa, 2021).

Predicting customer turnover is crucial to managing the multi-billion-dollar search advertisements industry. A common machine learning application is the ensemble model. It takes cues from the human cognitive system and integrates many relatively weak models to improve their prediction accuracy. Wang et al. (2019) look at the efficacy of using a GBDT ensemble model to forecast a customer's churniness in light of their behavior toward search adverts. For the GBDT, it extract both moving and still characteristics. It considers evolving characteristics a time series of client actions (such as impressions and clicks). When designing static components, Wang et al. (2019) consider the end user’s preferences. Combining the static and dynamic characteristics yields an AUC value of 0.8410 for the test set. It evaluated the performance of prediction using a big amount of consumer data from the Bing Ads platform. The proposed model has been successfully implemented everyday on Bing Ads and can properly predict which customers would churn in the near future (Huang et al., 2015).

Since the early 2000s, there has been a substantial increase in both the relevance and quantity of writings addressing the topic of customer turnover in the telecom industry (Yoo & Bai, 2013). This study's objective was to undertake a quantitative bibliometric retrospection of journals that made it into the ABDC journal quality list in order to examine past studies published in those journals concerning customer churn research in the telecoms sector (Bhattacharyya & Dash, 2021). By analyzing bibliometric information, Bhattacharyya and Dash (2021) provide insight into the publishing patterns, article types, stakeholders, most popular research methods, and areas of interest during thirty years (1985–2019). Existing research makes heavy use of quantitative methods (Nardi, 2018). The study's main argument is that academics have been blind to the metatheoretical repercussions of operating under a logical positivism paradigm for far too long. In addition, this research points out future research needs and the need to expand beyond feature selection and modeling in the study of customer attrition (Bhattacharyya & Dash, 2021).

The term "customer churn" describes how a corporation loses customers. The churn rate, which is utilized for forecasting expansion, is currently seen as being as crucial as the company's net income. Companies are doing all they can to maintain a low churn rate in the face of intensifying market rivalry (Bi et al., 2016). As a result, anticipating customer turnover has become more important for retaining current clients and predicting future trends. Agrawal et al. (2018) demonstrate how a Deep Learning technique may be used to a Telco dataset to predict churn. In order to establish a non-linear classification model, a Neural Network with many layers was created. The churn prediction model incorporates customer, support, use, and environment data into its predictions. Possible attrition and the variables that may influence it are forecasted. The final weights are then applied to these characteristics by the trained model, which makes a churn prediction for that specific client. The rate of success was 80.03 percent. Because it also includes churn characteristics, the model may be used to investigate the causes of customer defection and devise strategies to counteract them (Geetha & Kumari, 2012).

Customer attrition is one of the most difficult issues that may damage a company's revenue and growth plan. In a recent poll by Gartner Tech Marketing, ninety-one percent of C-level respondents see customer turnover as a major issue. But just 43% have made extra investments to accommodate client growth. As a result, keeping the consumers you already have is crucial to expanding your business (Kelly et al., 2017). Models for predicting customer turnover using machine learning approaches have been proposed by several authors in the past, each with its unique take. For a fuller comprehension of the field, De et al. (2021) also summarize the prediction methods, datasets, and performance attained by these investigations. Even though hybrid and ensemble strategies have proven mostly beneficial at boosting model performance, the study indicates a need for criteria on model evaluation measures (Massaoudi et al., 2021). Despite the prominence of quantitative methodologies, the customer business interaction instances such as chat transcripts has not received nearly enough attention (Matthews, 2017). The paper's findings will aid new and established academics in determining where their research efforts are best directed and in raising industry knowledge of cutting-edge developments in churn prediction algorithms powered by machine learning (Vo et al., 2021).

In the fast-growing, intensely competitive telecommunications business, customer churn is a tough problem. It's of interest to researchers and business executives alike who are trying to figure out how to tell apart churning customers from those who aren't. The major reasons for this are the urgent need for companies to keep the present customers and the high expense of attracting new ones (Ref & Guan, 2012). A literature review reveals that the telecommunications industry lacks rules-based Customer Churn Prediction (CCP) techniques (Amin et al., 2017). Churn clients may be differentiated from non-churn customers, as well as those who may or may not churn in the future. According to actual investigations, RST based on GA may be utilized to extract implicit information in the form of decision rules from the benchmark telecom dataset (Amin et al., 2017). The suggested approach provides a globally optimal solution for CCP in the telecom sector when compared to a range of cutting-edge methodologies, according to comparative results (Shukla et al., 2020). A successful customer retention strategy may become an important part of telecom sector strategic planning and decision-making, which this research shows how attribute-level research may assist build (Amina et al., 2017).

Data mining relies heavily on the study of categorization. It is important to identify rules that can properly categorize records that have uncertain class membership based on data records that belong to established classes (Padhy et al., 2012). When it comes to assessing the probability of each categorization, many of them are not developed with this in mind. Consequently, they cannot be used to anticipate churn. Using such an app, people can not only tell whether and when a customer is likely to transfer service providers, but also whether or not that customer will actually do so. This enables a carrier to assess whether to deliver targeted promotions to customers who are projected to churn more often (Ascarza et al., 2018). It has been shown that DMEL is a powerful tool for discovering interesting categorization rules from a variety of data sets. The churn rate may be reliably estimated by using real telecom subscriber data (Lu et al., 2012).

Integration of longitudinal behavioral data with static consumer data might be challenging for churn prediction. Before utilizing longitudinal behavioral data in a prediction model, it is usual practice to turn the data into static data (Chen et al., 2012). The hierarchical multiple kernel support vector machine (H-MK-SVM) is an entirely novel machine learning technique (Lepot et al., 2017).. As a consequence of the sparse non-zero coefficients associated with the chosen variables, the variables are sparse, the H–MK–SVM training approach additionally comprises feature selection and temporal subsequence selection.. Three real-world datasets were utilized in computational experiments. According to computational studies, the H-MK-SVM outperforms existing classifiers in terms of performance measures (Chen et al., 2012).

Algorithms that forecast client turnover focus on those who have a high likelihood of attrition. Predicting churn requires a model that's both accurate and reasonable. These are the three most important parts of the system. Using a precise model, the customers who are most likely to depart may be identified and targeted with accuracy, while a well-defined rule set makes it possible to pinpoint the root causes of customer churn and devise an industry-specific retention strategy (Suh & Alhaery, 2016). Verbeke et al. (2011) undertake a comprehensive literature review on the use of data mining to customer attrition prediction modeling. It has been shown that approaches for churn prediction have received little attention for their clarity and intuitiveness. Using churn prediction models, comparisons are being made between two new data mining methodologies and older established rule induction techniques, such as C4.5 and RIPPER. AntMiner+ and ALBA may be used to develop precise and transparent classification rule sets. AntMiner+ enables the incorporation of domain knowledge by putting monotonicity restrictions on the final rule set. ALBA, on the other hand, employs an intelligible and highly accurate non-linear support vector machine model (Burkart & Huber, 2021). Research shows that ALBA helps students learn categorization strategies more effectively, leading to better-performing model designs. Accurate, clear and defensible are only some of the advantages of using AntMiner+ in this research (Joo, 2012).

Telecom churner predictions are tough because of the enormous datasets involved. As a result of the resulting financial losses, telecom sector stakeholders have lately been significantly more interested in churn prediction (Almuqren, 2021). Simulating the telecom churn prediction problem required a mix of mRMR, Fisher's ratio and F-score approaches (Idris et al., 2013).

The majority of explanatory features are returned by the mRMR approach, despite the fact that it is considerably easier to calculate. In addition, RotBoost's Adaboosting approach improves the accuracy of predictions by handling challenging conditions. The simulation findings demonstrate that the RotBoost-mRMR approach (CP-MRB) effectively handles the high dimensionality of telecom datasets. Predicting telecom customer turnover is a challenging topic, and CP-MRB is a useful modeling technique for it (Gopal & MohdNawi, 2021).

Chapter 3 Systematic literature analysis

3.1 Approach

This chapter conducts a literature review on customer churn prediction. Therefore, the first task is collecting relevant literature on the domain being analyzed to build a comprehensive body of knowledge (Moro et al., 2019). This is to identify the research gap and see where this research may contribute to the existing body of knowledge. Google Scholar is one of the most popular search engines for searching academic articles and publications (Harzing, 2013). The following search query: “Telecom” OR “Customer churn forecast” OR “Data mining” OR “ML” was chosen for querying its database for articles. The filters used included setting the timeframe period for publications/articles from 2010 to the present and keeping out patents. The number of hits is 17,800, and the 40 most relevant articles published in journals were gathered for deeper analysis with roots in the famous Google search engine. Only articles/publications from experiments using data-driven approaches for customer churn forecasting, for example, empirical analyzes based on real data were considered. Each of the articles was checked carefully to investigate what method was used for data analysis, the timeframe, and from which country the data came. These three dimensions comprised the three key elements for the critical analysis and comparative analysis of the literature. The study aspires better understand the inherent laws of the telecom market business and obtain a control method for telecom customer management risk. Table 3-1 answers the following questions: Which are the most used techniques in the customer churn forecast? Thus, it is possible to verify that DT, SVM, and LR are the three most popular and valuable methods. By analyzing each method shows that these three methods are efficient techniques for accurately extracting implicit information from the database.

Table 3-1: 40 relevant articles that characterize the customer churn prediction

Reference	Main Goal	Dataset	Techniques	Outcomes
Abbasimehr, 2011 (Telecom)	Built two ANFIS models for Customer Churn Prediction.	Dataset of 5000 customers	ANFIS	ANFIS shows acceptable performance in terms of accuracy and comprehensibility, and it is an appropriate choice for churn prediction applications.
Ahmad & Aljoumaa, 2019 (Telecom)	Develop a churn prediction model and use customer social network in the prediction model by extracting Social Network Analysis (SNA) features.	more than 70 Terabyte data from SyriaTel company.	ML, DT, Random Forest, Gradient Boosted Machine Tree "GBM" and Extreme Gradient Boosting "XGBOOST".	The best results were obtained by applying XGBOOST algorithm. This algorithm was used for classification in this churn predictive model.
Amin et al., 2014 (Telecom)	Study the trade-off in the selection of an effective classification model for customer churn prediction.	3333 Instances with 85.51% non-churn & 14.49% churns.	Rough Set Theory	Rough Set as a multi-class classifier will provide more accurate results for binary/ classification problem.
Amin et al., 2015 (Telecom)	Formalize a three-phase customer churn prediction method.	Dataset of 3333 instances.	IGAE, CAE, RDR	The proposed method could be a worthy alternate for churn prediction in telecommunication industry.
Amina et al., 2017 (Telecom)	Proposed a RST approach to offers a globally optimal solution for customer churn prediction in telecom sector	A publicly available dataset which consists of 3333 instances.	Exhaustive Algorithm (EA), Genetic Algorithm (GA), Covering Algorithm (CA) and the LEM2 algorithm (LA).	RST based on GA is the most efficient technique for extracting implicit knowledge in the form of decision rules from the publicly available dataset.

Au et al., 2003 (Telecom)	Develop a new data mining algorithm (DMEL), to handle classification problems of which the accuracy of each predictions made has to be estimated.	A database of 100000 subscribers' data	DMEL	DMEL is able to effectively discover interesting classification rules and predict churn accurately when applied to real telecom subscriber data.
Ballings & Poel., 2012 (Newspaper)	Investigate time window optimization with respect to predictive performance in a newspaper company	129,892 instances where 75% for estimation and 25% for validation.	LR, classification trees and bagging.	Analysts can significantly decrease data-related burdens, such as data storage, preparation and analysis.
Bock & Poel., 2012 (NA)	To Apply and evaluate GAMensPlus on six real life datasets and churn prediction projects.	Dataset of supermarket , DIY , Bank, telecom and mail order garments. The instances are from 3827 to 43,305.	GAMensPlus, bagging, random forests, RSM, logistic regression, GAM.	GAMensPlus achieves good classification performance that is performing at least as good as the benchmark algorithms.
Brandusoiu & Toderean, 2013 (Telecom)	Propose an advanced methodology for predicting customers churn in mobile telecommunications Industry.	Dateset for 3333 customers	SVM	SVM based model performs best, having 88.56% accuracy.

Chen et al., 2012 (NA)	Propose a framework for customer churn prediction directly using longitudinal data.	Foodmart: 8842 instances;
		Adventure：
		633 instances;
			SVM, NN, DT, random forests, boosting, LR, proportional hazard model	The H-MK-SVM directly using longitudinal behavioral data shows better performance than currently available classifiers.

		Telecom: 3399 instances.		
Chitra & Subashini., 2011 (Banking)	Figure out a solution for the churn problem in banking sector using data mining technique.	Dataset of 10,000 customers.	DT, DT Construction algorithm, CART	This research predicts the future churn for banking customers and can help make intervention strategies based on churn prediction to reduce the lost revenue by increasing customer retention.
Cousseme nt & Bock., 2013 (Online gambling)	Investigate if churn prediction is a valuable option in the CRM palette of the online gambling companies.	Dataset of 3729 gamblers.	DT, random forests, GAM & GAM as ensemble	Churn prediction is an useful strategy to recognize and profile the customers at risk. Moreover, the performance of the ensembles is more robust and better than the single models.
Cousseme nt et al., 2017 (Telecom)	Study the data preparation techniques to improve the prediction performance for the generally used logit model.	Dataset of 30,104 Customers.	LR, bagging, Bayesian network, Naive Bayes, DT, NN, RF, SVM, SGB	The data-preparation approach actually affects churn prediction performance;And the enhanced LR also is competitive with more advanced single and ensemble DMEL.
Huang & Kechadi., 2013 (Telecom)	Propose a new hybrid model-based learning system to obtain more accurate predictive results.	Dataset of 104,199 customer with 6,056 churners and 98,143 non-churners.	Chi-Square, DT (C 4.5), LR, k-NN, SVM, OneR, PART, SePI, k-NN-LR, KM-BoostedC 5.0	The hybrid model-based learning system is very promising and outperform the existing models.
Huang et al., 2010 (Telecom)	Present a new multiobjective feature selection method for	Dataset of 18,600 customers	DT C4.5, NSGA-II, FBSM	The proposed feature selection method is efficient for churn

	churn prediction in telecommunication service field, based on the optimisation method NSGA-II.	from Ireland telecom.		prediction with multiobjectives.
Idris et al., 2013 (Telecom)	Propose an intelligent churn prediction system for telecom by employing efficient feature extraction technique and ensemble method.	50k instances which is comprised of 190 numerical and 70 nominal features.	Random Forest, Rotation Forest, Decorate, and RotBoost ensemble with mRMR	Compared to Fisher's ratio and f-score, mRMR returns more suitable features for ensemble and CP-MRB enhances prediction of the churners in telecom industry.
Jadhav & Pawar, 2011 (Telecom)	Build a decision support system using data mining technology for churn prediction in Telecommunication Company.	895 customers	BPNN (Back Propagation NN algorithm.)	The proposed model is capable of predicting customers churn behavior well in advance.
Keramati et al., 2014 (Telecom)	Adopt data mining classification techniques including Decision Tree, ANN, K-Nearest Neighbors(KNN), and SVM so as to compare their performances.	Dataset of 3150 customers.	DT, Artificial Neural Network (ANN), KNN, SVM	Above 95% accuracy for Recall and Precision is achievable and a new method for extracting influential features in dataset was introduced and experienced.
Kim et al., 2014 (Telecom)	Propose a new procedure of the churn prediction by examining the communication patterns among Subscribers.	Dataset contained 89,412 Instances, of which 9.7% are churners.	LR and MLP NN	Established an advanced approach applying SPA method as propagation means.
Kirui et al., 2013 (Mobile Telephony)	Propose a new set of features with the aim of improving the recognition rates of possible churners.	106405 instances with 5.6% Churns	C4.5 DT, Naive Bayes and Bayesian Network.	All predictive models performed with improved prediction rate. Naive Bayes and Bayesian Network achieved higher true positive rate while C4.5 DT performed better in high accuracy rate.

Kisioglu & Topcu., 2011 (Telecom)	Build a model by Bayesian Belief Network to identify the customers' behaviors with a propensity to churn.	Dataset of 2000 subscribers.	Bayesian Belief Network (BBN), DT(CHAID algorithm)	BBN mainly deals with the causal relationships between the factors and is the most efficient way to demonstrate such relations
Lee et al., 2011 (Telecom)	Establish an accurate and compact predictive model for the churn prediction using a partial least square (PLS)-based method on highly correlated data sets among variables.	Dataset of 100,000 observations .	PLSall, PLSkernel, Logit0.15, DT, ANN, random model	The proposed model brings more accurate performance than traditional prediction models and recognizes key variables to better study churning behaviors.
Lima et al., 2011 (NA)	Present how a back testing framework can be applied for churn evaluation, enabling the validation and monitoring process for the generated churn models.	Training dataset of 10000 customers; Test dataset of 5000 customers; and Out-of-time dataset of 5000 customers.	LR, DT	This analysis can provide the foundation to make the model acceptable for future evaluation and highlight the applicability of the backtesting procedures for churn analysis.
Long et al., 2012 (Online Social Network)	Build a prediction model based on a clustering scheme to investigate the potential churn of users.	Dataset of 100,000 users.	DT-based classifier, k-means clustering algorithm	The churn and nonchurn prediction accuracies of ~65% and ~77% are reached respectively.
Lu et al., 2012 (Telecom)	Carry out a realworld study on customer churn prediction and present the use of boosting to improve a customer churn prediction model.	Dataset from a telecommunication company which includes a segment of mobile customers (in the number of millions) .	Boosting, LR	Boosting provides a good separation of churn data;so boosting is recommended for churn prediction analysis.

Moeyersoms & Martens., 2015 (Energy)	Apply high cardinality attributes to predict churn in energy industry.	Dataset of more than 1,000,000 Customers.	C4.5 DT, logistic regression, SVM	It leads to better prediction models with more data, which is not showed for "traditional" data.
Olle & Cai., 2014 (Telecom)	Propose a hybrid learning model to predict churn in mobile telecommunication networks.	2000 instances with 23 Variables.	LR and Voted perceptron,	Developed a hybrid model to predict churn with the most accurate result.
Owczarczuk., 2010 (Telecom)	Investigate the effectiveness of the popular data mining models to predict customers churn of the Polish cellular telecommunication company.	Dataset includes : the train sample – 85,274 observations , the calibration sample – 36,824 observations and the test sample – 45,497 observations .	LR, linear regression, Fisher linear discriminant analysis and decision trees.	linear models, particularly LR, are a great choice when modelling churn of the prepaid clients. Decision trees are unsteady in high percentiles of the lift curve, which is not recommend to use.
Petkovski et al., 2016 (Telecom)	Study the main reasons for churn in telecommunication sector in Macedonia.	Dataset of 34 million records from a Macedonia telecommunication Company.	Chi-Squared, C4.5, KNN, Naïve Bayes and Logistics Regression.	The classification models obtained from C4.5, KNN and LR have over 90% accuracy. The highest accuracy is achieved with LR with 94,351% accuracy.
Qureshi et al., 2013 (Telecom)	Propose widely used data mining techniques for the identification of customers who are possible to churn.	Dataset of 106,000 telecom customers.	Linear Regression, LR, ANN (ANNs), K-Means Clustering, DT.	In the condition of the data set used, DT are the most accurate classifier algorithm while identifying potential churners.

Sanchez & Asimakopoulos., 2012 (mobile communications)	Investigate the effects of MNP implementation on competition in the European mobile communications industry.	Dataset of 301 observations.	Descriptive statistics, Pearson correlation analysis, empirical analysis	The churn rates of subscriber are negatively affected by both the level of charges levied on subscribers liking to keep their current number (porting) when switching mobile providers and the length of time required to switch.
Saradhi & Palshikar., 2011 (NA)	Propose a case study for establishing and comparing predictive employee churn models, and present a simple value model to identify how many of the churned employees were ''valuable''.	Dataset of 1575 employees.	SVM, Random forests, Naïve Bayes, DT, LR.	This work has the potential for making better employee retention plans and improving employee satisfaction.
Shaaban et al., 2012 (NA)	Propose a model based on DM techniques to help a CRM department to keep track its customers and their behavior against churn.	Dataset with 23 attributes and 5000 instances	DT , SVM and NN.	DT accuracy: 77.9%, SVM accuracy: 83.7%, NN accuracy: 83.7%, the best output for the data set in hand is SVM technique.
Sharma & Kumar, 2011 (Telecom)	Develop a NN based approach to predict customer churn in subscription of cellular wireless services.	2427 customers,	ANN	NN-based approach can predict customer churn with accuracy more than 92%.
Tang et al., 2014 (Financial service)	Illustrate that efficient use of information can increase value to financial services industry and improve the prediction of customer attrition.	Dataset of 19,774 customers.	Probit-hazard model	It is necessary for the researchers and the financial service industry to collect and use derived financial information besides the information which is directly observable.

Tsai & Chen., 2010 (Telecom)	Propose the key processes of developing MOD customer churn prediction models using data mining techniques.	Dataset of 37,882 MOD customers.	NN and DT.	Applying association rules allows the DT and NN models to bring better prediction performances over a chosen validation dataset.Especially, the DT model has better performance than the NN model.
Vafeiadis et al., 2015 (Telecom)	Propose a comparative study on the most popular ML approaches applied to the challenging problem of telecom customer churning prediction.	5000 Instances	ANN, SVM, DT, Naïve Bayes, Regression analysis-LR analysis.	SVM-POLY using AdaBoost was the best overall classifier.
Verbeke et al., 2012 (Telecom)	Develop a novel, profit centric performance measure and analyze the impact of classification technique, oversampling, and input selection on the performance of a customer churn prediction model.	The smallest data contains 2180 and the largest up to 338874 observations	C4.5, CART, ADT, RF, LMT, Bag, Boost, kNN10, kNN100, NN, RBFN, RIPPER, PART, Logit, NB, BN, linSVM, rbfSVM, linLSSVM, rbfLSSVM, VP	Oversampling doesn't enhance classifiers' performance on telecom customer churn prediction and a large group of classifiers is found to yield comparable performance.
Verbeke et al., 2011 (Telecom)	Apply two novel data mining techniques to churn prediction Modeling.	5000 observations	C4.5, RIPPER	Both AntMiner+ and ALBA are shown to induce accurate as well as comprehensible classification rulesets.
Zhang et al., 2012 (Telecom)	Present a new prediction model which is based on interpersonal influence and that combines the propagation process and customers' personalized characters.	Dataset from a leading mobile service provider and include more than	LR, DT and NN.	Traditional classification models that incorporate interpersonal influence can greatly enhance prediction accuracy, and the proposed prediction model surpasses the traditional models.

3.2 Literature review

Churn prediction is an efficient way for detecting customers who are about to defect. A company's retention marketing plan may effectively target consumers who are most likely to depart (Abbasimehr, 2011).

The accuracy of churn prediction models should not be the primary consideration when evaluating them. Prediction models for churn should be accurate and understandable. Churn prediction modeling is tested against standard rule-based classifiers like C4.5 and RIPPER by using the ANFIS. FIS based on subtractive clustering and FCM based FIS were created in this work. The ANFIS-FCM and ANFIS-Subtractive models both performed well, according to the results (Abbasimehr, 2011).

Amin et al. (2014) employ rough set theory as a one-class classifier and a multi-class classifier to evaluate the trade-off in selecting an efficient classification model for predicting customer turnover. In experiments, four unique rule generating algorithms were evaluated (i.e. exhaustive, genetic, covering and LEM2). A general collection of one-class classifiers and multi-class classifiers based on evolutionary algorithms outperforms the other three rule generation approaches. Using publicly accessible datasets, this research found that a rough set as a multiclass classifier is more accurate than a rough set as a single classifier for binary/multi-class classification tasks (Amin et al., 2014).

Predicting customer attrition may be difficult, but Amin et al. (2015) help to standardize a three-phase process. Initial features are selected by an automated procedure that reduces the number of features to a manageable number while simultaneously increasing the relevance of those that remain (Singh et al., 2019). This results in a limited but highly correlated collection of features. Phase two is used to build a KBS utilizing a Ripple Down Rule (RDR)-based learner. Prudential analysis is used to solve the problem of brittleness in churn KBS by this student who acquires information about observed customer churn behavior. Any time a situation arises that isn't covered by the knowledge database, this analysis will alert the decision-maker. The last phase proposes a mechanism for evaluating Knowledge Acquisition (KA) in KB systems using a Simulated Expert (SE). When applied to publicly accessible datasets, the findings show that the proposed technique may serve as an acceptable replacement for churn prediction in the telecom sector (Fujo et al., 2022).

The key topic of the research is: How long should a customer's event history be to properly predict customer churn? While the majority of studies in predictive churn modeling seek to enhance models via data augmentation or algorithm development, Ballings and Poel. (2012) concentrate on a new dimension: the optimization of time windows in relation to predictive performance. This is the first book to provide the time window selection approach and examine the relevant literature. Using logistic regression, classification trees, and bagging with classification trees, Ballings and Poel. (2012) evaluate the impact of expanding customer event history from one to sixteen years. There is only a tiny gain in predictive performance each year, so the corporation in this study may safely delete 69% of its data with little influence on its capacity to predict the future (Johannes et al., 2014). The amount of time spent on data preparation, storage, and analysis may be drastically decreased as a result of this technique. In the era of big data, it is crucial to reduce computer complexity (Hashem et al., 2015).

To develop a dependable model for forecasting customer turnover, it is required to use a categorization strategy that satisfies both performance and interpretability criteria (De et al., 2012). Many applications and data mining contests have shown that ensemble classifiers outperform single classifiers in terms of performance in recent research (De et al., 2021). In addition, due of their increasing complexity, these models are sometimes difficult to decipher. In this paper, a classification ensemble based on generalized additive models (GAMs) is described and assessed (De et al., 2012). GAMensPlus strikes a compromise between performance and interpretability in churn prediction models. GAMens, a framework for analyzing model interpretability based on Bagging, the Random Subspace Method, and semi-parametric GAMs as component classifiers, now adds generalized feature significance scores and bootstrap confidence bands for smoothing splines (Marra & Wood, 2012). Six real-world churn prediction projects' data sets were used to demonstrate the competitiveness of the suggested method in comparison to a collection of well-known benchmark algorithms in terms of four evaluation factors (Keramati et al., 2014). In a case study using data from a European bank, it is shown that the strategy is beneficial for shedding light on the elements that influence customer attrition (Schramm et al., 2015). When it comes to churn predictors, it is shown how the analyst may decide how important they are in terms of the model's prediction accuracy using generalized feature significance ratings. Second, it has been shown that GAMensPlus is capable of recognizing nonlinear connections between predictors and churn probability (Bock & Poel., 2012).

As a result of the difficulties caused by global competition, client turnover is one of the most important concerns for businesses of all kinds. There is a 30% turnover rate in the telecommunications industry (Gunasekaran et al., 2011). To address this issue, it is necessary to use predictive algorithms to identify clients at danger of defection (Rachid et al., 2018). Mobile telecommunications customer churn may be estimated using a sophisticated method shown in this paper. The dataset's call information records include 3333 records with a total of 21 characteristics. Using four kernel functions, this research builds the prediction models using the SVM approach. Gain is a statistic for evaluating and comparing the performance of different models (Brandusoiu & Toderean, 2013).

Currently, turnover is a key problem in the banking industry. Because it costs money to recruit new customers, losing existing ones may be highly expensive (Osterwalder & Pigneur, 2010). In this article, Chitra and Subashini (2011) provide a data mining-based solution to the churn issue in the banking industry. In order to transform valuable data into knowledge, predictive data mining techniques might be useful A bank's ability to accurately predict customer attrition necessitated the usage of Classification and Regression Trees (Chitra & Subashini., 2011).

Online gambling has become one of the most lucrative sectors of the entertainment industry, resulting in intense competition and overcrowded marketplaces. Therefore, it is crucial to keep gamblers delighted (Kim & Mauborgne, 2014). Churn prediction is a possible new approach for assessing customer retention in customer relationship management (CRM) (Wassouf et al., 2020). It is the process of identifying, based on past behavior, which workers are most likely to leave the organization. This study examines churn prediction as a CRM tool for online gaming enterprises (Coussement & Bock., 2013). Using real-world data from bwin poker players, CART decision trees, random forests, and GAMens are compared to their ensemble counterparts, random forests, and GAMens. According to the study, a churn prediction model may be used to identify and profile at-risk clients. Individual model performance is less reliable than group performance (Vo et al., 2021).

The objective of data preparation is to transform categorical and continuous independent variables into a format suitable for further analysis (Coussement et al., 2017). Coussement et al. (2017) examine a number of data-preparation strategies in an effort to improve the prediction performance of the widely used logit model. In this study, churn prediction models are used, and an upgraded logit model is compared to eight cutting-edge data mining techniques using common data, including cross-sectional data from a leading European telecom operator (Gubela et al., 2020). The outcomes are congruent with the obtained conclusions. Analysts are becoming more conscious that the data preparation technique they use affects the effectiveness of churn prediction (Kayaalp, 2017). Both single and ensemble data mining methods that are more advanced can compete with the upgraded logistic regression. Finally, there are management advice and proposals for further research, and demonstration that the results may be used in different business situations. (Ucbasaran et al., 2013).

Three studies using telecom datasets were undertaken. Experiments are done to establish whether weighted k-means clustering may provide better outcomes in data partitioning, and the findings are compared with those of other well-known modeling approaches (Duwairi & Abu, 2015). In a third set, the suggested hybrid-model system is examined and contrasted with a number of other hybrid categorization systems recently presented (Seera & Lim, 2014). In addition, the findings were compared to benchmarks collected from the library of the University of California, Irvine (UCI). The results indicate that the hybrid model-based learning system outperforms existing models (Huang & Kechadi., 2013).

To increase churn prediction rates in the landline telecommunications business, Huang et al. (2010) propose a new set of characteristics based on three novel input window approaches. Henley segmentation and demographic profiling are among the new features. As predictors, DT, multilayer perceptron neural networks, and SVM are utilized to assess these new properties and window strategies (Khemakhem et al., 2018). According to the findings of the trials, churn prediction in landline telecommunications services is enhanced by combining new features and approaches (Shen et al., 2014).

Data mining and knowledge discovery from databases have been driven by the hunt for answers to real-world problems since its start (Liu & Motoda, 2012). The purpose of this work is to construct a decision support system for forecasting customer turnover in a telecoms firm using data mining methods (Jadhav & Pawar, 2011). Telecommunications companies stand to lose a lot of money if any of their customers do this. Any telecommunications company must cope with this rising challenge. Managing these frightening scenarios with a normal information system becomes more difficult as a company expands. This necessitates the development of a highly developed, customized and sophisticated decision support system. This paper explains how to use data mining to build a decision support system. Months in advance, the proposed model can predict customer turnover behavior (Bhattacharyya et al., 2011).

In today's telecom sector, identifying customers who are willing to transfer providers is essential. Telecom companies are increasingly concerned in predicting customer churn. It's hard to overstate the value of having a reliable indicator of client behavior in such a competitive market (Angelova & Zekiri, 2011). In addition to evaluating and comparing these tactics against one another, Keramati et al. (2014) also made comparisons across different well-known data mining apps utilizing data from an Iranian mobile service provider. Based on the analysis of several methods, Keramati et al. (2014) developed a hybrid approach which greatly enhanced the effectiveness of numerous assessment criteria. Recall and Precision accuracy of more than 95% may be achieved using the suggested methods. This research also demonstrated and tested a brand-new approach to extracting key characteristics from large datasets (Yao et al., 2020).

Predicting customer turnover is essential to the client retention plans of telecommunications firms, which is one of the most crucial components of CRM (Anaam et al., 2021). The objective of churn prediction is to identify probable future consumers who may opt to leave (Thanuja et al., 2011). Traditional techniques of detecting potential churning clients do not take the relationship between customers into consideration; instead, just the personal information of customers is employed. Because telecom subscribers are linked to other customers, the attributes of the individuals in the network may have an influence on sales. As a result, Kim et al. (2014) created a novel method for predicting customer churn based on subscriber communication patterns and call detail data. When the initial energy of churners (the quantity of information sent) is changed differently in churn date or centrality, propagation is more efficient (Karnstedt et al., 2010).

Many factors contribute to the high rate of customer turnover in the telecom market, including fierce competition and new technologies, cheap switching costs, and government-enacted deregulation (Khayyat, 2017). It is imperative that the industry's players develop reliable and trustworthy prediction models in order to identify and attract prospective churners in order to keep as many customers as possible. Kirui et al. (2013) provide a new set of parameters designed to increase the detection rates of potential churners. The data is compiled using call logs and client profiles, and then split into contract-related, call pattern description, and call pattern change description categories. According to experimental results, all of the models utilized were better at predicting outcomes (Langarizadeh & Moghbeli, 2016).

In the highly competitive mobile telecommunications sector, marketing managers want a business intelligence model that enables them to maintain an ideal (or near-optimal) level of customer turnover while minimizing the costs of marketing campaigns (Stone & Woodcock, 2014). Using highly correlated data sets across factors for PLS-based churn prediction is the first step toward an optimal churn management program for marketing managers, according to the conclusions of this research (Lee et al., 2011). Initial trials indicate that the proposed model is much better than traditional models and reveals essential aspects for comprehending churning behavior (Kim et al., 2017). There are also some simple marketing efforts, such as how to deal with overages and complaints, that are offered and discussed (Chen & Chen, 2015).

A company's competitive advantage is based on its ability to evaluate its customer relationships and the benefits they provide. More and more models for customer churn have been developed over time. Although these models might be complicated, they're often generated on demand when a consumer survey is requested. Using a backtesting technique, this paper shows how churn assessment may be made easier by validating and monitoring churn models that have been developed (Lima et al., 2011).

Churn is the phrase used to characterize an individual's departure from an online social network (OSN) (Nikolaou et al., 2015). The early identification and analysis of customer turnover allows the quick implementation of retention solutions (such as interventions, customised services, and enhanced user interfaces) that are useful in avoiding client churn (Zahid et al., 2019). This article provides a clustering-based prediction methodology for analyzing the probability of customer churn. In the experiment, this research utilizes data from 77,448 real-name OSN users to assess the technique (Long et al., 2012). Using the information, a set of 24 attributes may be gleaned. A decision tree classifier is used to predict churn and non-churn users for the next month. Based on their online social networking behavior, the k-means algorithm is utilized to cluster real churn users into many groups (Long et al., 2012). Results show that churn and nonchurn prediction accuracy is achieved at 65% and 77%, respectively. A total of five types of churning consumers are identified based on their OSN habits, and some suggestions for retaining these clients are provided (Ascarza et al., 2018).

Building digital CRM solutions is a new trend in the global economy, spurred on by the proliferation of digital systems and related information technologies (Reicher & Szeghegyi, 2015). This trend is especially evident in the telecommunications sector, where companies are progressively transitioning to the digital era (Lee et al., 2018). The capacity of contemporary telecom CRM systems to anticipate customer loss is a crucial characteristic. Lu et al. (2012) analyze the performance of a customer churn prediction model in the real world and suggest the usage of boosting to improve this performance. Lu et al. (2012) uncovered a small group while searching for customers who could be especially susceptible. Each cluster's churn prediction model is developed using logistic regression as the fundamental learner. A single model of logistic regression is used to examine the findings. Testing demonstrates that boosting may also be used to successfully segregate churn data for use in a forecasting study (Hanif, 2019).

High-cardinality attributes, such as bank account numbers, are categorical traits with a vast number of possible values (Moeyersoms & Martens, 2015). A predictive modeling scenario might benefit greatly from such qualities since it could be useful to know whether individuals use the same bank account number or live in the same town (Mougan et al., 2021). Predictive modeling seldom incorporates high-cardinality traits, despite their clear and palpable advantages. This is due to the fact that these characteristics cannot be included into the majority of classification models. Presented are potential transformation functions from various contexts and domains that might be utilized to integrate high-cardinality data into prediction models. Using a one-of-a-kind data collection from a large energy provider with over one million clients, (Moeyersoms & Martens., 2015) shows that the inclusion of such factors does indeed increase the model's predictive performance. Three reasons why more data is better than "traditional" data may be found experimentally. As a consequence, this research contributes to the field of big data analytics as well (Gandomi & Haider, 2015).

The purpose of this study is to evaluate the efficacy of well-known data mining methods for predicting customer defection from a Polish mobile phone service provider (Owczarczuk., 2010). Compared to past research on this issue, the following characteristics distinguish this study: All the percentiles of the lift curve are included in the tests, and the test population is made up of prepaid consumers (prior research focused on postpaid customers), who have a higher churn rate, are less stable, and are less well-documented (no application, demographic, or personal data) (Howland et al., 2017). Prepaid client attrition may be effectively modeled using linear models, notably logistic regression. Because of their instability at the higher percentiles of the lift curve, decision trees should not be used at all (Rzepakowski & Jaroszewicz, 2012).

In the last two decades, mobile communication has advanced to the point where it has surpassed all other forms of communication. In many countries, particularly industrialized ones, the market is so competitive that each new consumer must be wooed away from rivals. Due to governmental norms and standards in mobile communication, users may now easily switch carriers, generating a competitive market (Andrews et al., 2012). Mobile carriers have recently moved their attention to customer retention (Angelova & Zekiri, 2011), and churn prediction is perhaps the most important business intelligence (BI) tool for detecting clients who are leaving or switching to a competitor (Olszak, 2015). This study aims to provide frequently used data mining techniques for detecting clients eager to transfer service providers (Qureshi et al., 2013). In order to detect potential churners, these techniques examine historical data for trends. This study used many well-known approaches, including DT and ANNs (Wang et al., 2010). This study was assisted by information obtained from the Customer DNA website. It includes information on the traffic and use of 106,000 users over the period of three months. This study also examines the use of resampling to solve the issue of class imbalance. When it comes to spotting potential churners in the data set this research used, decision trees outperform all other algorithms (Kiguchi et al., 2022).

Service providers have considerably reduced the fees they charge customers for switching providers during the last two decades. Mobile Number Portability (MNP) is anticipated to substantially increase competition in the mobile sector as a result of a major legislative initiative to lower switching costs (Visser et al., 2014). When evaluating whether or not to switch service providers, the implementation of MNP has varied widely throughout the European Union countries. This is notably true in terms of porting timeframes and consumer prices. This study investigates the effect of MNP deployment on the competitiveness of the European mobile market (Ominike, 2016). Research shows that consumers who wish to preserve their existing number (known as porting) when they switch cell carriers face higher churn rates if they are paid a price to do so (Sanchez & Asimakopoulos., 2012).

In most industries, losing a client may have a negative effect on revenue and brand reputation, while attracting new customers can be difficult. Predicting customer attrition with reliable predictive models can help plans for customer retention. Predictive models of customer attrition have been developed using a variety of major machine learning technologies, which this research compares and evaluates (Leiria et al., 2011). Because of employee attrition, which is related but not the same as customer turnover, businesses suffer from delays, dissatisfied customers, and wasted time and money in the process of finding, interviewing and training new employees (Saradhi & Palshikar., 2011). This research will conduct a case study to evaluate and contrast different models for forecasting employee attrition in the future. It also provides a simple value model for workers in order to estimate how many of the churned employees were "useful." Improved employee satisfaction and a better strategy for retaining personnel might result from this approach (Fatt et al., 2010).

The objective of churn prediction is to identify clients who are considering switching service providers (Ahmed & Linen, 2017). A corporation spends five to 10 times more on customer retention than on client acquisition (Khan, 2013). It is possible to employ predictive algorithms to identify consumers who are likely to churn in the near future. In this research, Data Mining (DM) methods are employed to construct a novel prediction model (Shaaban et al., 2012). Identifying domains, selecting data, investigating data sets, classifying clustering and applying knowledge are the six processes outlined in the approach. It relies on a dataset of 5000 cases and 23 characteristics. In all, 4000 samples were used to train the model and 1000 examples were used for testing. When using a retention strategy, churners are categorized into three distinct groups. In this study, data mining methods such as DT, SVM, and NN are implemented using the free source application WEKA (Goeschel, 2016).

Because of the rapid expansion of the market in every sector, service providers now have a bigger base of customers. New rivals, innovative business methods, and better offerings have increased the cost of acquiring customers (Williamson, 2010). The benefit of maintaining existing customers in such a time-consuming setup has been realized by service providers. Because of this, service providers must work to reduce churn, which is when a customer decides to cease utilizing the services of a firm. A number of methods are used by the scientists to halt this churning. Research in the communication industry and other fields that heavily rely on consumer involvement are examined in this essay (Sharma & Kumar, 2011).

The importance of the client has been widely acknowledged in financial planning and resource allocation, particularly in the financial services sector (Mention & Bontis, 2013). According to prior research, direct observable data may be utilized to accurately predict customer attrition probabilities (Gomber et al., 2018). Consumer behavior isn't taken into consideration in these research. Tang et al. (2014) demonstrate how the financial services industry may benefit from better information use and improve customer attrition prediction. Derived data may enhance projections and help us better understand client churn. As a result, this research concludes that derived financial information should be collected and exploited by both researchers and the financial services industry (Perols, 2011).

In addition to normal TV services, the interactive system multimedia on demand (MOD) offers a variety of value-added services (Tsai & Chen., 2010). In mobile communications, data mining methods have been extensively used to create DT and NN as models for forecasting client attrition (Dalvi et al., 2016). However, a large portion of this field's work is on developing prediction models from scratch. The pre-processing step of data mining, which tries to eliminate erroneous data or information, is seldom taken into consideration by research. This research presents a methodology for constructing MOD customer churn prediction models using data mining methods (Tsai & Chen., 2010). Prediction accuracy, precision, and recall are among the four evaluation metrics not taken into account while assessing the performance of the model (Arisholm et al., 2010). The data comes from a single telecommunications carrier in Taiwan that offers MOD services. The DT model outperforms the NN model in particular. In the DT model, several important and crucial rules are also presented for marketing and management reasons, showing the components causing a high proportion of customer turnover (Tsai & Chen., 2010).

Vafeiadis et al. (2015) compare the predominant machine learning algorithms used to predict customer churn in the telecoms sector. Using cross-validation, each model was evaluated using a publically available dataset in the early phases of the project. In the second phase, research on boosting's influence on performance was undertaken. Vafeiadis et al. (2015) ran Monte Carlo simulations for each method with a large number of parameters to determine the best effective parameter combinations. According to the statistics, SVM-POLY with AdaBoost, the best overall classifier, had F-measures of more than 84% and 97% accuracy (Singh & Singh, 2020).

By identifying customers who are most likely to depart, retention initiatives may be made more successful and expenses associated with customer turnover can be decreased. This is the objective of models used to predict client attrition (Yang et al., 2012). In the first part of this study, a completely new profit-oriented performance metric is devised, which assesses how much profit may be made by enrolling customers with the highest expected attrition rates in a retention campaign (Verbeke et al., 2012). By selecting the appropriate model and customer inclusion percentage, the unique approach greatly boosts earnings as compared to statistical techniques (Abdou & Pointon, 2011).

As part of a detailed benchmarking effort, eleven real-world data sets from telecommunications providers throughout the globe were used to evaluate the usefulness of various categorisation methodologies. Research shows that just a few characteristics are needed to correctly predict customer attrition, and oversampling has little to no effect on performance.

In the end, it turns out that a huge number of classifiers have similar performance (Verbeke et al., 2012).

3.3 The literature review summary

The literature review focuses on churn prediction modeling, which is used to detect customers who are likely to leave a company. The review covers various aspects of churn prediction modeling, including evaluation criteria, techniques, and optimization. The accuracy and interpretability of churn prediction models are crucial considerations when evaluating them. The literature suggests that ensemble classifiers, such as generalized additive models, strike a balance between performance and interpretability. Furthermore, the optimization of time windows in relation to predictive performance is a new dimension in churn prediction modeling. After conducting a review of relevant literature, it was found that increasing the amount of customer event data analyzed from one year to sixteen years resulted in only a negligible improvement in predictive accuracy. Therefore, it may be possible to reduce the amount of time and resources required for data preparation, storage, and analysis by focusing on more recent customer data. Finally, the telecommunications industry has a high turnover rate, and predictive algorithms are essential in identifying customers at risk of defection.

3.4 The methodology used for selecting the literature

The methodology used for selecting the literature in the above literature review involved the use of specific search queries and a defined research range. The search queries used included "Telecom," "Customer churn forecast," "Data mining," and "ML." These search queries were used to identify relevant literature on the domain being analyzed.

The research range included the 40 most relevant articles/publications published in journals from 2010 to the present with patents, aiming at choosing reliable and valid sources.. This range was selected to ensure that the literature review covered the most recent and relevant research on the topic of customer churn prediction in the telecom industry.

The literature review mainly focused on three directions: Customer Churn Prevention, Data Mining Theory, and Methods of Churn Prediction. The literature was selected based on its relevance to these three directions and its ability to contribute to the existing body of knowledge on the topic.

All the literature included in the text is related to churn prediction, which is the main topic discussed. Each paper provides insights, approaches, or techniques that could be helpful in predicting customer churn. These literature comes from different authors, institutions, and countries, providing a diverse range of perspectives and approaches to churn prediction. Overall, the selection process aimes at providing a comprehensive and up-to-date review of the literature on churn prediction, focusing on the latest research, diverse approaches, and reliable sources.

3.5 Discussion

The 40 articles were published in a total of 30 different journals (Table 3-2 shows those from which more than one article was selected), corresponding to 11 different publishers (Table 3-3 for those publishers with more than one selected). Such numbers prove that telecom forecasting is not limited to specific telecom literature, even though telecom gets the largest share, with 68% of articles; on the contrary, the investigations found a more extensive range of sciences, with a particular emphasis on the bank, energy, and online social network literature (Table 3-4). The fact that big data in the telecom industry are currently helpful for discovering cutting-edge information technologies makes it an exciting subject for empirical investigations to evaluate novel data modeling approaches and applications (Mikalef et al., 2019). Even so, it is a leading journal that covers telecom literature, such as Expert Systems with Applications, which accommodates the highest number of publications focusing on telecom forecasting. From the perspective of the publisher, Elsevier and ieeexplore.ieee.org are currently the two publishers ahead in telecom forecasting journal article publications. Based on the 40 articles analyzed, three main aspects were analyzed:

(1) the main goal and outcome of each study;

(2) the dataset (from where the data were extracted and data volume); and

(3) the techniques adopted.

Since all the articles present empirical data-driven experiments, it is interesting to understand from which years the data was gathered for the experiments to evaluate if the periods are recent enough. It shows that most of the articles perform experiments based on data from the yearly 2011's, with few articles before 2010. One of the key dimensions of data-driven knowledge discovery is data recency, especially considering that telecom customers' behavior has changed over the years. Therefore, using recent data decreases the risk of negatively influencing models built on these data for forecasting telecom business demand.

In addition, artificial intelligence techniques such as SVM (adopted ten times) and NN (applied four times) now appear as the dominant method. It would be attractive to observe future reserves for artificial intelligence applications to telecom customer churn forecasting.

Table 3-2: Journals from which more than one article was selected

Journal	**No. of articles**
Expert systems with Applications	10
IEEE Transactions on Industrial.	4
European Journal of operational research	3
Decision support systems.	2
Neurocomputing	2

Table 3-3: Publishers from which more than one article was selected

Publisher	No. of articles
Elsevier	26
ieeexplore.ieee.org	6
Springer	3
researchgate.net	2
Citeseer	1
arxiv.org	1

Table 3-4: Research domain from journals from which articles were selected

Research domain	No. of articles
Telecom	28
Banking	1
Energy	1
Financial Service	1
Online Social Network	1
Newspaper	1
Online Gambling	1

Predicting telecom customer churn has been a hot topic for a long time, which is also becoming one of the most critical issues for telecom companies since they need to understand customers' different need to avoid customer churn and profit loss.

The present investigation provides a recent literature review on data-based empirical research for forecasting customer churn by providing a summary of the literature covering 40 relevant publications, mainly from 2010 up to June 2019, thus a very recent timeframe. The present chapter reviews the most recent trends in this domain, focusing on the future regarding customer churn prediction and trying to find the research gap.

The findings show that DT, SVM, and LR are the three most popular and valuable methods. Besides, artificial intelligence techniques are already practically used to predict customers' behavior. Especially, artificial NN is outstandingly recognized as a competent prediction method. In addition, the literature found is not limited to telecom journals, verifying that telecom themes are also of interest for a more extensive range of social sciences (e.g., Banking) and that telecom data comprises a vital asset for evaluating novel for prediction modeling technologies. Based on this chapter above, a customer churn model will be established to predict whether the telecom customer will be lost or retained. The model will combine data mining technology with the rich data resources of the telecom industry and the latest Marketing theories, which will not only maximize customer acceptance of telecom packages within a manageable risk range but also help increase the company's business volume and revenue. It would also be attractive to study which trends will emerge in customer churn prediction in the future.

Chapter 4 Hypotheses and Proposed Model

4.1 Contextualization

Loss of customers is a major issue for telecoms since it reduces revenue (Bach et al., 2021). The telecommunications industry is a very competitive one, and as a result, it is becoming more difficult to keep clients in an increasingly competitive worldwide market. Companies pay a lot of money in marketing to bring in new customers, but keeping an existing one costs far less in the long run (Kim et al., 2020). Because of this, keeping customers from leaving is high on the list of priorities for telecom providers.

The term "customer churn" describes the phenomenon of a company's clients defecting to a rival. By predicting customer churn, businesses may learn what drives customers away and develop strategies to keep them as customers for as long as possible, so maximizing revenue (Xie et al., 2009). Understanding the reasons why a client may want to cancel service is crucial for telecommunications firms since it gives them an edge in the market.

Previous studies have attempted to understand the factors that could have impact to the customer churn. For instance, Kisioglu and Topcu, (2011) mentioned that customer churn is affected by factors such as average time of calls, average billing fee and tariff type. Oghojafor et al. (2012) proposed the conclusion that high call rates, inadequate service facilities, off-beam advertising medium, availability of better service provider, and unappealing service plans are the primary contributors to customer turnover in Nigeria's telecommunications industry. Besides that, According to the data, poor voice quality, spam messages, poor network quality, and unexpected fees are the leading causes of customer turnover in Pakistan's telecommunications sector (Akmal, 2017). In spite of the fact that many academic investigations have attempted to identify the causes of customer defection. However, no study has tried to investigate the impacts of the factors specificlly for Chinese telecom industry.

Facing this identified gap in the literature, this study aims to identify and investigate the specific factors that impact telecom customers churn in Chinese telecom industry such as the total fee receivable, the fixed monthly cost, the local fee, the roaming fee, China Unicom’s network fee, the fee with China Mobile positively, the fixed-line fee positively, the total monthly caller MOU, the total monthly called MOU and the total local called MOU. Then hypotheses and proposed model will be made and related literature review will be conducted to support the hypotheses. Our study extends the previous work by Zhang, (2018) by innovatively showing how these factors could impact chinese telecom customers churn. Thus, we propose the following research questions: What factors will lead to customer loss? With this research's findings in hand, telecom managers should be better equipped to identify the causes of customer churn and develop effective retention policies.

4.2 Factors affecting customer churn

Technological progress is crucial in determining who will be the market leader and achieving better market performance (Asimakopoulos & Whalley, 2017). Meanwhile, technological progress has already changed the competition and the game’s rules in the telecom industry. In the past, telecom operators generally won customers through price competition. However, today’s consumers pay more attention to differentiated and value-added services, which has increased switching costs while making consumers more loyal (Aydin & Özer, 2005). In the telecom section, technological progress could help companies identify customers with a high risk of churn and to establish a business strategy with customer retention as the core goal, which will make the companies healthier and allow for long-term operation (Almana et al., 2014). Finally, the development of telecommunication technologies has also brought about more market competition and higher customer churn rates. The customer churn rate for telecom customers in the European market has reached 30%, while in Asia, it has reached 60% (Olle & Cai, 2014).

A Bayesian belief network analysis concluded that the average tariff amount would affect customer churn. The other factors are the average call time and tariff type (Kisioglu & Topcu, 2011). The tariff structure will affect customers’ perceptions of value, affecting customer churn (Iyengar et al., 2011). A multilayer perceptron (MLP) analysis of a sample of five thousand Jordanian telecom customers concluded that the monthly tariff is the most significant factor affecting customer churn (Mahajan et al., 2017). Tariffs for domestic calls are essential in predicting customer loss (Shukla et al., 2021). The telecom section has two types of pricing: two-part tariff and pay-per-use pricing. Compared with two-part tariffs, pay-per-use pricing can reduce the customer churn rate by 10.5% (Iyengar et al., 2011). Discriminant analysis and t-test of one thousand Indian telecom customers concluded that the tariff rates for calls and customer satisfaction with the telecom service offered are the key factors determining customer churn (Mahajan et al., 2017).

As competition in the telecommunications market intensifies, providing tariff price promotions and differentiated services for key customers will be an efficient method to avoid customer churn (Jahanzeb & Jabeen, 2007). In the Korean market, the tariff rate is one of the critical factors determining customer churn. Tariffs and customer care services are the main factors influencing customer satisfaction and churn, as shown using discriminant and regression analyses (Kim & Yoon, 2004). Service quality in the telecom industry refers to Internet signal quality. Good service quality will improve customer satisfaction and loyalty, lowering the risk of customer loss (Kim et al., 2004).

Additionally, it will also help to attract new customers. A factor analysis and regression analysis concluded that tariffs and service quality are critical factors in prepaid customer churn. Hence, companies must monitor and improve service quality (Mahajan et al., 2017).

Customer retention and loss are influenced by the customers' sociodemographic characteristics and satisfaction (Mahajan et al., 2017). The customers' sociodemographic data, for example, regarding gender, could be used to predict whether customers will be lost or not (Verbeke et al., 2012). Age and gender will influence telecom customers' preferences and behavior. People under thirty years value customer service quality, value-added services, and mobile service fees. The tariff is not a critical factor in determining churn for this segment. However, those older than thirty years pay more attention to tariff pricing, which will largely influence their retention or loss (Seo et al., 2008).

Forecasting client turnover is a crucial requirement for many businesses, and it demands a comprehensive grasp of customer attrition rates (Ascarza et al., 2018). This study introduces a novel prediction model that takes into account both the spread of interpersonal impact and the particular characteristics of consumers (Hung et al., 2010). As part of their contribution, (Zhang et al., 2012) examined how interpersonal influence affects prediction accuracy, taking into account factors that other researchers have found useful. And they compared a number of models based on machine learning and statistical methods to ensure the validity of the evaluation. More than one million consumers of a prominent conventional and network-based mobile telecommunications service provider are represented in this study's data collection. Inclusion of interpersonal influence in empirical investigations has been demonstrated to increase the accuracy of standard classification approaches, but the suggested prediction model exceeds the conventional methods (Connelly & Ones, 2010).

In the telecom business, employee turnover is a huge issue. As a result, businesses are emphasizing the development of precise and reliable prediction algorithms to identify consumers who are likely to churn in the near future (Ullah et al., 2019). This study aims to identify the primary causes of customer turnover in Macedonia's telecommunications industry (Petkovski et al., 2016). The proposed technique for churn prediction analysis includes understanding the company, data processing, analysis, and implementation of several classification algorithms (Ullah et al., 2019). The results of a Macedonian telecoms firm will tremendously help the management and marketing teams of other telecommunications companies throughout the nation and the globe (Mirkovski et al., 2016).

Huang et al. (2012) provide a fresh set of characteristics for forecasting landline customer attrition. These characteristics include call data, line information, payment data and complaints. Huang et al. (2012) apply seven distinct prediction techniques to the problem of customer churn based on these updated attributes: LR, linear classification, naive Bayes, DT, multilayer perceptron NN, SVM, and the evolutionary data mining algorithm. The results of experiments comparing the new feature set to seven modeling approaches for churn prediction were compiled. The results of the experiments show that the unique characteristics developed using the six modeling methods are superior to the ones already in use for predicting customer turnover in the telecommunications industry (Lu et al., 2012).

4.3 Hypotheses and Proposed Model

Customer consumption tags distinguish customers by expense-related information, such as monthly fee, package type, or mobile terminal price (Jia et al., 2019). Precision marketing can be performed using telecom data to classify and identify customers. Using such information will allow telecom operators to concentrate on the target customers and convert them into potential customers. This could significantly optimize marketing expenses and avoid customer churn (Jia et al., 2019).

Expense-related data could be applied to understand the reasons for customer loss. Customers with similar consumption–expense behaviors have similar reasons for churn. Users with similar expense-related characteristics could be segmented into groups to conduct an analysis (Xu et al., 2021). Thus, we propose the following hypotheses:

H1: The total fee receivable for the month positively impacts customer loss;

H2: The fixed monthly cost has a positive impact on customer loss;

H3: The local fee has a positive impact on customer loss;

H4: The roaming fee has a positive impact on customer loss;

H5: China Unicom’s network fee has a positive impact on customer loss;

H6: The fee with China Mobile positively impacts customer loss;

H7: The fixed-line fee positively impacts customer loss.

Taiwan’s telecommunication industry has experienced fierce competition since it removed the restriction of wireless telecom services, and customer churn management has become the operators’ focus to retain telecom customers by satisfying their needs. One main challenge is predicting customer churn (Hung et al., 2006).

Using empirical analysis, different data mining methods that can be used to allocate ‘propensity-to-churn’ scores were evaluated from customer and operator perspectives. The results showed that call data, NN and DT methods could be applied for accurate customer churn prediction models. Furthermore, the customers’ recent six-month transactions can be applied to predict customer churn for the coming month. The call data can also be included in the transaction data. Thus, we proposed the following hypotheses:

H8: The total monthly caller MOU positively impacts customer loss;

H9: The total monthly called MOU has a positive impact on customer loss;

H10: The total local called MOU positively impacts customer loss.

The Data Warehouse system, which accumulates telecom data, such as SMS, was used to increase the customer retention rate for SyriaTel. Generally, all SMS and MMS data that indicate customer behavior should be used, as it is unknown which features will be valuable in predicting churn (Ahmad et al., 2019).

The SMS and MMS data for daily, weekly, and monthly users were aggregated for the research to identify related variables and see how they relate to each other. Three charts were built using three kinds of weights: (1) the standardized SMS and MMS quantities; (2) the standardized customer calling times; (3) the mean of the first two standardized weights. Two features for each chart were produced by applying the SenderRank and PageRank algorithms according to the directed charts (Ahmad et al., 2019).

The Indian liberalization and globalization process has influenced the telecom industry. The marked leader Airtel was selected to conduct a case study through its value proposition approach by concentrating on new value-added services such as the new SMS Pack plan (Dwivedi & Sharma, 2011). Consequently, the following hypotheses were also assessed:

H11: China Unicom's SMS quantity positively impacts customer loss;

H12: China Mobile's SMS quantity positively impacts customer loss;

H13: China Telecom's SMS quantity positively impacts customer loss.

Our hypotheses are listed in Table 4-1

Table 4-1: The hypotheses of the study.

Hypotheses	Description
H1	The total fee receivable for the month positively impacts customer loss.
H2	The fixed monthly cost has a positive impact on customer loss.
H3	The local fee has a positive impact on customer loss.
H4	The roaming fee has a positive impact on customer loss.
H5	China Unicom's network fee has a positive impact on customer loss.
H6	The fee with China Mobile has a positive impact on customer loss.
H7	The fixed-line fee has a positive impact on customer loss.
H8	The total monthly caller MOU has a positive impact on customer loss.
H9	Total monthly called MOU has a positive impact on customer loss.

H10	The total local caller MOU has a positive impact on customer loss.
H11	China Unicom's SMS quantity has a positive impact on customer loss.
H12	China Mobile's SMS quantity has a positive impact on customer loss.
H13	China Telecom's SMS quantity has a positive impact on customer loss.

Chapter 5 Approach and results

5.1 Methodology

5.1.1. Data Collection

Client data were provided by three major Chinese telecommunication operators: China Mobile, China Unicom, and China Telecom. These data included the information for 4126 clients from 2007 to 2018, and anonymous demographic information, business information, and basic metadata regarding the clients' fees, calls, and SMS and MMS activity.

The information from the dataset is shown in Table 5-1.

Table 5-1: Dataset information.

Information	Characterization
Demographic and business information	Sex Age Career Non-fixed monthly fee Fixed monthly fee
Phone calls (mobile and fixed line)	Monthly minutes in local calls Monthly minutes in long distance calls using mobile phone roaming service Monthly minutes in calls using fixed line Minutes of usage (MOU)
SMS	Number of SMS
MMS	Number of MMS

5.1.2. Data Analysis

For data analysis, we used SPSS. Factor analysis, Pearson correlation, chi-square, and discriminant and LR analysis methods were used to predict customer churn (Lee et al., 2017).

The meanings of the independent variables from F1 to F6 are shown in Table 5-2.

Table 5-2: The meanings of independent variables - adapted from (Zhang, 2018)

Independent Variable	Meaning
F1	Common factor of non-monthly fixed cost
F2	Common factor of monthly fixed cost
F3	Common factor of the calls MOU
F4	Common factor of long-distance and roaming call
F5	Common factor of SMS
F6	Common factor of China Unicom's MMS

5.1.3. Dataset Description

The samples' sex characteristics are shown in Table 5-3 and Figure 5-1. Of the 4126 customers, 1184 were females (28.7%), and 2942 were males (71.3%).

Table 5-3: Sex characteristics of the sample.

	Number	Proportion	Valid Proportion	Accumulative Proportion
F	1184	28.7	28.7	28.7
M	2942	71.3	71.3	100.0
Total	4126	100.0	100.0	

Sex: M, male; F, female.

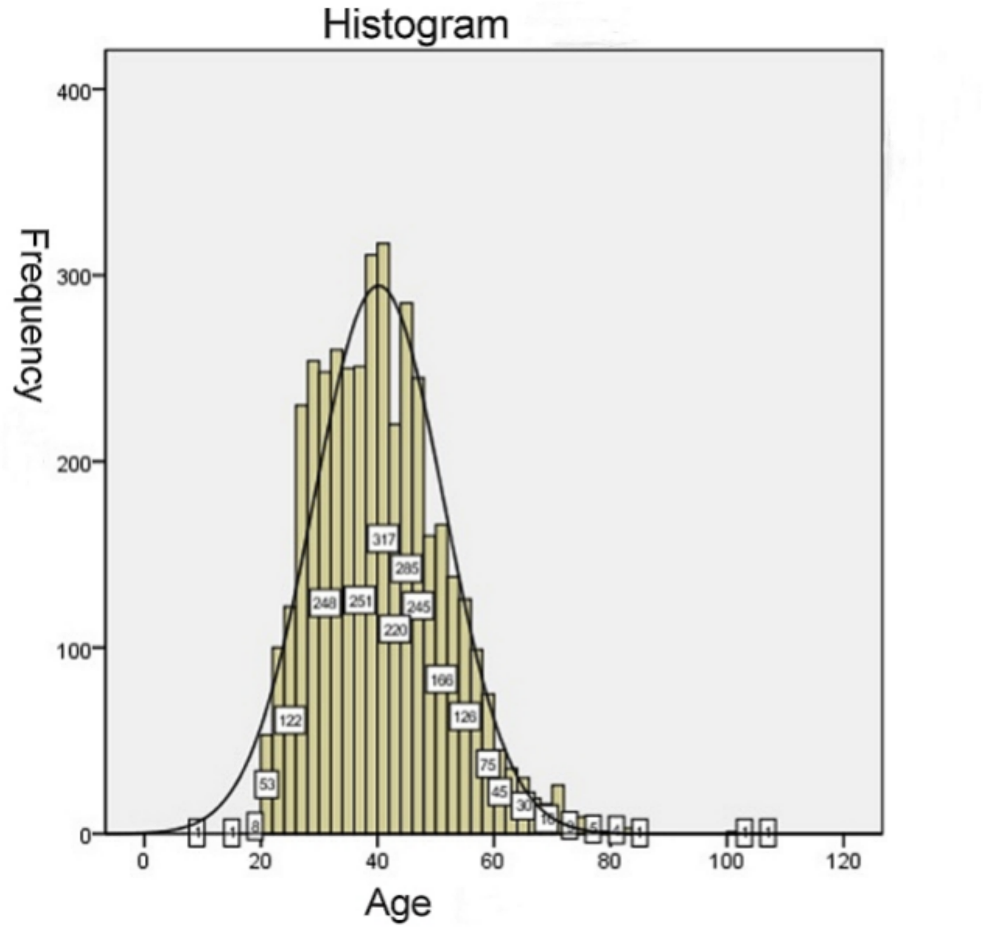

Figure 5-1: Age distribution of the sample.

Among the 4126 customers, the ages ranged from 9 to 107. However, the most common ages ranged from 20 to 60, representing 95% of the total. Customers aged 40 were most represented, with 165 cases (4%).

5.2 Discussion and results

5.2.1 Factor Analysis to Characterize Expense, Call, and SMS Attributes

5.2.1.1 Expense Factor Analysis

5.2.1.1.1 Variable Selection

Factor analysis refers to the concept that significant and measured variables can be decreased to less latent variables with common variance (Bartholomew et al., 2011). Some factors are unobservable and unmeasurable, but variables can be reduced into the same group based on similar characteristics to test the relationships (Barton et al., 1973). Expense data, such as monthly fees, package type, or mobile terminal price data, can distinguish customers into different customer consumption tags (Jia et al., 2019). Cost and expense management is critical to the operation of companies, and the factor analysis approach could be used to study the expense and cost data and to understand the relationships between the variables (Xue & Hong, 2016). Telecom customer cost data, such as wireless data fees, are suitable for factor analysis and could be used to understand customer behavior (Zhang et al., 2011). Thus, the telecom customers' expense data were selected for the following factor analysis. All expense-related factors, including the (1) total fee receivable for the month, (2) fixed monthly costs, (3) local fee, (4) roaming fee, (5) Unicom's network fee, (6) China Mobile's fee, and (7) fixed-line fee, were used to conduct the factor analysis and analyze the characteristics of the cost factors. Later, Kaiser–Meyer–Olkin (KMO) and Bartlett tests were applied to identify whether these factors are suitable for factor analysis.

5.2.1.1.2 Research Hypothesis Testing: KMO and Bartlett Sphericity Tests

The KMO and Bartlett tests were carried out to identify whether the data could be used to conduct a factor analysis with good effect. If the KMO measures of sampling adequacy are > 0.5 or the value of Sig is < 0.05, the data can be used to conduct a factor analysis with good effect. The KMO and Bartlett test results for expense data are shown in Table 5-4. The KMO measures of sampling adequacy were 0.599 > 0.5, and the value of Sig was 0.000 < 0.05. Therefore, it was concluded that the data were suitable for factor analysis.

Table 5-4: KMO and Bartlett tests.

Test of KMO and Bartlett		
KMO measures of sampling adequacy		0.599
Bartlett testing of sphericity	Value of Chi-square	22244.842
	Value of df	21
	Value of Sig.	0.000

5.2.1.1.3 Common Factor Variance of Expenses

Factor analysis needs to extract overlapping information for variables to reduce them. This requires that the original variables must have strong correlations with each other. If there is no overlapping information between the variables, they cannot be integrated and concentrated, and there is no need to perform the factor analysis.

We applied the common factor variance to judge the degree of information condensing via factor analysis (Table 5-5). The common extracted factor values reached a maximum value of 87.8% and a minimum of 57.8%, with most being more significant than 60%. The effect was good, and each variable's information loss was low. It can be concluded that the results were representative and reliable.

Table 5-5: Common factor variance results - adapted from (Zhang, 2018)

	Initial Value	Extraction Value
Total fee receivable for the month	1	0.857
Fixed monthly cost	1	0.878
Local fee	1	0.682
Roaming fee	1	0.578
Unicom network fee	1	0.590
Fee with China Mobile	1	0.838
Fixed line fee	1	0.672
Principal component analysis applied to extract values.		

5.2.1.1.4 Total Interpretation Variance

The cumulative variance of the first two factors was 72.798%, suggesting that most of the observed variables were fully represented (Table 5-6). Therefore, the common factors F1 and F2 were selected.

Table 5-6: Total interpretation variance.

Component	Eigenvalues Starting Value			Squared Sum Extraction Loading			Square Sum Rotation Loading		
	Sum	Variance %	Accumulative %	Sum	Variance %	Accumulative %	Sum	Variance %	Accumulative %
1	4.086	58.369	58.369	4.086	58.369	58.369	4.054	57.910	57.910
2	1.010	14.429	72.798	1.010	14.429	72.798	1.042	14.888	72.798
3	0.912	13.024	85.822						
4	0.473	6.756	92.578						
5	0.298	4.263	96.841						
6	0.171	2.447	99.288						
7	0.050	0.712	100.000						

Figure 5-2 shows a screen plot. The horizontal axis shows the component numbers, while the vertical axis shows the eigenvalues. The eigenvalues for the first two common factors, 1 and 2, were more significant than 1, which meant they were suitable for analysis.

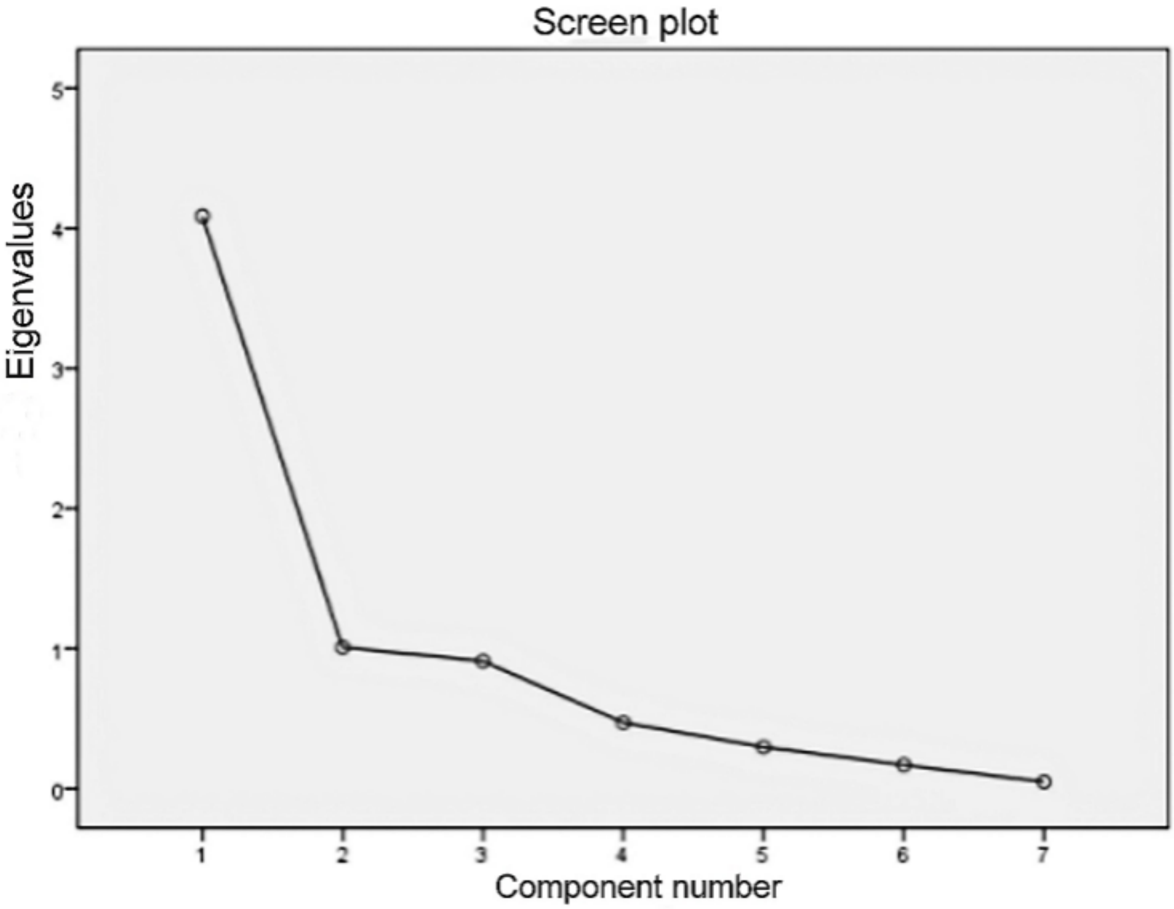

Figure 5-2: Scree plot.

5.2.1.1.5 Component Matrix

A component score coefficient matrix is shown in Table 5-7. F2 had a more significant load for the number of fixed monthly costs. Additionally, except for the small load on the fixed monthly cost, the first factor has the same load on the other cost factors. Therefore, the first factor, F1 can explain the non-monthly fixed cost factor.

Table 5-7: Component score coefficient matrix.

	Ingredient	
	1	2
Total fee receivable for the month	0.217	0.100
Fixed monthly cost	−0.063	0.918
Local fee	0.217	−0.241
Roaming fee	0.157	0.226
Unicom network fee	0.198	−0.123
Fee with China Mobile	0.229	−0.043
Fixed line fee	0.195	0.062

Principal component analysis applied to extract values.

Therefore, we confidently concluded that F1 (common factor of non-monthly fixed costs) and F2 (common factor of monthly fixed costs) could characterize the expense attributes. The formulas used are shown below, which were adapted from (Zhang, 2018):

F1 = 0.217×Total fee receivable for the month − 0.063×Fixed monthly cost + 0.217×Local fee + 0.157×Roaming fee + 0.198×Unicom network fee + 0.229×Fee with China Mobile + 0.195×Fee with fixed line

F2 = 0.100×Total fee receivable for the month + 0.918×Fixed monthly cost − 0.241×Local fee + 0.226×Roaming fee − 0.123×Unicom network fee − 0.043×Fee with China Mobile + 0.062×Fee with fixed-line.

5.2.1.2 Factor Analysis of Telecom Customer Calls

5.2.1.2.1 Variable Selection

Customer call data, such as total monthly, long-distance, and roaming calls, are suitable for factor analysis to investigate the main factors influencing customer preference for the service provider (Paulrajan & Rajkumar, 2011). Factor analysis was conducted on several variables, including customer call data, to identify the main factors determining customer loyalty. It was concluded that better call quality and service would positively influence customer loyalty (Jessy, 2011). Thus, the telecom customers’ call data were selected for the following factor analysis. The following call-related factors were used to conduct the factor analysis and analyze the characteristics of cost factors:

(1) total monthly traffic MOU;

(2) total monthly caller MOU;

(3) total monthly called MOU;

(4) total local MOU;

(5) total local called MOU;

(6) total long-distance MOU;

(7) total roaming MOU.

Later KMO and Bartlett sphericity tests were applied to identify whether these factors were suitable for factor analysis.

5.2.1.2.2 Research Hypotheses Testing: KMO and Bartlett Sphericity Tests

The KMO and Bartlett test results for call data are shown in Table 5-8. The KMO measures of sampling adequacy were 0.555 > 0.5, and the value of Sig was 0.000 < 0.05. It was concluded that the data were suitable for factor analysis.

Table 5-8: KMO and Bartlett test results.

KMO and Bartlett’s Test		
KMO measures of sampling adequacy		0.555
Bartlett testing of sphericity	Value of Chi-square	102964.374
	Value of df	21
	Value of Sig.	0.000

5.2.1.2.3 Common Factor Variance

The common factor variance results are shown in Table 5-9. The common extracted factor values ranged between 47.5% and 99.6%. Most of these extraction values were greater than 80%, revealing an ideal overall effect. The results were considered scientific and representative, as each variable’s loss rate was low.

Table 5-9: Common factor variance.

	Initial Value	Extraction Value
Total monthly traffic MOU	1	0.996
Total monthly caller MOU	1	0.882
Total monthly called MOU	1	0.931
Total local MOU	1	0.978
Total local called MOU	1	0.961
Total long-distance MOU	1	0.620
Total Roaming MOU	1	0.475
Principal component analysis applied to extract values.		

5.2.1.2.4 Total Interpretation Variance

The cumulative variance reached 83.463% (Table 5-10), suggesting that most observed variables were represented. Therefore, most of the original information was replaced by factors F1 and F2.

Table 5-10: Total interpretation variance.

Component	Eigenvalue Starting Value			Squared Sum Extraction Loading			Square Sum Rotation Loading		
	Sum	Variance %	Accumulative %	Sum	Variance %	Accumulative %	Sum	Variance %	Accumulative %
1	4.713	67.330	67.330	4.713	67.330	67.330	4.191	59.869	59.869
2	1.129	16.133	83.463	1.129	16.133	83.463	1.652	23.594	83.463
3	0.910	13.000	96.464						
4	0.245	3.502	99.966						
5	0.002	0.026	99.992						
6	0.001	0.008	100.000						
7	1.148	0.000	100.000						

The scree plot is displayed in Figure 5-3. The horizontal axis shows the component numbers, while the vertical axis shows the eigenvalues. The feasibility of the first two common factors was revealed, as the eigenvalues of the first two common factors, 1 and 2, were more significant than 1.

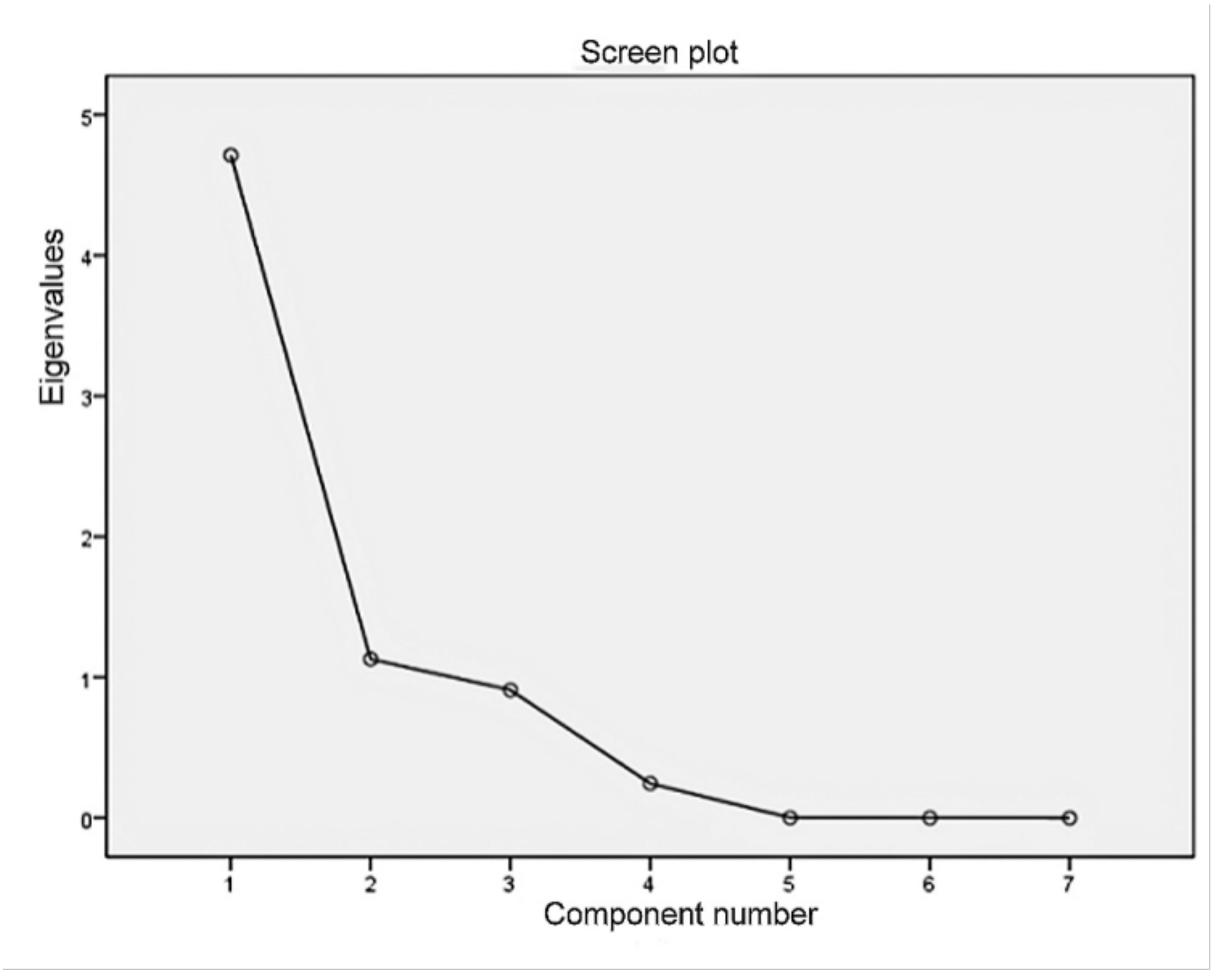

Figure 5-3: Scree plot.

5.2.1.2.5 Component Matrix

The component score coefficient matrix is shown in Table 5-11. F4 had more significant loads for the total long-distance MOU and total roaming MOU. Therefore, long-distance and roaming calls were resumed as the second factor F4. Additionally, the total monthly called MOU, total local MOU, and total local called MOU numbers showed significant loads for the first factor F3. Therefore, the first factor F3 can explain the called MOU factor.

Table 5-11: Component score coefficient matrix.

	Ingredient	
	1	2
Total monthly traffic MOU	0.179	0.120
Total monthly caller MOU	0.073	0.317
Total monthly called MOU	0.257	−0.100
Total local MOU	0.283	−0.160
Total local called MOU	0.294	−0.199
Total long-distance MOU	−0.119	0.553
Total Roaming MOU	−0.160	0.540

Principal component analysis applied to extract values.

Therefore, we confidently concluded that F3 (common factors of the called MOU) and F4 (common factors of long-distance and roaming calls) characterize the call attributes. The formulas for calculation were as below, which were adapted from (Zhang, 2018):

F3 = 0.179×Total monthly traffic MOU + 0.073×Total monthly caller MOU + 0.257×Total monthly called MOU + 0.283×Total local MOU + 0.294×Total local called MOU − 0.119×Total long-distance MOU − 0.160×Total Roaming MOU

F4 = 0.120×Total monthly traffic MOU + 0.317×Total monthly caller MOU − 0.100×Total monthly called MOU − 0.160×Total local MOU − 0.199×Total local called MOU − 0.553×Total long-distance MOU − 0.540×Total Roaming MOU

5.2.1.3 SMS of Telecom Customers Factor Analysis

5.2.1.3.1 Selection of Variables

Factor analyses are performed to explore the factors influencing telecom customer experiences using certain variables, including customer SMS and MMS data (Subramanian & Palaniappan, 2016). Customer SMS data relating to the SMS quantity in the telecom package are suitable for use in factor analyses, which could help telecom companies to identify the factors that impact customer satisfaction and loyalty (Alam & Rubel, 2014). The telecom sector has achieved impressive development in Bangladesh. Customer SMS data has been used in factor analyses, helping to understand the relationship between SMS data and customer loss (Amin et al., 2019). Thus, the telecom customers' SMS and MMS data in the data source were selected to conduct the following factor analysis. All SMS-related factors, including (1) China Unicom's SMS quantity, (2) China Mobile's SMS quantity, (3) China Telecom's SMS quantity, (4) China Unicom's MMS quantity, and (5) CRBT, were used to conduct the factor analysis and analyze the characteristics of the cost factors. Later, KMO and Bartlett tests were applied to identify whether these factors could be used to conduct the factor analysis.

5.2.1.3.2 Research Hypothesis Testing: KMO and Bartlett Tests of Sphericity

The KMO and Bartlett test results for SMS data are shown in Table 5-12. The KMO measures of sampling adequacy were 0.567 > 0.5, and the value of Sig was 0.000 < 0.05. It was concluded that the data were suitable for factor analysis.

Table 5-12: Test of KMO and Bartlett.

Test of KMO and Bartlett		
KMO measures of sampling adequacy		0.567
Bartlett testing of sphericity	Value of Chi-square	636.772
	Value of df	10
	Value of Sig.	0.000

5.2.1.3.3 Common factor variance

The results of common factor variance are shown in Table 5-13. The common factor extracted revealed results more significant than 50%. The results were considered scientific and representative, as each variable's loss rate was low.

Table 5-13: Common factor variance - adapted from (Zhang, 2018)

	Initial Value	Extraction Value
China Unicom SMS quantity	1	0.580
China Mobile SMS quantity	1	0.594
China Telecom SMS quantity	1	0.545
China Unicom MMS quantity	1	0.556
CRBT	1	0.629
Principal component analysis applied to extract values.		

5.2.1.3.4 Total Variance of Interpretation

The cumulative variance was 50.087% (Table 5-14), suggesting that most observed variables were represented. Therefore, most of the original information was replaced by factors F1 and F2.

Table 5-14: Total interpretation variance.

Component	Eigenvalue Starting Value			Squared Sum Extraction Loading			Square Sum Rotation Loading		
	Sum	Variance %	Accumulative %	Sum	Variance %	Accumulative %	Sum	Variance %	Accumulative %
1	1.458	29.151	29.151	1.458	29.151	29.151	1.313	26.265	26.265
2	1.047	20.937	50.087	1.047	20.937	50.087	1.191	23.822	50.087
3	0.971	19.423	69.51						
4	0.815	16.291	85.801						
5	0.71	14.199	100						

The scree plot is displayed in Figure 5-4. The horizontal axis shows the component numbers, while the vertical axis shows the eigenvalues. The feasibility of the first two common factors was revealed, as the eigenvalues of the first two common factors, 1 and 2, were more significant than 1.

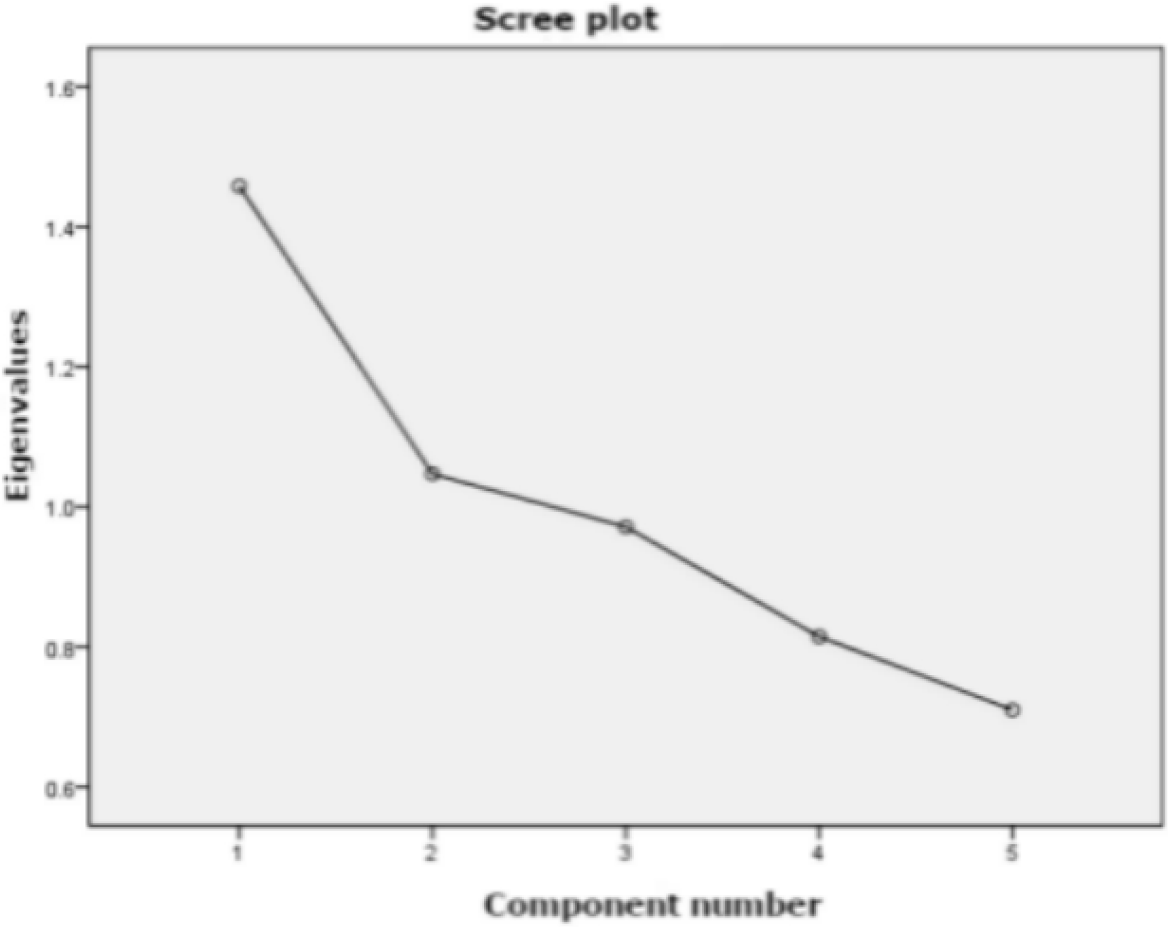

Figure 5-4: Scree plot.

5.2.1.3.5 Component Matrix

The component score coefficient matrix is shown in Table 5-15. F6 had more significant loads for China Unicom's MMS quantity and CRBT. Therefore, MMS and CRBT were resumed as the second factor F6. Moreover, the first factor F5 showed more significant loads for China Unicom's SMS quantity, China Mobile's SMS quantity, and China Telecom's SMS quantity. Therefore, the SMS quantity can be explained by the first factor F5.

Table 5-15: Component score coefficient matrix.

	Ingredient	
	1	2
China Unicom’s SMS quantity	0.596	−0.106
China Mobile’s SMS quantity	0.570	0.035
China Telecom’s SMS quantity	0.295	−0.034
China Unicom’s MMS quantity	0.011	0.614
CRBT	−0.120	0.685

Principal component analysis applied to extract values.

Therefore, it can be concluded that F5 characterizes the SMS, while F6 characterizes MMS and CRBT. The used formulas were as follows:

F5 = 0.596×China Unicom SMS quantity’s SMS + 0.570×China Mobile SMS quantity’s SMS + 0.295×China Telecom SMS quantity’s SMS + 0.011×China Unicom MMS quantity’s MMS − 0.120×CRBT

F6 = −0.106×China Unicom SMS quantity’s SMS + 0.035×China Mobile SMS quantity’s SMS − 0.034×China Telecom SMS quantity + 0.614×China Unicom MMS quantity + 0.685×CRBT

5.2.2 Discriminant Telecom Customer Loss Model

5.2.2.1 Empirical Analysis for the Discriminant Model

5.2.2.1.1 Discriminant Attributes

According to the data, the discriminant analysis revealed an appropriate discriminant model. The model refers to the discrimination between the sample and the parent. First, historical data are established from the samples' discriminant distances. Then, each sample's data are replaced with the discriminant distance to calculate the actual distance.

5.2.2.1.2 Analysis of Discriminant Model

The discriminant model's eigenvalues were analyzed to identify the discriminating judgment power of the function. Then, Wilks' lambda discriminant test was applied to confirm the significance of the discriminant function, i.e., whether the discriminant function was valid. Afterward, Fisher's linear discriminant function was used for the telecom customer loss prediction equation, indicating the key factors (F1, F2, F3, F4, F5, and F6) that could influence the telecom customer churn. Finally, an accuracy test was conducted for the discriminant function to investigate the accuracy of the discriminant equation.

Eigenvalues of the discriminant function

The discriminant model was used in the analysis. In the table below, when the discriminant model's eigenvalue is higher, the model's discriminating judgment power is higher. The last column represents the canonical correlation coefficient, while the results reveal an acceptable range due to the discriminant function's eigenvalue (0.030) and canonical correlation (0.171) (Table 5-16). In the Table 5-16, "a" means that the former canonical discriminant function was used in the analysis.

Table 5-16: Eigenvalues.

The Function	The Eigenvalues	Variance Percentage	Accumulative Percentage	Canonical Correlation
1	0.030a	100	100	0.171

Wilks' Lambda discriminant test

Wilks' lambda is the ratio of the within-group sum of squares to the total sum. The value is one when the group means for all observations are equal; it is close to zero when the within-group variation is small compared to the total variation. Thus, a considerable Wilks' lambda value indicates that the means of each group are more or less equal; a small Wilks' lambda value shows that the means of each group are different. It can be seen from Table 5-17 that the first discriminant function explained 97.1% of all variations. Moreover, the value of Sig. was 0.000 < 0.05, meaning that this discriminant function was significantly established.

Table 5-17: Wilks' lambda values.

Function Testing	Wilks' Lambda Value	Chi-square Value	Value of df	Value of Sig.
1	0.971	121.638	7	0

Fisher's linear discriminant function test

Y1 and Y2 represent the customer churn and customer existence, respectively (Table 5-18).

Table 5-18: Classification function coefficients - adapted from (Zhang, 2018)

	The Loss or Retain of Customers	
	Customer Loss-Y1	Customer Retain-Y2
Factor score for F1	−1.518	−0.1
Factor score for F2	0.257	0.176
Factor score for F3	0.588	0.135
Factor score for F4	6.021	0.291
Factor score for F5	−0.712	−0.211
Factor score for F6	−1.051	−0.02
Gender	5.963	6.397
(The constant)	−16.592	−4.81

The Fisher linear discriminant equation

The established Fisher discriminant equation was as follows, which was adapted from (Zhang, 2018):

$Y1 = -16.592 - 1.518 \times F1 + 0.257 \times F2 + 0.588 \times F3 + 6.021 \times F4 - 0.712 \times F5 - 1.051 \times F6 + 5.963 \times Gender$

$Y2 = -4.810 - 0.100 \times F1 + 0.176 \times F2 + 0.135 \times F3 + 0.291 \times F4 - 0.211 \times F5 - 0.020 \times F6 + 6.397 \times Gender$

The discriminant model indicates the top factors that could be used to forecast telecom customer churn. The classification is considered Y1 if the result is one, revealing customer churn. If the result is zero, the classification is Y2, suggesting customer retention.

Accuracy test for discriminant function

One hundred random samples from the dataset were chosen to conduct the accuracy test. The results are shown in Table 5-19. Half of them were lost customers, and half were retained customers. The one hundred random samples were imported into the telecom customer churn discrimination model. Then, the predicted customers churn results were obtained to judge the prediction accuracy rate of the model.

Table 5-19: Discriminant result checklist - adapted from (Zhang, 2018)

Test Results	Customer Retain	Customer Loss	Sum
Total	50	50	100
Successful prediction	36	39	75
Failure prediction	14	11	25
The accuracy rate	0.720	0.780	0.750

From the above Table, we can see that the overall prediction accuracy rate was 75%. Among the 50 retained customers, 36 were predicted successfully. The accuracy rate was 72%. Furthermore, among the 50 churn customers, 39 of them were predicted successfully, and the accuracy rate was 78%.

5.2.3 LR Model of Telecom Customer Churn Prediction

It can be seen from Table 5-20 that a total of 19 items, such as the Total fee receivable for the month, are independent variables. Moreover, filter_$, which means the customer is lost or retailed, is the dependent variable for binary LR analysis to build the customer loss prediction model. When filter_$ is one suggests that the customer is lost. When filter_$ is 0, it suggests that the customer will be retained. Based on these results, we can estimate whether or not a customer will stay with a telecommunications service provider based on the information in the dataset. The model formula is: ln(p/1 − p) = −2.056 − 0.002×Total fee receivable for the month − 0.308×Fixed monthly cost − 0.077×Local fee+ 0.023×Roaming fee + 0.041×Unicom network fee + 0.031×Fee with China Mobile + 0.032× Fee with fixed-line + 0.003×China Unicom SMS quantity + 0.004×China Mobile SMS quantity + 0.003×China Telecom SMS quantity + 0.009×China Unicom MMS quantity + 0.238×CRBT − 0.539×Total monthly traffic MOU − 0.016×Total monthly caller MOU − 0.057×Total monthly called MOU + 0.559×Total local MOU + 0.039×Total local called MOU + 0.548×Total long-distance MOU + 0.510×Total Roaming MOU (where p represents the probability that filter_$ is 1, which indicates that the customer will be lost. Furthermore, 1-p represents the probability that filter_$ is 0, which indicates that the customer will be retained).

Table 5-20: Binary LR.

Item	Regression Coefficients	Standard Error	z Value	Wald χ^2	p Value	OR Value	OR Value 95% CI
Total fee receivable for the month	−0.002	0.005	−0.402	0.162	0.688	0.998	0.988~1.008
Fixed monthly cost	−0.308	0.027	−11.564	133.734	0.000	0.735	0.698~0.774
Local fee	−0.077	0.010	−7.979	63.665	0.000	0.926	0.908~0.943
Roaming fee	0.023	0.021	1.051	1.104	0.293	1.023	0.981~1.067
Unicom network fee	0.041	0.010	3.988	15.906	0.000	1.041	1.021~1.062
Fee with China Mobile	0.031	0.009	3.639	13.243	0.000	1.032	1.014~1.049
Fee with fixed line	0.032	0.010	3.254	10.590	0.001	1.032	1.013~1.052
China Unicom SMS quantity	0.003	0.001	3.466	12.014	0.001	1.003	1.001~1.004
China Mobile SMS quantity	0.004	0.001	7.168	51.379	0.000	1.004	1.003~1.005
China Telecom SMS quantity	0.003	0.006	0.511	0.261	0.609	1.003	0.992~1.014
China Unicom MMS quantity	0.009	0.002	4.250	18.058	0.000	1.009	1.005~1.013
CRBT	0.238	0.031	7.599	57.740	0.000	1.268	1.193~1.348
Total monthly traffic MOU	−0.539	0.252	−2.143	4.592	0.032	0.583	0.356~0.955
Total monthly caller MOU	−0.016	0.006	−2.711	7.348	0.007	0.984	0.973~0.996
Total monthly called MOU	−0.057	0.023	−2.529	6.395	0.011	0.945	0.904~0.987
Total local MOU	0.559	0.252	2.217	4.916	0.027	1.749	1.067~2.867
Total local called MOU	0.039	0.022	1.812	3.284	0.070	1.040	0.997~1.086
Total long-distance MOU	0.548	0.251	2.182	4.763	0.029	1.730	1.057~2.830
Total Roaming MOU	0.510	0.254	2.010	4.041	0.044	1.665	1.013~2.736
Intercept	−2.056	0.121	−16.974	288.116	0.000	0.128	0.101~0.162
Dependent variable: filter_$							

According to the parameter test, the regression coefficient of the total monthly fee receivable was −0.002, but this was not significant since $z = -0.402$ and $p = 0.688 > 0.05$. This suggests that the total fee receivable for the month will not affect filter_$. Thus, hypothesis 1 was rejected, meaning that the total monthly fee receivable does not positively impact customer loss.

The regression coefficient of the fixed monthly cost was −0.308, which was significant since z = −11.564 and p = 0.000 < 0.05, suggesting that the fixed monthly cost will have a significant negative impact on customer churn. Moreover, the dominance ratio (OR value) was 0.735, suggesting that when the fixed monthly cost increases by one unit, the decrease in Y is 0.735 times. Thus, hypothesis 2 was rejected, suggesting that the monthly fixed cost does not positively impact customer loss.

The summary analysis showed that Unicom's network fee, China Mobile's network fee, fixed-line fee, China Unicom's SMS quantity, China Mobile's SMS quantity, China Unicom's MMS quantity, CRBT, total local MOU, total long-distance MOU, and total roaming MOU have a significant favorable influence on the customer churn. On the other hand, the fixed monthly cost, local fee, total monthly traffic MOU, total monthly caller MOU, and total monthly called MOU negatively impact the customer churn. However, the total fee receivable for the month, roaming fee, China Telecom's SMS quantity, and total local called MOU do not affect the customer churn. Therefore, H1, H2, H3, H4, H8, H9, H10, and H13 were rejected, while H5, H6, H7, H11, and H12 were confirmed.

In Table 5-21, the model's overall prediction accuracy is shown to be 93.94%, and the model's fit is acceptable. The LR analysis and hypothesis tests show that expense, SMS, and call information influence customer churn. Moreover, the accuracy-test for the LR prediction model proved that it has good prediction performance, with an accuracy rate of 93.94%. Thus, it is possible to estimate whether or not a customer will stay with a telecommunications service provider based on information from the data. This investigation indicates that the LR method could accurately predict customer churn.

Table 5-21: Binary LR prediction accuracy rate.

		Forecast Value		Forecast Accuracy	Forecast Error Rate
		0	1		
True value	0	3823	36	99.07%	0.93%
	1	214	53	19.85%	80.15%
Summary				93.94%	6.06%

5.2.4 Discussion

A growing number of service providers are focusing on increasing their customer base as the telecommunications sector continues to thrive. In today's extremely competitive industry, retaining existing customers is a critical issue for organizations (Jawaria et al., 2010). According to the telecom industry estimate, it costs far more to acquire a new customer than it does to maintain an existing one. Therefore, accumulating data from the telecom sector may aid in forecasting client loyalty and, ultimately, turnover. Telecommunications companies need to take the necessary steps to begin acquiring their related clients if they want to maintain a stable market value (Dahiya & Bhatia, 2015).

Customers who are considering switching to a rival or canceling their subscription are referred to as "customer churn" in the ICT business. Predicting and controlling this kind of conduct is crucial for markets and competitiveness in the real world (Mustafa et al., 2021). Hudaib et al. (2015) examine three distinct hybrid techniques with the objective of developing a dependable and effective churn prediction model. The three models are based on two phases: the phase of clustering and the phase of prediction. Hudaib et al. (2015) begin by filtering the acquired customer data. In the subsequent phase, it attempts to predict how consumers will behave. The first model employs the k-means algorithm for data filtering, and Multilayer Perceptron Artificial Neural Networks for prediction (MLP-ANN). Model 2 uses MLP-ANN, a hierarchical clustering approach (Al et al., 2015). The third alternative is a combination of MLP-ANN and self-organizing maps (SOM). Accuracy and churn rate metrics are used to evaluate the three models and and upon inspection, it is clear that the three hybrid versions fare better than the two common variants (Le et al., 2011).

A novel strategy for analyzing and forecasting client attrition has been offered. In the financial sector, the technique employs a data mining methodology. An growing turnover rate of almost 1.5 million annual consumers has prompted this. Client attrition forecasting is also known as churn customer prediction (Hassani et al., 2018). Karvana et al. (2019) compared 5 distinct categorization strategies using a dataset with 57 features. The experiments were repeated multiple times, each time comparing groups of students from various academic levels. There was a 50/50 split in the results of a comparison using a support vector machine. Predicting client turnover at an Indonesian private bank is best accomplished with class sample data. Any business that wants to reduce customer attrition may use the information gained from this modeling to inform their strategies (Tamaddoni et al., 2010).

Decision makers and the machine learning community find Customer Churn Prediction (CCP) difficult since churn and non-churn consumers often have similar characteristics. Classifiers' accuracy varies across regions of a dataset, as observed in separate tests on customer turnover and related data (Jeyakarthic & Venkatesh, 2020). The confidence and accuracy of a classifier's prediction tend to go hand in hand in such cases. If a method is given to measure the classifier's confidence for separate zones within the data, it is possible to calculate the expected accuracy of the classifier prior to classification (Bokulich et al., 2018). This research presents a new CCP technique (Amin et al., 2019), the dataset is divided into zones based on the distance factor, with each zone holding either I data with high confidence or (ii) data with low certainty, which may be used to predict whether or not a customer would churn (Xu et al., 2021). Evaluation metrics (such as accuracy, f-measure, precision, and recall) applied to multiple publicly available datasets from the Telecommunications Industry (TCI) demonstrate that I the distance factor is strongly correlated with the certainty of the classifier, and (ii) the classifier obtained high accuracy in the zone with greater distance factor's value (i.e., customer churn and non-churn with high certainty) than in the zone with smaller distance factor's value (i.e., customer (i.e., customer churn and non-churn with low certainty) (Amin et al., 2019).

To effectively anticipate customer churn, one firm (the target) may need data from another company (the source), which is where the field of CCCP comes in. Before developing a CCCP model, it is common practice to convert the cross-company data into a set of target company data with a normal distribution that is consistent with the CCCP data (Grover et al., 2018). Nonetheless, the best strategy for data transformation in CCCP is still unknown. Furthermore, in the telecommunications industry, there has not been a full investigation into how different data transformation techniques affect the performance of CCCP models using various classifiers. A model for customer churn prediction (CCCP) was developed using data transformation techniques (i.e., log, z-score, rank, and box-cox), and it not only provided a thorough comparison to validate the impact of these transformation techniques on CCCP, but also evaluated the efficiency of underlying baseline classifiers (i.e., Naive Bayes (NB), K-Nearest Neighbor (KNN), and Gradient Boosting (GB)). Using datasets made available by the telecom sector for research reasons, experiments were conducted (Ishaq et al., 2021). The results showed that most data transformation techniques (including log, rank, and box-cox) greatly boost CCCP's efficiency. However, when compared to the other data transformation strategies, the Z-Score strategy did not outperform the others. As an extra bonus, it is observed that the CCCP model based on NB excels on converted data, whereas DP, KNN, and GBT performed on par, and the SRI classifier did not achieve statistically significant outcomes (Amin, 2019).

Caigny et al. (2020) investigate how the addition of textual data to CCP models may increase their accuracy. It contributes to the current research by comparing CNNs to state-of-the-art approaches for text analysis in CCP (Li et al., 2021). First, the results are consistent with previous research indicating that adding textual input to a CCP model improves its predictive accuracy (Ramesh et al., 2022). CNNs are superior than state-of-the-art techniques when it comes to text mining in CCP (Lee & Hsiang, 2020). But it is hard to construct churn prediction models that can compete with those based on traditional structured data using just unstructured textual data (Gandomi & Haider, 2015).

Telecommunications firms, facing increased competition in the mobile industry, must prioritize retention efforts or risk losing customers. Communications firms may reasonably devise measures to prevent the loss of consumers in a two-month time frame. But it will cause a great deal of unpredictability and bring up a lot of uncertainty. Predicting client attrition is challenging for two key reasons. In the first place, there is a major asymmetry in the customer turnover dataset. In addition, as the feature space includes several dimensions, dimension reduction is necessary (Khalid et al., 2014). Li et al. (2016) propose a unique supervised one-side sampling approach for pre-processing the imbalanced data set as a solution to these difficulties. Using the random forest method for dimension reduction and feature selection. In this study, a C5.0 decision tree classifier was employed to forecast customer churn during the following two to three months (Li et al., 2016). About 2.7 million 4G wireless subscriber records are used for research. The results indicate an accuracy of 80.42% and a recall rate of 52.43%. The suggested model yields positive prediction results that may be put to use in the real world to prevent the loss of paying consumers (Barberis, 2013).

Telecom businesses, like many others in an increasingly crowded market, depend extensively on predictive churn models to identify potential customers who may cancel their service. To identify the most lucrative churn model, earlier research has developed the EMPC (Van et al., 2020). But the architecture of the model does not explicitly include profit considerations. Therefore, this research provides a classifier called ProfLogit, which uses a genetic algorithm to optimize the EMPC during training; ProfLogit's internal model structure is similar to that of a lasso-regularized logistic model (Praseeda & Shivakumar, 2021). Furthermore, it offers the predicted profit maximizing fraction-based recall and accuracy metrics that are threshold-independent. The method is geared on building churn models for retention efforts that maximize profits for businesses. Another interesting discovery is that ProfLogit, like the lasso, selects features depending on how profitable they are to the model rather than how accurate they are (Stripling et al., 2018).

Telcos have a significant challenge in retaining customers, who often defect to competing services. The reason for this is because consumers are now being recognized as a company's most valuable asset. As a result, a growing number of businesses are allocating resources into research and development of systems that may effectively anticipate client attrition (Umayaparvathi & Iyakutti, 2012). If businesses could foresee which customers have the possibility to churn, they may provide retention solutions and optimize their marketing efforts to great effect (Ascarza et al., 2018). Using Particle Swarm Optimization and a Feedforward neural network, Faris (2018) proposes a smart hybrid model for churn prediction. If this study aims to increase the neural network's capacity to predict, it may employ PSO to optimize the network's structure and fine-tune its input data weights simultaneously (Faris, 2018). In addition, the proposed model incorporates a sophisticated oversampling method to account for the data's unequal class distribution. The results of the assessment indicate that the proposed model improves the client coverage rate (Zhao et al., 2021). The model is also highly interpretable, with feature weights providing insight into the relative value of individual features during categorization (Murdoch et al., 2019).

In the telecommunications sector, managing customer correlation and maintaining customers are crucial, and customer churn is a big contributor to both of these issues (Kisioglu & Topcu, 2011). Data mining methods may be used in order to better predict client churn (Ahmad, 2011). This study includes both a PPFCM-based clustering suggestion and an ANN-based churn prediction (Gopal & MohdNawi, 2021). In the clustering module, the input dataset instances are grouped using a probabilistic possibilistic fuzzy C-means clustering method (Du, 2010). Clustered test data are used to choose the ANN classifier with the highest accuracy, as determined by minimum distance or similarity measurements. Finally, the output score value is used to forecast client turnover. Experiments are run in three distinct phases: (1) to implement the PPFCM clustering method; (2) to evaluate the classification result; and (3) to verify the suggested hybrid model presentation. When compared to other models, the suggested hybrid PPFCM-ANN model outperforms them all (Vijaya et al., 2020).

High customer turnover in online stores is accompanied with an unbalanced customer churn data set. Wu and Meng (2016) introduce an enhanced SMOTE and AdaBoost-based e-commerce customer churn prediction model with the dual goals of better predicting which customers would churn and making it easier to distinguish between churning and non-churning ones. The churn data is first processed using an enhanced version of SMOTE that includes oversampling and undersampling techniques to deal with the imbalance issue, and then the AdaBoost algorithm is used to make predictions. Finally, the empirical research on B2C E-commerce platform demonstrates that the model outperforms the established customer churn prediction algorithms (Ulas et al., 2023).

A systematic churn prediction model is required to monitor client churn in order to maintain operations (Provost & Fawcett, 2013). Sivasankar and Vijay (2019) examine the potential of Synthetic Minority Oversampling TEchnique (SMOTE) to mitigate the data imbalance. Area Under the Curve (AUC), sensitivity, and specificity are among the measures used to evaluate the prediction ability of classifiers (Hightower et al., 2010). In terms of predicting churners, simulation results indicate that the recommended technique based on SMOTE, co-relation, and ensembling beats simply applying learners to the raw dataset. Since churn is a major problem in the telecommunications business, this technique may be useful (Mishra & Rani, 2017).

The mobile Internet sector often deals with class-imbalanced datasets. Relative Weight, ,Standardized Regression Coefficients and Random Forest (RF) were the feature selection approaches, Oversampling, Undersampling, and Synthetic Minority Oversampling were the balancing methods, and Standardized Regression Coefficients were the classification method. The integrated models include a feature selection approach, a balancing mechanism, and a classification strategy (Gui, 2017).

Both feature selection and balancing algorithms' efficacies were tested against the original dataset. According to the data collected throughout the experiments, the lowest value of Cost = 1085 was achieved when SRC was used in conjunction with the SMOTE method. The most crucial characteristics for achieving the lowest possible telecommunications costs were determined by the application of the Cost calculation to all models. When used together, these models have the potential to optimize profit while minimizing expenditures associated with client retention, hence lowering churn rates (Devriendt et al., 2021).

The term "customer churn" describes the process through which a corporation loses customers. As a growth estimation tool, churn rate has risen in prominence to rival financial profit as a key performance indicator (Agrawal et al., 2018). Companies are doing all they can to maintain a low churn rate in the face of intensifying market rivalry. As a result, forecasting customer turnover has become more important, not only for keeping current customers, but also for anticipating how they could behave in the future. This study use Deep Learning to anticipate churn on a Telco data set (Sivasankar & Vijaya, 2019). In order to establish a non-linear classification model, a Neural Network with many layers was created (Bach et al., 2015). The churn prediction model takes into account customer, support, use, and environment data to make predictions. It is possible to foretell the elements that will lead to customer defection and the likelihood of it happening. The trained model then applies the final weights to these factors, enabling a forecast of the probability of churn for that particular customer. The rate of success was 80.03 percent. The model's inclusion of churn elements enables businesses to investigate the causes of churn and devise strategies to eradicate it (Agrawal et al., 2018).

Obtaining new clients is more expensive than keeping old ones, according to decision-makers and business experts. For business analysts and CRM analysts, understanding the reasons of customer churn and the behavior patterns of prior churn customers is critical (Zaslavsky et al., 2013). The authors of this paper provide a churn prediction model using classification and clustering techniques (Ullah et al., 2019). As a first stage, the proposed model categorizes churn customer data using classification algorithms; the RF method demonstrated good results, correctly identifying occurrences in 88.63% of all tests (Idris et al., 2012). The major duty of the CRM system is to reduce client churn by establishing and executing effective retention policies (Linoff & Berry, 2011). The proposed technique splits churning customer data further by clustering churning customers using cosine similarity to present cluster-based retention offers. This article also contains important churn criteria for identifying the sources of churn (Ullah et al., 2019). By recognizing and capitalizing on the most crucial churn causes gathered from customer data, customer relationship management (CRM) solutions enable firms to significantly improve their marketing efforts (Verhoef et al., 2010). Utilizing metrics such as accuracy rate and precision, the proposed churn prediction model's performance is evaluated. Prior models using the RF algorithm and k-means clustering to describe consumers were surpassed by the recommended churn prediction model. Further, the rules gene explains why customers depart (Ullah et al., 2019).

Due to the sensitive nature of the information, churn data from the telecom industry is rarely made public. Orange, a French telecoms operator, offers the KDD Cup 09 competition a telecom customer attrition data set in 2009 (Idris et al., 2012). As part of the KDD Cup 09, participants are tasked with reducing the number of features from which they may build a classification model, and this novel feature reduction approach is being utilized to investigate the effect that various characteristics have on this prediction. Zhao et al. (2017) introduce the K-local maximum margin (KLMM) technique for feature extraction. By studying the partition rules of the diversification subspace, the potential field structure may be built. When seen via the lens of data source in the scalability dimension, the basic link between data attributes and classification results is shown (Sundaram et al., 2019). By reducing the dimension, the resulting features may make churn prediction in telecom data simpler on the eyes (Kantardzic, 2011). To account for the anisotropy of features, the KLMM technique modifies the auto selection sigma factor. To determine which attribute weight is most crucial, the potential function is applied to the set of attribute weights. Data points may be separated into their respective classes using the retrieved features using KLMM, as shown through experiments and analyses (Giordani et al., 2018).

Using techniques from social network analytics, the telecommunications sector may successfully foresee client attrition. It has been shown in particular that relational learners tailored to this challenge improve predictive model accuracy (Xevelonakis & Som, 2012). Oskarsdóttir et al. (2017) conduct a statistical analysis to determine the effect of relational classifiers and collective inference methods on the predictive ability of relational learners. In conclusion, Oskarsdóttir et al. (2017) investigate how network topology influences model performance, and their findings indicate that edge and weight definitions do impact the output of predictive models.

Predictive algorithms are increasingly being used in customer retention initiatives to identify at-risk customers in a large user base. Predicting customer attrition may be framed as a binary classification issue from a machine learning viewpoint. The purpose of developing classification algorithms is to correctly predict the likelihood of client defection based on prior behavior (Guelman et al., 2012). The predictive churn models are often selected based on accuracy-related performance parameters, such as the area under the ROC curve (AUC) (Höppner et al., 2020). Misclassification costs and advantages from a proper classified are typically overlooked in these models, making them misaligned with the main business goal of profit maximization. Thus, it is desirable to build lucrative churn prediction models that can also be interpreted. A new indicator, EMPC, has been introduced to assist in choosing the best churn model (Stripling et al., 2018). To this end, Óskarsdóttir et al. (2017) provide a novel classifier that builds the EMPC measure in from the start. ProfTree is the method for developing profitable decision trees using an evolutionary algorithm. This research demonstrates the superior profitability of ProfTree over traditional accuracy-driven tree-based approaches in a benchmark analysis using real-world datasets from a variety of telecommunications service providers (Höppner et al., 2020).

Several research have shown the usefulness of relational learning in networked data. Together, relational classifiers and collective inference techniques form relational learners, which permit inference of network nodes given the presence and strength of linkages to other nodes (Natarajan et al., 2012). Telecommunications businesses have already begun using these tools to make forecasts about customer turnover, suggesting that doing so may improve forecasting accuracy. Óskarsdóttir et al. (2016) apply a variety of relational learners to several CDR datasets originating from the telecommunications industry and compare their performance on each dataset with the objective of ranking them all and independently examining the effects of relational classifiers and collective inference methods. The best relational classifier, according to this study, is the network-only link-based classifier, which develops a logistic model by assessing each node's linkages as opposed to using standard measures of association (Verbeke et al., 2014).

Keeping current clients has evolved into a top concern. However, a successful retention strategy requires careful planning and the identification of both at-risk and valuable consumers. The boosting methodology, a nonparametric approach, provides greater prediction in the majority of cases (i.e., over a wide range of sample sizes, purchase frequencies, and churn percentages) (Tamaddoni et al., 2016). In addition, logistic regression is preferred in situations with low churn rates. Finally, parametric probability models excel if the size of the client base is quite small (Hayes et al., 2015).

Academics have created churn prediction models for the telecom industry to monitor and manage customer migration and keep current customers. The marketing literature is unanimous in its conclusion that keeping current clients is preferable than seeking out new ones. As a result, market researchers may evaluate the data and develop a proper model for foreseeing customer attrition and measuring client attitude (Melian et al., 2022). It turns out, however, that most approaches to solving this issue did not take into account the intricate web of relationships between client characteristics and attrition. Vijayaraman and Chellappa (2016) offer a unique Neural Network-based mathematical model for forecasting customer attrition and assessing consumer sentiment. To predict who would leave and how they feel about leaving, the suggested approach uses a pattern-finding technique based on past data. ERNN and JRNN are implemented and their performance is compared in order to forecast the churn rates of mobile phone customers. Both algorithms are put through their paces in an in-depth experimental research utilizing real-time data acquired from mobile phone subscribers of Indian (Sigloch, 2018). The results show that JRNN outperforms ERNN in terms of overall effectiveness. MATLAB is used to verify the experimental analysis. Accurate churn prediction is essential to minimizing wasted resources and maximizing return on investment by focusing retention efforts exclusively on customers who are actively in the process of switching service providers. The experiments have shown that RNNs are the most effective method for this task (Che, 2018).

Companies in a wide variety of sectors have a vested interest in churn prediction, the process of determining which consumers will eventually stop using a service. With the growing quantity of large-scale, diverse data these companies acquire on the characteristics and actions of consumers, new approaches to churn prediction become feasible (Deligiannis & Argyriou, 2020). This article presents a unified analytic strategy for finding churn predictors (Stripling et al., 2018). In this method, supervised learning algorithms are given a feature-set that has been produced using brute-force feature engineering and then modified through feature selection. Khan et al. (2015) used many terabytes of data from a big mobile phone network to develop a system that can predict if a subscriber would churn with 89.4 percent accuracy. The approach identified some apparent, and a few unexpected, early warning indications of churn (Williams et al., 2012).

For survival, mobile phone companies in a competitive market must concentrate on keeping their current clientele. Backiel et al. (2016) look at how adding social network data to churn prediction algorithms might improve precision, timeliness, and profitability. Customer qualities are used in traditional model building, however for prepaid consumers, this information is typically lacking. Another option is to examine up-to-date and comprehensive caller-record graphs. In order to construct the call graph and extract useful characteristics for use in classification models, a method was devised. This method's scalability and utility are shown by applying it to a dataset consisting of over 30 million monthly calls and 1.4 million consumers in the telecoms industry. The findings suggest that by using network characteristics, performance may be improved beyond that of using local features, while keeping a high degree of interpretability and usability (Krajewski et al., 2022).

Mobile telecom carriers face a number of difficult difficulties, but customer turnover is among the most detrimental to both income and subscriber base. To a lesser extent than how well you can anticipate who would go, when you make that prediction is also crucial to the effectiveness of your retention programs (Zhang & Liang, 2011). Models for monthly churn prediction were published in prior works on the topic, with an emphasis on customers' static behavior; even studies that took into account customers' dynamic behavior focused primarily on the monthly level of activity. However, customer behavior may shift over the course of a month, and a customer can exhibit varied behaviors in the days leading up to a churn decision. As a result, examining behavioral variables once a month has a detrimental impact on predictive accuracy since it fails to account for variations in behavior across individual days within a month (Armeli et al., 2010). To solve these issues, Alboukaey et al. (2020) suggest daily churn prediction as opposed to monthly churn prediction, with the former based on the client's dynamic daily activity and the latter on his static monthly behavior. Alboukaey et al. (2020) provide four models for forecasting attrition. The findings demonstrated that daily models greatly outperform their monthly counterparts in identifying potential churners before they occur. Even more so, the LSTM-based model achieves much superior results than the CNN-based model (Li et al., 2020).

In this study, Machado et al. (2019) assess the efficacy of two Gradient Boosting Decision Tree Models, XGBoosting and LightGBM, the latter of which has never been used to predict customer loyalty. Machado et al. (2019) use these techniques to make forecasts about the loyalty of a financial institution's credit card clients. The data set is publicly accessible because to a Kaggle competition. LightGBM outperforms XGBoosting when it comes to predicting customers' loyalty as measured by the root-mean-squared error (RMSE) (Rayhan et al., 2019).

Research on telecom churn prediction is still going strong. Since social network analytics has grown popular, and since earlier benchmarking studies have shown a relatively flat maximum performance effect of predictive modeling methodologies, we examine the relationship between the two variables. Researchers have moved their attention to further extending and studying the universe of relevant features (Bonchi et al., 2011). While most research have shown that increasing the number of characteristics used to make predictions improves accuracy, few have addressed practical concerns such data accessibility and computing cost. Mitrovi et al. (2018) present a novel, re-usable method for discovering optimal feature type combinations using Pareto multi-criteria optimization. The findings provide various takeaways that might be used by professionals in the field as standards or guidelines (Mitrović et al., 2018).

The pace at which a service provider loses clients is known as churn rate (Almana et al., 2014). Users of freemium online games are a good example of this since they may simply stop playing at any time. Game developers are on the lookout for churn detection and prediction technologies that will allow for prompt intervention from upper management (Kim et al., 2017). This may be done by examining game logs. In this piece, Machado et al. (2019) analyze data from the free-to-play game The Settlers Online to draw some conclusions. Based on standard churn and disengagement criteria found in the game analytics literature, four distinct labeling methodologies were used to enable churn detection. Features were calculated from the raw game data to create predictive classifiers. There were a total of eight distinct machine learning methods used, all of which provided binary classifications. Machado et al. (2019) compared the outcomes across all algorithms and classification methods. Area under curve values greater than 0.99 indicated that random forests with sliding windows were the best solution for the dataset, allowing for prediction accuracies of 97%. This was validated by testing on a separate dataset, and in the discussion Machado et al. (2019) provide advice on how to best use feature engineering, labeling methodologies (especially disengagement), and machine learning algorithms to forecast churn. Game developers and researchers who are interested in comparable topics would benefit from the suggestions (Rothmeier et al., 2020).

Worldwide airline profits have been falling due to the global crisis and the rise of low-cost airlines. This study presents a novel method for finding important service attributes (SAs) that might avoid customer churn (Franke & John, 2011). The weights of SAs are extracted using multiple regression, and the most important SAs for predicting customer retention are identified using logistic regression. In conclusion, the importance-performance analysis is carried out to provide direction with feedback (Jairak & Praneetpolgrang, 2013). Additionally, a support vector machine is used to guard client retention's dependability. The following is a summary of the key findings: a) obtaining feedback from passengers about their experiences with airlines, b) identifying which SAs should be prioritized for improvement in order to boost passenger happiness, and c) using the discovered key predictors to foretell client retention (Wang & Fong, 2016).

In this article, Zhao et al. (2018) explain, demonstrate, and analyze supervised machine learning algorithms for forecasting employee turnover rates. Using the following procedures, numerical experiments are conducted: (1) a decision tree method; (2) a random forest method; (3) a gradient boosting trees method; (4) an extreme gradient boosting method; (5) a logistic regression method; (6) a support vector machine method; (7) a neural network method; (8) a linear discriminant analysis method; (9) a Nave Bayes method; (10) a simulated human resources dataset representing small, medium, and large employee populations. To evaluate which of these supervised machine learning algorithms is best successful in predicting employee turnover, we will compare their performance, Zhao et al. (2018) apply a rigorous and in-depth assessment procedure based on statistical measures. In addition, trustworthy recommendations on how to choose, use, and interpret these techniques are offered for analyzing human resources data sets of varied sizes and complexity (Ritchie et al., 2015).

5.2.4.2 Summary

According to the literature review, the majority of the publications offer empirical data-driven experiments, with the majority of them based on data from 2010 and after. The freshness of the data is an essential component in data-driven knowledge discovery, particularly given how telecommunications consumers' behavior has evolved over time (Sun et al., 2018). As a result, using current data decreases the danger of models developed on this data adversely influencing predictions of telecommunications industry demands.

Furthermore, AI algorithms including SVM, DT, ANN, XGboost, and NN are the most used customer churn prediction modeling tools. It will be exciting to see how AI is used to telecom customer churn forecasts in the future.

In comparison to the previous literature, this study develops a unique customer churn prediction model using Fisher's discriminant equation and LR analysis for forecasting telecommunications customer attrition. According to the data, the churn model for telecoms established by regression analysis has a higher prediction accuracy (93.94%) and yields better results. China Mobile, China Unicom, and China Telecom are the three main Chinese telecom carriers whose data was used in this research.

This study benefits the academic community by identifying research gaps, presenting current trends in customer churn predictions, and assisting in the understanding of how to design accurate and successful marketing plans. Today’s market is getting more competitive (Chadha & Kapoor, 2009). Telecom firms must make key choices and build effective retention strategies to prevent customer churn, since it is far less costly to keep current customers (Kim et al., 2020). This study will assist telecom businesses in accurately predicting customer attrition and taking preventative measures, hence improving profitability.

Chapter 6 Conclusions

6.1 Contributions

Telecom customer churn is a central issue for telecom companies since it decreases profits (Bach et al., 2021). Furthermore, preventing customer churn is imperative. The global telecom industry is becoming more saturated, and companies are increasingly struggling to retain customers (Chadha & Kapoor, 2009). Currently, most companies invest heavily in marketing to attract new customers. However, keeping existing customers is cheaper than acquiring new customers (Kim et al., 2020). Thus, it is becoming more critical and a significant concern for telecommunication companies to prevent customer churn. This research aims to improve customer targeting using customer segmentation approaches based on data science. To achieve this aim, this research was divided into two stages. In the first stage, this study presents a literature review on customer churn prediction based on 40 relevant articles published between 2010 and June 2020. In a second stage, data were collected from three major Chinese telecom companies to create a churn prediction model to predict telecom client churn through customer segmentation using Fisher discriminant equations and LR analysis.

Moreover, the results of this study will give telecom managers the ability to accurately predict customer behavior and loss and optimize their strategies to improve customer retention rates. Meanwhile, the findings will help companies reduce costs and optimize their budgets. Furthermore, it will be possible for telecom managers to improve customer targeting through this study's results and increase the profits of telecom companies. From the scientific perspective, it should be highlighted that this thesis resulted in two published articles:

- **Zhang, T.**, & Moro, S. (2021). Research trends in customer churn prediction: A data mining approach. In *World Conference on Information Systems and Technologies* (pp. 227-237). Published in the Springer Book Series "Trends and Applications in Information Systems and Technologies. Advances in Intelligent Systems and Computing", ranked in **Quartile 3 in the Scopus** database. Citations received (as of 2022-09-21): Google Scholar: 2; Scopus: 0; Web of Science: 0.

- **Zhang, T.**, Moro, S., & Ramos, R. F. (2022). A Data-Driven Approach to Improve Customer Churn Prediction Based on Telecom Customer Segmentation. *Future Internet*, 14(3), 94. Ranked in **Quartile 2 in the Scopus** database.

Citations received (as of 2022-09-21): Google Scholar: 4; Scopus: 3; Web of Science: 1.

Table 6-1: Visibility information for the journal article.

Abstract Views	1262
Full-Text Views	1627
Download Times	1103
Citations *	3

* Source: Crossref, 29 September 2022

6.2 Theoretical implications

6.2.1 Theoretical review and implications of customer churn prediction

The objective of customer churn prediction is to identify subscribers who are likely to cancel their service subscription (Dalvi et al., 2016). The once-rapidly developing mobile communications industry has lately stagnated and become saturated, with intense competition (Bhattacharyya & Dash, 2020). Consequently, telecom companies are changing their focus from attracting new customers to keeping their current clients (Liu et al., 2011). Knowing which of your customers are most likely to switch to a competitor is consequently highly advantageous (Fuchs & Schreier, 2011). Information extracted from the telecommunications industry may be utilized to better understand and address the reasons of customer churn (Ullah et al., 2019). Dalvi et al. (2016) suggest using data mining and machine learning methods, namely LR and DT, to create a churn prediction model for the telecoms business. Comparing the efficacy of several algorithms on the present dataset.

Due to increased rivalry and market saturation, the cost of acquiring new consumers has risen dramatically over the last several years, while the cost of keeping existing ones has remained relatively constant. Therefore, retaining existing customers is a top priority for many businesses. Keeping customers happy calls for churn control, and that in turn necessitates a reliable churn prediction model (Dälken, 2014). Several techniques, including as logistic regression, neural networks, evolutionary algorithms, and decision trees, have been used to the churn prediction issue (Tsai & Lu, 2010). This research presents a hybrid technique that combines data fusion and feature extraction to give more accurate forecasts of customer retention (Thakkar & Chaudhari, 2021). Then, two algorithms were trained with varied feature sizes and tested on test data using the data preparation and feature selection processes described before. Finally, weighted voting was used to aggregate the results from the various classifiers. Results from applying the approach to actual data from a telecommunications business demonstrated its efficacy (Jamalian & Foukerdi, 2018).

There is intense rivalry among service providers in the modern telecommunications industry for the attention of new clients. As a result of clients leaving for other services, several companies have seen a decline in profits. One of the most common approaches used to maintain client loyalty over time is a dedicated customer retention campaign (Jurisic & Azevedo, 2011). However, this comes at a significant price, therefore businesses should instead concentrate on recognizing early on whether clients are at risk of churning. Since there hasn't been a lot of study into customer churn using machine learning methods, Ismail et al. (2015) will investigate the possibility of employing an artificial neural network to enhance churn prediction. The outcome shown that neural networks are superior than statistical models for making predictions (91.28% accuracy). The findings imply that a neural network learning algorithm may be a viable alternative to conventional statistical forecasting approaches for predicting client attrition (Ismail et al., 2015).

Customer retention should be a major goal for any company that depends on repeat business. To optimize the service provider's profit, it is vital to construct accurate churn prediction models (Ghorban & Tahernejad, 2012). Rodan and Faris (2015) offer a method for forecasting customer churn in the telecommunications industry by combining an ESN with a SVM training algorithm. Both a widely-used publicly-available dataset and a smaller, locally-obtained dataset are used to train and evaluate the suggested method. The experimental results reveal that ESN with SVM readout outperforms other well-known machine learning models for forecasting customer churn (Rodan & Faris, 2015).

Widely utilized are techniques for predicting customer attrition that combine excellent predictive accuracy with understandable explanations, such as decision trees and logistic regression (Dalvi et al., 2016). Despite these advantages, logistic regression struggles with interaction effects across variables, while decision trees struggle with linear connections (Fife & Onofrio, 2022). Therefore, we introduce the logit leaf model (LLM) as an unique hybrid approach for enhanced data categorization (Fathian et al., 2016). The LLM is predicated on the idea that discrete models created on subsets of data rather than the whole dataset result in greater prediction performance without compromising the interpretability of the models produced in the leaves (Narteni et al., 2022). First, client segments are selected using decision criteria, followed by the development of a model for each of these subbranches (Hailu, 2021). The predictive accuracy and interpretability of this innovative hybrid method are compared to those of DT, LR, RF, and logistic model trees (Lee & Jun, 2018). LLM outperforms its component techniques, LR and DT, and is comparable to more advanced ensemble methods such as RF and logistic model trees (De et al., 2018). This paper addresses comprehensibility using a case study for which the LLM has many significant benefits over decision trees or logistic regression (Caigny et al., 2018).

Project managers have a lot riding on customer retention, especially in the increasingly competitive and crowded telecommunications industry. Due to the high cost of customer acquisition, churn prediction has become a key part of strategic decision making and planning in the telecommunications business (Wilson et al., 2018). If you want to keep the clients that are likely to leave your service, it's crucial that you anticipate their behavior and take steps to keep them as customers. This study is a further attempt to establish churn prediction rules by using rough set theory, a rule-based approach to decision making (Rodan & Faris, 2015). The efficiency of four unique algorithms was empirically evaluated. The performance of the approach based on a genetic algorithm for rough set classification is the most satisfying. In addition, the results of testing the suggested method on a publically accessible dataset demonstrate its efficacy in identifying and predicting consumers likely to churn, providing invaluable insight for business leaders (Amin et al., 2015).

Increases in strategic and analytic capabilities are necessary to keep up with the explosive expansion of telecom data and the severe rivalry among telecom providers for client retention. Classifying and predicting prospective churners from a big customer set allows for the creation of lucrative and tailored retention initiatives, which are essential for maintaining a long-term connection with valuable clients (Du et al., 2021). Several churn prediction models have been proposed in the past in an effort to precisely identify customers who are likely to churn in order to achieve the aforementioned objective (Richter et al., 2010). However, the limitations of these previously stated models make it impossible to apply them directly for exact prediction (Rudin, 2019). During model creation, the majority of earlier work's feature selection algorithms neglected the information-rich variables provided in the call details record (Azeem et al., 2017). As a second point, Rodan and Faris (2015) relied solely on statistical techniques to choose which details to highlight. Even though statistics has been successfully applied to many different fields, it is important to remember that statistics on its own is not enough to provide accurate findings; domain expertise is also necessary. Thirdly, the benchmark datasets that have been used to evaluate the prior models do not accurately reflect the noise and vast number of missing values seen in real-world telecom data. Fourth, the True Positive (TP) rate, which stresses a model's ability to effectively classify the proportion of churners vs non-churners, was disregarded by the previously used evaluation measures (Valluri et al., 2022). Classifiers employed in older models primarily disregarded fuzzy classification approaches, which perform rather well with noisy data sets. Fuzzy classifiers have been compared against a number of common classifiers, such as Neural Network, Linear regression, C4.5, SVM, AdaBoost, Gradient Boosting, and Random Forest, to establish their superiority (Azeem et al., 2017).

Personality and character traits reveal shoppers' purchasing tendencies. These traits and characteristics may be found in a wide range of people. Learning, beliefs, attitude, Culture, and social pressures are all factors in character, as are factors like quality, motivation, occupation, and income, perception, psychology, personality, reference groups, and demographics (Kosinski et al., 2013). These days, data mining is often utilized to examine consumer buying behaviors via the use of several algorithms and techniques. Slowly but surely, the data mining industry has grown and expanded into a wide variety of different fields. Information about a customer's day-to-day habits, such as how much time and energy they invest in making purchasing decisions, is recorded as bits of data in a database (Jukić et al., 2015). The most frequently purchased goods and the total amount bought are also taken into account. Customers' consent is not required for the collection of this information. Maheswari and Priya (2017) utilize the dataset to examine and classify customers according to their buying habits. SVM algorithm is used to do the classification. Experimental findings are evaluated, demonstrating that the suggested technique provides a more thorough understanding of a customer's actions (Maheswari & Priya, 2017).

The telecommunications business has a significant difficulty in retaining its customers. The ability to accurately estimate customer turnover may be critical to a company's bottom line, since a thorough investigation into customer attrition can provide useful insights for retaining profitable clientele (Nitzan & Libai, 2011). Mashraie et al. (2020) use actual data from a partner firm to evaluate the efficacy of several churn prediction algorithms. Logistic regression, SVMs, RFs, and DTs are all available as prediction models. In addition, the push-pull-mooring (PPM) architecture is used to examine the impact of features on customer retention based on the push, pull, and mooring angles (Rhoudri & Benazzou, 2021). PPM analysis is conducted using partial least squares (PLS) regression (Perera et al., 2021). Additionally, both churners and non-churners' behaviors are examined. Logistic regression was determined to be the most accurate predictor of customer turnover (Mashraie et al., 2020).

Many businesses worry greatly about losing customers. The telecom industry is especially prone to this pattern (Weiss, 2010). To prevent consumers from departing, telecom businesses need a proactive approach (Ahmad et al., 2019). Best feature selection for developing models is not implemented in the current works. Pamina et al. (2019) add to the creation of a churn prediction model, which may be used by telecom companies to anticipate and prepare for the loss of consumers. Focus is placed on the recall assessment metric that provides a workable answer to an urgent business issue. Multiple measures, including as ROC, F-score and Accuracy, are used to evaluate the effectiveness of the model. The approach is to provide the best possible recall value, which may be applied to practical business issues. This research found that DT model 3 performed the best in the experiments (Pamina et al., 2019).

Predicting which customers would abandon a telecom service is challenging (Idris et al., 2012). This study introduces a novel customer churn prediction system for telecom named FW-ECP to assist with this challenging task (Idris & Khan, 2017). The FW-innovation ECP's is that it is able to use the accumulative knowledge of an ensemble classifier constructed from many base classifiers to choose features in a unified, but innovative, manner. As part of the filtering process, Idris and Khan (2017) use undersampling based on Particle Swarm Optimization and feature selection based on multi-resolution multi-rank (mRMR) to mitigate the impact of uneven class distribution and high dimensionality. As part of the Wrapper step, it uses a Genetic Algorithm to get rid of any remaining unnecessary or duplicate characteristics (Moslehi & Haeri, 2020). Once the new feature space has been identified, it is exploited using methods such as RF and SVM. The higher prediction outcomes are from the method of FW-superior ECP. Orange and Cell2Cell datasets' initial feature spaces have been reduced from 260D and 76D, respectively, to 24D and 18D. This study reveals that the AUCs for the Orange and Cell2Cell datasets are 0.85 and 0.82, respectively, using FW-ECP (Idris & Khan, 2017).

In recent years, the prominence of the knowledge-based economy has increased, particularly in online shopping applications that record users' purchases and comments (Zolfaghar & Aghaie, 2012). Logs might be processed using machine learning techniques to glean hidden insights. The information is used to help industries and organizations learn more about customer behavior, as well as identify new possibilities and risks (Sivarajah et al., 2017). Researchers stand to benefit from a better understanding of how to anticipate electronic consumer behavior. Even though Idris and Khan (2017) had already experienced internet purchasing before to the coronavirus pandemic, the number of online purchases during the outbreak surged considerably. Due to the fast dissemination of COVID-19, it is important to avoid being public places and remain inside (Bahety et al., 2021). Online shoppers' actions are directly influenced by these problems. Safara (2020) provides a prediction model that uses machine learning methods to anticipate the activities of consumers. The model constructed utilizing DT ensembles and Bagging fared the best at forecasting consumer behavior (95.3% accuracy), according to the data (Miguéis et al., 2018). To further investigate the factors that are most influential on the number of web-based acquisitions made during the epidemic, a correlation analysis is conducted (Safara, 2020).

6.2.2 Summary

Telecom companies must understand the reasons for customer churn since customer churn is directly related to the companies' profit. The process, combined with the big data accumulation in the telecom industry and the increasingly mature data mining technology, motivates the development and application of a customer churn model to predict customer behavior. Therefore, the telecom company can effectively predict the churn of customers and avoid customer churn. The literature review results showed that the most widely used data mining techniques are DT, SVM, and LR.

There is very little knowledge about how telecom customers' opinions regarding the services provided by their telecom company impact customer churn. We aimed to cover this research gap using a Fisher discriminant analysis and a LR analysis of telecom customer churn related to diverse factors. Moreover, the discriminant function and LR analysis predict telecom customer churn (Alzubaidi & Shamery, 2020). In this study, through a Wilks' lambda discriminant test, we concluded that the discriminant equation is valid and can explain the reasons for churn. Furthermore, the accuracy-test proved the LR equation valid and can explain the reasons for churn. Serrano et al. (2013) highlighted that previous telecom customer churn studies have mainly applied factor analysis, cluster analysis, and other methods, while telecom customer churn studies conducted using Fisher discriminant analysis and LR analysis remain scarce, even in top journals. This new investigation should solve this problem.

6.3 Managerial implications

6.3.1 Managerial application and implications of customer churn prediction

There is intense rivalry among service providers in the modern telecommunications industry for the attention of new clients. As a result of clients leaving for other services, several companies have seen a decline in profits. One of the most common approaches used to maintain client loyalty over time is a dedicated customer retention campaign (Jurisic & Azevedo, 2011). However, this comes at a significant price, therefore businesses should instead concentrate on recognizing early on whether clients are at risk of churning. Since there hasn't been a lot of study into customer churn using machine learning methods, this study will investigate the possibility of employing an artificial neural network to enhance churn prediction. The outcome shown that neural networks are superior than statistical models for making predictions (91.28% accuracy) (Gao et al., 2018). The findings imply that a neural network learning algorithm may be a viable alternative to conventional statistical forecasting approaches for predicting client attrition (Ismail et al., 2015).

For practitioners and academics in the telecommunications industry to prosper in the face of strong competition and retain the present loyal customers, it is crucial that they be able to forecast potential churn customers using predictive modeling techniques (Khan et al., 2019). Successful prediction models may be used to pinpoint the clientele most likely to remain loyal. The number of unsatisfied customers contemplating leaving the firm might be reduced by allocating resources specifically to retaining them. In this research, the authors suggest an artificial neural network method for identifying potential churners (Khan et al., 2019). This model can process a large number of features from a telecom company's data collection, including demographic information, billing details, and use trends (Bahrami et al., 2020).

In recent years, businesses, especially in the telecommunications industry, have focused more on predicting whether or not a client would depart (Ascarza et al., 2018). The major focus is on churn in the telecommunications industry, with the goal being an exact estimation of the survival and hazard functions for customers over time. The second objective is to identify customers who are about to leave and forecast how long they will remain (Zhang & Chang, 2021). Ahmed and Linen (2017) examine churn prediction methods in order to discover churn conduct and verify churn causes. This article presents the major churn prediction methods presently in use and illustrates that hybrid models, as opposed to individual algorithms, provide the most accurate churn prediction. The purpose of this research is to better understand why customers depart (Vo et al., 2021). The telecom sector may then utilize this information to better satisfy the demands of high-risk consumers and reverse the churn decision (Adebiyi et al., 2016).

Because of increased competition on a worldwide scale, businesses in all sectors are more worried about losing customers. The telecoms industry ranks highest, with a turnover rate of 30 percent (Yu et al., 2019). By using predictive algorithms, telecoms companies can keep track of their consumers at danger of leaving. When there is a significant gap in the sample sizes of various groups within a dataset, there is unequal data (Leevy et al., 2018). When dealing with the class-imbalance issue (CIP), over/under sampling is the most prevalent method (Duarte et al., 2021). This paper examines six popular sampling methods and evaluate their relative merits. The empirical results demonstrate that MTDF and rules-generation based on genetic algorithms had the best overall predictive performance (Amin et al., 2016).

A project manager's ability to foresee how clients would act is crucial. Data-driven sectors, such as the telecommunications industry, may use data mining methods to glean insights on their customers' likely future actions (Vassakis et al., 2018). In addition to helping with classification accuracy, feature elimination may have a significant impact on the dataset's computing cost and complexity (Salo et al., 2019). The proposed study applies the Minimum Redundancy Maximum Relevance (mRMR) technique for feature extraction, which does more than just choose the optimal subset of features; it also reduces the features space (Khalid et al., 2014). Decision-makers and researchers may utilize the results as proof of the predictive efficacy of oversampling methodologies and rules-generation algorithms, enabling them to choose the most successful strategy (Amin et al., 2015).

As market rivalry increases, the importance of customer churn control as a source of business advantage is growing (Bressler, 2012). Existing churn prediction algorithms, however, do not perform well when dealing with huge data in the business (Ullah et al., 2019). Moreover, decision makers are often confronted with inaccurate operations management (Bi et al., 2016). In response to these issues, the semantic-driven subtractive clustering method (SDSCM) is presented as a novel clustering technique (Bi et al., 2016). Experimental findings reveal that SDSCM has more semantic clustering power (Cui et al., 2015). The Hadoop MapReduce framework is then used to create a parallel SDSCM algorithm. In the example study, the parallel SDSCM algorithm suggested has a faster execution speed than the other approaches. In addition, this study gives marketing ideas in line with clustering findings and simulates a reduced marketing activity to maximize profits (Bi et al., 2016).

Keeping clients happy in the modern telecommunications sector is a cutthroat business. So, in order to stay competitive, you need a reliable churn prediction system to spot your clients just before they decide to go (Mahapatra & Singh, 2021). Customers who are predicted to leave may be kept around by acting on the specific reasons why they are likely to do so. As a result, a churn prediction system's duty has expanded to include the deciphering of customers' churning behavior in addition to the prediction of churners (Slof et al., 2021). Idris and Iftikhar (2019) integrate the searching capabilities of genetic programming (GP) with the classification capabilities of AdaBoost in order to construct a high-performance churn prediction system with enhanced churn detection skills. As part of this procedure, AdaBoost-based learning is utilized to identify and compare the frequency with which certain features arise in the assessments of several GP expressions (Kamarudin et al., 2017). Particle swarm optimization (PSO)-based undersampling is used to provide an equitable training set to the GP-AdaBoost-based prediction system (Almasi & Saniee, 2018). The churn prediction system (ChP-GPAB) is the result of integrating an undersampling technique based on particle swarm optimization with GP-AdBoost. This combination allows for improved churner learning and the identification of causal elements in customer churn. The suggested ChP-GPAB system is tested, evaluated, and compared using two industry-standard telecom data sets.The results reveal that the proposed ChP-GPAB system detects churning causes (Idris & Iftikhar, 2019).

6.3.2 Summary

Following the outcomes of the literature research, this work creatively constructs discriminant and LR models to predict telecom client attrition using customer segmentation data from three large Chinese telecom businesses. The results showed that LR analysis is efficient in building the telecom customer churn model. The model’s prediction accuracy has achieved 93.94%, confirming that the conclusions of the literature review are correct, which provides an efficient and reliable method for telecom companies to predict customer churn.

Previous literature has never conducted an in-depth study of the specific factors influencing subscriber churn in China, specifically, the relationship between monthly fixed cost, local costs, the service quality of Internet, fixed-line and CRBT products, the call time for long-distant calls, the numbers of SMS and MMS in the telecom package and subscriber churn. This paper conducts a relevant study that provides operators with a means to prevent customer. This paper also provides recommendations for operators to prevent customer churn.

According to the results of this paper, the recommendations are for telecom companies to decrease their monthly fixed costs and local costs to increase the possibility of retaining their telecom customers. Additionally, the managers of telecom companies have already realized the value and importance of improving the service quality of the Internet, fixed-line, and CRBT products, as well as the call time for long-distant calls and the numbers of SMS and MMS in the telecom package, which has previously been proven to have a positive influence on telecom customers retention.

6.4 Limitations and future research

Many investigators are dedicating themselves to work on the topic of how to predict telecom customer churn. It is also becoming one of the most important problems for telecom firms to resolve, as they must identify the causes of client churn and work to prevent it (Ullah et al., 2019). In the future, more updated and related literature reviews should be conducted to keep up with the latest and relevant telecom churn prediction methods and further refine our research.

The dataset includes the information for 4126 clients from 2007 to 2018. However, it has been nearly four years since then. Because of the COVID-19 pandemic, the telecom market and customer consumption habits may differ significantly. Therefore, more current data should be gathered to improve the model's accuracy further and move the model more in line with the current market situation. Furthermore, repeated data testing approach can further improve the model.

Moreover, data were collected from three operators. Data from other operators may increase the reliability of the model. Finally, additional variables could be applied to improve its predictability.

References

Abbasimehr, H., Setak, M., & Tarokh, M. J. (2011). A neuro-fuzzy classifier for customer churn prediction. International Journal of Computer Applications, 19(8), 35-41.

Abdou, H. A., & Pointon, J. (2011). Credit scoring, statistical techniques and evaluation criteria: a review of the literature. Intelligent systems in accounting, finance and management, 18(2-3), 59-88.

Abdullah, M., Ali, N., Hussain, S. A., Aslam, A. B., & Javid, M. A. (2021). Measuring changes in travel behavior pattern due to COVID-19 in a developing country: A case study of Pakistan. Transport Policy, 108, 21-33.

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE access, 6, 52138-52160.

Addetia, A., Crawford, K. H., Dingens, A., Zhu, H., Roychoudhury, P., Huang, M. L., ... & Greninger, A. L. (2020). Neutralizing antibodies correlate with protection from SARS-CoV-2 in humans during a fishery vessel outbreak with a high attack rate. Journal of clinical microbiology, 58(11), e02107-20.

Adebiyi, S. O., Oyatoye, E. O., & Amole, B. B. (2016). Improved customer churn and retention decision management using operations research approach. EMAJ: Emerging Markets Journal, 6(2), 12-21.

Adwan, O., Faris, H., Jaradat, K., Harfoushi, O., & Ghatasheh, N. (2014). Predicting customer churn in telecom industry using multilayer preceptron neural networks: Modeling and analysis. Life Science Journal, 11(3), 75-81.

Aftab, S., Ahmad, M., Hameed, N., Bashir, M. S., Ali, I., & Nawaz, Z. (2018). Rainfall prediction using data mining techniques: A systematic literature review. International journal of advanced computer science and applications, 9(5).

Agarwal, S., Farid, H., Gu, Y., He, M., Nagano, K., & Li, H. (2019, June). Protecting World Leaders Against Deep Fakes. In CVPR workshops (Vol. 1, p. 38).

Agrawal, S., Das, A., Gaikwad, A., & Dhage, S. (2018, July). Customer churn prediction modelling based on behavioural patterns analysis using deep learning. In 2018 International conference on smart computing and electronic enterprise (ICSCEE) (pp. 1-6). IEEE.

Ahmad, A. K., Jafar, A., & Aljoumaa, K. (2019). Customer churn prediction in telecom using machine learning in big data platform. Journal of Big Data, 6(1), 1-24.

Ahmad, U. (2011). What makes customers brand loyal: A study on telecommunication sector of Pakistan. International journal of business and social science, 2(14).

Ahmed, A., & Linen, D. M. (2017, January). A review and analysis of churn prediction methods for customer retention in telecom industries. In 2017 4th International Conference on Advanced Computing and Communication Systems (ICACCS) (pp. 1-7). IEEE.

Ahmed, M., Afzal, H., Majeed, A., & Khan, B. (2017). A survey of evolution in predictive models and impacting factors in customer churn. Advances in Data Science and Adaptive Analysis, 9(03), 1750007.

Akmal, M. (2017). Factor Causing Customer Churn: A Qualitative Explanation of Customer Churns In Pakistan Telecom Industry (Doctoral dissertation, MS Thesis], March).

Al Amin, M., Jewel, M. M. H., & Fouji, M. H. (2019). Influencing factors of customer attitude towards SMS marketing-a case of mobile telecommunication industry in bangladesh. Jagannath Univ. J. Bus. Stud., 1, 65-78.

Alam, N., & Rubel, A. K. (2014). Impacts of corporate social responsibility on customer satisfaction in telecom industry of Bangladesh. ABC Journal Of Advanced Research, 3(2), 93-104.

Alboukaey, N., Joukhadar, A., & Ghneim, N. (2020). Dynamic behavior based churn prediction in mobile telecom. Expert Systems with Applications, 162, 113779.

Al-Debei, M. M., Akroush, M. N., & Ashouri, M. I. (2015). Consumer attitudes towards online shopping: The effects of trust, perceived benefits, and perceived web quality. Internet Research.

Aldeen, Y. A. A. S., Salleh, M., & Razzaque, M. A. (2015). A comprehensive review on privacy preserving data mining. SpringerPlus, 4(1), 1-36.

Aldowah, H., Al-Samarraie, H., & Fauzy, W. M. (2019). Educational data mining and learning analytics for 21st century higher education: A review and synthesis. Telematics and Informatics, 37, 13-49.

Aljumah, A. A., Ahamad, M. G., & Siddiqui, M. K. (2013). Application of data mining: Diabetes health care in young and old patients. Journal of King Saud University-Computer and Information Sciences, 25(2), 127-136.

Alkhatib, K., & Abualigah, S. (2020, April). Predictive model for cutting customers migration from banks: based on machine learning classification algorithms. In 2020 11th International Conference on Information and Communication Systems (ICICS) (pp. 303-307). IEEE.

Almana, A. M., Aksoy, M. S., & Alzahrani, R. (2014). A survey on data mining techniques in customer churn analysis for telecom industry. International Journal of Engineering Research and Applications, 4(5), 165-171.

Al-Mashraie, M., Chung, S. H., & Jeon, H. W. (2020). Customer switching behavior analysis in the telecommunication industry via push-pull-mooring framework: A machine learning approach. Computers & Industrial Engineering, 144, 106476.

Almasi, M., & Saniee Abadeh, M. (2018). A new MapReduce associative classifier based on a new storage format for large-scale imbalanced data. Cluster Computing, 21(4), 1821-1847.

Al-Mohair, H. K., Saleh, J. M., & Suandi, S. A. (2015). Hybrid human skin detection using neural network and K-means clustering technique. Applied Soft Computing, 33, 337-347.

Almuqren, L. (2021). Twitter Analysis to Predict the Satisfaction of Saudi Telecommunication Companies’ Customers (Doctoral dissertation, Durham University).

Alwis, P. K. D. N. M., Kumara, B. T. G. S., & Hapuarachchi, H. A. C. S. (2018). Customer Churn Analysis and Prediction in Telecommunication for Decision Making.

Alzubaidi, A. M. N., & Al-Shamery, E. S. (2020). Projection pursuit Random Forest using discriminant feature analysis model for churners prediction in telecom industry. International Journal of Electrical & Computer Engineering (2088-8708), 10(2).

Amado, A., Cortez, P., Rita, P., & Moro, S. (2018). Research trends on Big Data in Marketing: A text mining and topic modeling based literature analysis. European Research on Management and Business Economics, 24(1), 1-7.

Amin, A., Al-Obeidat, F., Shah, B., Adnan, A., Loo, J., & Anwar, S. (2019). Customer churn prediction in telecommunication industry using data certainty. Journal of Business Research, 94, 290-301.

Amin, A., Anwar, S., Adnan, A., Nawaz, M., Alawfi, K., Hussain, A., & Huang, K. (2017). Customer churn prediction in the telecommunication sector using a rough set approach. Neurocomputing, 237, 242-254.

Amin, A., Anwar, S., Adnan, A., Nawaz, M., Howard, N., Qadir, J., ... & Hussain, A. (2016). Comparing oversampling techniques to handle the class imbalance problem: A customer churn prediction case study. IEEE Access, 4, 7940-7957.

Amin, A., Khan, C., Ali, I., & Anwar, S. (2014, November). Customer churn prediction in telecommunication industry: with and without counter-example. In Mexican international conference on artificial intelligence (pp. 206-218). Springer, Cham.

Amin, A., Khan, C., Ali, I., & Anwar, S. (2014). Customer churn prediction in telecommunication industry: with and without counter-example. In Mexican international conference on artificial intelligence (pp. 206-218). Springer, Cham.

Amin, A., Rahim, F., Ali, I., Khan, C., & Anwar, S. (2015). A comparison of two oversampling techniques (smote vs mtdf) for handling class imbalance problem: A case study of customer churn prediction. In New contributions in information systems and technologies (pp. 215-225). Springer, Cham.

Amin, A., Rahim, F., Ramzan, M., & Anwar, S. (2015, May). A prudent based approach for customer churn prediction. In International conference: Beyond databases, architectures and structures (pp. 320-332). Springer, Cham.

Amin, A., Rahim, F., Ramzan, M., & Anwar, S. (2015). A prudent based approach for customer churn prediction. In International conference: Beyond databases, architectures and structures (pp. 320-332). Springer, Cham.

Amin, A., Shah, B., Khattak, A. M., Moreira, F. J. L., Ali, G., Rocha, A., & Anwar, S. (2019). Cross-company customer churn prediction in telecommunication: A comparison of data transformation methods. International Journal of Information Management, 46, 304-319.

Amin, A., Shehzad, S., Khan, C., Ali, I., & Anwar, S. (2015). Churn prediction in telecommunication industry using rough set approach. In New trends in computational collective intelligence (pp. 83-95). Springer, Cham.

Amin, D. M., & Garg, A. (2019). Performance analysis of data mining algorithms. Journal of Computational and Theoretical Nanoscience, 16(9), 3849-3853.

Anaam, E. A., Magableh, M. N. Y., Hamdi, M., Hmoud, A. Y. R., & Alshalabi, H. (2021). Data mining techniques with electronic customer relationship management for telecommunication company. Amazonia Investiga, 10(48), 288-304.

Andrews, J. G., Claussen, H., Dohler, M., Rangan, S., & Reed, M. C. (2012). Femtocells: Past, present, and future. IEEE Journal on Selected Areas in communications, 30(3), 497-508.

Angelova, B., & Zekiri, J. (2011). Measuring customer satisfaction with service quality using American Customer Satisfaction Model (ACSI Model). International journal of academic research in business and social sciences, 1(3), 232.

Anoopkumar, M., & Rahman, A. M. Z. (2016, March). A Review on Data Mining techniques and factors used in Educational Data Mining to predict student amelioration. In 2016 International Conference on Data Mining and Advanced Computing (SAPIENCE) (pp. 122-133). IEEE.

Arbenina, M. (2021). Spatial data mining and machine learning techniques to understand cross-border relocations of headquarters in Europe.

Arisholm, E., Briand, L. C., & Johannessen, E. B. (2010). A systematic and comprehensive investigation of methods to build and evaluate fault prediction models. Journal of Systems and Software, 83(1), 2-17.

Armeli, S., Conner, T. S., Cullum, J., & Tennen, H. (2010). A longitudinal analysis of drinking motives moderating the negative affect-drinking association among college students. Psychology of Addictive Behaviors, 24(1), 38.

Ascarza, E. (2018). Retention futility: Targeting high-risk customers might be ineffective. Journal of Marketing Research, 55(1), 80-98.

Ascarza, E., Neslin, S. A., Netzer, O., Anderson, Z., Fader, P. S., Gupta, S., ... & Schrift, R. (2018). In pursuit of enhanced customer retention management: Review, key issues, and future directions. Customer Needs and Solutions, 5(1), 65-81.

Asimakopoulos, G., & Whalley, J. (2017). Market leadership, technological progress and relative performance in the mobile telecommunications industry. Technological forecasting and social change, 123, 57-67.

Au, W. H., Chan, K. C., & Yao, X. (2003). A novel evolutionary data mining algorithm with applications to churn prediction. IEEE transactions on evolutionary computation, 7(6), 532-545.

Aydin, S., & Özer, G. (2005). The analysis of antecedents of customer loyalty in the Turkish mobile telecommunication market. European Journal of marketing.

Ayele, W. Y. (2020). Adapting CRISP-DM for idea mining: a data mining process for generating ideas using a textual dataset. International Journal of Advanced Computer Sciences and Applications, 11(6), 20-32.

Azeem, M., Usman, M., & Fong, A. C. M. (2017). A churn prediction model for prepaid customers in telecom using fuzzy classifiers. telecommunication Systems, 66(4), 603-614.

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K. R., & Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PloS one, 10(7), e0130140.

Backiel, A., Baesens, B., & Claeskens, G. (2016). Predicting time-to-churn of prepaid mobile telephone customers using social network analysis. Journal of the Operational Research Society, 67(9), 1135-1145.

Badawy, M., Abd El-Aziz, A. A., Idress, A. M., Hefny, H., & Hossam, S. (2016). A survey on exploring key performance indicators. Future Computing and Informatics Journal, 1(1-2), 47-52.

Baek, C., & Doleck, T. (2021). Educational data mining versus learning analytics: A review of publications from 2015 to 2019. Interactive Learning Environments, 1-23.

Baesens, B., Mues, C., Martens, D., & Vanthienen, J. (2009). 50 years of data mining and OR: upcoming trends and challenges. Journal of the Operational Research Society, 60(sup1), S16-S23.

Bahety, G., Bauhoff, S., Patel, D., & Potter, J. (2021). Texts don’t nudge: An adaptive trial to prevent the spread of COVID-19 in India. Journal of development economics, 153, 102747.

Bahrami, M., Bozkaya, B., & Balcisoy, S. (2020). Using behavioral analytics to predict customer invoice payment. Big data, 8(1), 25-37.

Baker, R. S., Martin, T., & Rossi, L. M. (2016). Educational data mining and learning analytics. The Wiley handbook of cognition and assessment: Frameworks, methodologies, and applications, 379-396.

Ballings, M., & Van den Poel, D. (2012). Customer event history for churn prediction: How long is long enough?. Expert Systems with Applications, 39(18), 13517-13522.

Banu, M. N., & Gomathy, B. (2013). Disease predicting system using data mining techniques. International Journal of Technical Research and Applications, 1(5), 41-45.

Barberis, N. C. (2013). Thirty years of prospect theory in economics: A review and assessment. Journal of Economic Perspectives, 27(1), 173-96.

Bartholomew, D. J., Knott, M., & Moustaki, I. (2011). Latent variable models and factor analysis: A unified approach. John Wiley & Sons.

Barton, K., Dielman, T. E., & Cattell, R. B. (1973). An item factor analysis of intrafamilial attitudes of parents. The Journal of Social Psychology, 90(1), 67-72.

Bhattacharyya, J., & Dash, M. K. (2020). Investigation of customer churn insights and intelligence from social media: a netnographic research. Online Information Review.

Bhattacharyya, J., & Dash, M. K. (2021). What do we know about customer churn behaviour in the telecommunication industry? A bibliometric analysis of research trends, 1985–2019. FIIB Business Review, 23197145211062687.

Bhattacharyya, J., & Dash, M. K. (2022). What do we know about customer churn behaviour in the telecommunication industry? A bibliometric analysis of research trends, 1985–2019. FIIB Business Review, 11(3), 280-302.

Bhattacharyya, S., Jha, S., Tharakunnel, K., & Westland, J. C. (2011). Data mining for credit card fraud: A comparative study. Decision support systems, 50(3), 602-613.

Bi, W., Cai, M., Liu, M., & Li, G. (2016). A big data clustering algorithm for mitigating the risk of customer churn. IEEE Transactions on Industrial Informatics, 12(3), 1270-1281.

Bokulich, N. A., Kaehler, B. D., Rideout, J. R., Dillon, M., Bolyen, E., Knight, R., ... & Gregory Caporaso, J. (2018). Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2’ s q2-feature-classifier plugin. Microbiome, 6(1), 1-17.

Bonchi, F., Castillo, C., Gionis, A., & Jaimes, A. (2011). Social network analysis and mining for business applications. ACM Transactions on Intelligent Systems and Technology (TIST), 2(3), 1-37.

Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. Information, communication & society, 15(5), 662-679.

Brandusoiu, I., & Toderean, G. (2013). Churn prediction in the telecommunications sector using support vector machines. Margin, 1, x1.

Breed, I. (2019). Hierarchical forecasting of engineering demand at KLM Engineering & Maintenance (Master's thesis, University of Twente).

Bressler, M. S. (2012). How small businesses master the art of competition through superior competitive advantage. Journal of Management and Marketing Research, 11(1), 1-12.

Brânduşoiu, I., Toderean, G., & Beleiu, H. (2016, June). Methods for churn prediction in the pre-paid mobile telecommunications industry. In 2016 International conference on communications (COMM) (pp. 97-100). IEEE.

Buenaño‐Fernandez, D., Villegas‐CH, W., & Luján‐Mora, S. (2019). The use of tools of data mining to decision making in engineering education—A systematic mapping study. Computer applications in engineering education, 27(3), 744-758.

Burkart, N., & Huber, M. F. (2021). A survey on the explainability of supervised machine learning. Journal of Artificial Intelligence Research, 70, 245-317.

Cai, H., Xu, B., Jiang, L., & Vasilakos, A. V. (2016). IoT-based big data storage systems in cloud computing: perspectives and challenges. IEEE Internet of Things Journal, 4(1), 75-87.

Cai, L., & Zhu, Y. (2015). The challenges of data quality and data quality assessment in the big data era. Data science journal, 14.

Calders, T., & Pechenizkiy, M. (2012). Introduction to the special section on educational data mining. Acm Sigkdd Explorations Newsletter, 13(2), 3-6.

Calvet Liñán, L., & Juan Pérez, Á. A. (2015). Educational Data Mining and Learning Analytics: differences, similarities, and time evolution. International Journal of Educational Technology in Higher Education, 12(3), 98-112.

Canlas, R. D. (2009). Data mining in healthcare: Current applications and issues. School of Information Systems & Management, Carne

Cerna, S., Guyeux, C., Royer, G., Chevallier, C., & Plumerel, G. (2020). Predicting fire brigades operational breakdowns: A real case study. Mathematics, 8(8), 1383.

Cetin, M., & Sevik, H. (2016). Assessing potential areas of ecotourism through a case study in Ilgaz Mountain National Park. Tourism-from empirical research towards practical application, 81-110.

Chadha, S. K., & Kapoor, D. (2009). Effect of switching cost, service quality and customer satisfaction on customer loyalty of cellular service providers in Indian market. IUP Journal of Marketing Management, 8(1), 23.

Chatfield, C. (1995). Model uncertainty, data mining and statistical inference. Journal of the Royal Statistical Society: Series A (Statistics in Society), 158(3), 419-444.

Chatti, M. A., Dyckhoff, A. L., Schroeder, U., & Thüs, H. (2012). A reference model for learning analytics. International journal of Technology Enhanced learning, 4(5-6), 318-331.

Chauhan, C., & Sehgal, S. (2017, May). A review: crime analysis using data mining techniques and algorithms. In 2017 International Conference on Computing, Communication and Automation (ICCCA) (pp. 21-25). IEEE.

Chaurasia, V., & Pal, S. (2014). Data mining approach to detect heart diseases. International Journal of Advanced Computer Science and Information Technology (IJACSIT) Vol, 2, 56-66.

Chawla, D., & Sodhi, N. (2011). Research methodology: Concepts and cases. Vikas Publishing House.

Che, W., Frey, H. C., Fung, J. C., Ning, Z., Qu, H., Lo, H. K., ... & Lau, A. K. (2020). PRAISE-HK: A personalized real-time air quality informatics system for citizen participation in exposure and health risk management. Sustainable Cities and Society, 54, 101986.

Che, Z., Purushotham, S., Cho, K., Sontag, D., & Liu, Y. (2018). Recurrent neural networks for multivariate time series with missing values. Scientific reports, 8(1), 1-12.

Chen, M., & Chen, Z. L. (2015). Recent developments in dynamic pricing research: multiple products, competition, and limited demand information. Production and Operations Management, 24(5), 704-731.

Chen, Z. Y., Fan, Z. P., & Sun, M. (2012). A hierarchical multiple kernel support vector machine for customer churn prediction using longitudinal behavioral data. European Journal of operational research, 223(2), 461-472.

Chitra, K., & Subashini, B. (2011). Customer retention in banking sector using predictive data mining technique. In ICIT 2011 The 5th International Conference on Information Technology.

Chu, T. (2022). Research on College Students’ Physique Testing Platform Based on Big Data Analysis. Mathematical Problems in Engineering, 2022.

Cios, K. J., & Kurgan, L. A. (2005). Trends in data mining and knowledge discovery. In Advanced techniques in knowledge discovery and data mining (pp. 1-26). Springer, London.

Cohen, M. C. (2018). Big data and service operations. Production and Operations Management, 27(9), 1709-1723.

Colgate, J. E., Edward, J., Peshkin, M. A., & Wannasuphoprasit, W. (1996). Cobots: Robots for collaboration with human operators.

Connelly, B. S., & Ones, D. S. (2010). An other perspective on personality: meta-analytic integration of observers' accuracy and predictive validity. Psychological bulletin, 136(6), 1092.

Cooke, N. A. (2012). Professional development 2.0 for librarians: Developing an online personal learning network (PLN). Library Hi Tech News.

Corne, D., Dhaenens, C., & Jourdan, L. (2012). Synergies between operations research and data mining: The emerging use of multi-objective approaches. European Journal of Operational Research, 221(3), 469-479.

Coussement, K., & De Bock, K. W. (2013). Customer churn prediction in the online gambling industry: The beneficial effect of ensemble learning. Journal of Business Research, 66(9), 1629-1636.

Coussement, K., Lessmann, S., & Verstraeten, G. (2017). A comparative analysis of data preparation algorithms for customer churn prediction: A case study in the telecommunication industry. Decision Support Systems, 95, 27-36.

Cui, X., Liu, S., & Jia, L. (2015, May). An improved method of semantic driven subtractive clustering algorithm. In 2015 IEEE 5th International Conference on Electronics Information and Emergency Communication (pp. 232-235). IEEE.

Dahiya, K., & Bhatia, S. (2015, September). Customer churn analysis in telecom industry. In 2015 4th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO)(Trends and Future Directions) (pp. 1-6). IEEE.

Dalvi, P. K., Khandge, S. K., Deomore, A., Bankar, A., & Kanade, V. A. (2016, March). Analysis of customer churn prediction in telecom industry using decision trees and logistic regression. In 2016 symposium on colossal data analysis and networking (CDAN) (pp. 1-4). IEEE.

Dangare, C. S., & Apte, S. S. (2012). Improved study of heart disease prediction system using data mining classification techniques. International Journal of Computer Applications, 47(10), 44-48.

Daniel, B. (2015). B ig D ata and analytics in higher education: Opportunities and challenges. British journal of educational technology, 46(5), 904-920.

Dash, S., Shakyawar, S. K., Sharma, M., & Kaushik, S. (2019). Big data in healthcare: management, analysis and future prospects. Journal of Big Data, 6(1), 1-25.

David, H., & Suruliandi, A. (2017). SURVEY ON CRIME ANALYSIS AND PREDICTION USING DATA MINING TECHNIQUES. ICTACT journal on soft computing, 7(3).

De Bock, K. W., & Van den Poel, D. (2012). Reconciling performance and interpretability in customer churn prediction using ensemble learning based on generalized additive models. Expert Systems with Applications, 39(8), 6816-6826.

De Caigny, A., Coussement, K., & De Bock, K. W. (2018). A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees. European Journal of Operational Research, 269(2), 760-772.

De Caigny, A., Coussement, K., De Bock, K. W., & Lessmann, S. (2020). Incorporating textual information in customer churn prediction models based on a convolutional neural network. International Journal of Forecasting, 36(4), 1563-1578.

De Caigny, A., Coussement, K., Verbeke, W., Idbenjra, K., & Phan, M. (2021). Uplift modeling and its implications for B2B customer churn prediction: A segmentation-based modeling approach. Industrial Marketing Management, 99, 28-39.

De, S., Prabu, P., & Paulose, J. (2021, November). Application of Machine Learning in Customer Churn Prediction. In 2021 Innovations in Power and Advanced Computing Technologies (i-PACT) (pp. 1-7). IEEE.

De, S., Prabu, P., & Paulose, J. (2021, September). Effective ml techniques to predict customer churn. In 2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA) (pp. 895-902). IEEE.

Delen, D. (2014). Real-world data mining: applied business analytics and decision making. FT Press.

Deligiannis, A., & Argyriou, C. (2020). Designing a Real-Time Data-Driven Customer Churn Risk Indicator for Subscription Commerce. International Journal of Information Engineering & Electronic Business, 12(4).

Deogun, J. S., Raghavan, V. V., Sarkar, A., & Sever, H. (1997). Data mining: Trends in research and development. Rough Sets and Data Mining, 9-45.

Devi, S. V. S. G. (2014). A survey on distributed data mining and its trends. International Journal of Research in Engineering & Technology (IMPACT: IJRET), 2(3), 107-120.

Devriendt, F., Berrevoets, J., & Verbeke, W. (2021). Why you should stop predicting customer churn and start using uplift models. Information Sciences, 548, 497-515.

Douglas, P., Rice, C., Runswick-Cole, K., Easton, A., Gibson, M. F., Gruson-Wood, J., ... & Shields, R. (2021). Re-storying autism: A body becoming disability studies in education approach. International Journal of Inclusive Education, 25(5), 605-622.

Du, K. L. (2010). Clustering: A neural network approach. Neural networks, 23(1), 89-107.

Du, R. Y., Netzer, O., Schweidel, D. A., & Mitra, D. (2021). Capturing marketing information to fuel growth. Journal of Marketing, 85(1), 163-183.

Duarte, K., Rawat, Y., & Shah, M. (2021). Plm: Partial label masking for imbalanced multi-label classification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 2739-2748).

Durak, G., OZKESKIN, E. E., & Ataizi, M. (2016). QR codes in education and communication. Turkish Online Journal of Distance Education, 17(2).

Duwairi, R., & Abu-Rahmeh, M. (2015). A novel approach for initializing the spherical K-means clustering algorithm. Simulation Modelling Practice and Theory, 54, 49-63.

Dwivedi, H., & Sharma, L. (2011). Leadership through innovation and creativity in marketing strategies of Indian telecom sector: a case study of airtel using factor analysis approach. International Journal of Business Administration, 2(4), 122.

Dwivedi, Y. K., Hughes, D. L., Coombs, C., Constantiou, I., Duan, Y., Edwards, J. S., ... & Upadhyay, N. (2020). Impact of COVID-19 pandemic on information management research and practice: Transforming education, work and life. International journal of information management, 55, 102211.

Dälken, F. (2014). Are porter’ s five competitive forces still applicable? a critical examination concerning the relevance for today’ s business (Bachelor's thesis, University of Twente).

Edward, M., George, B. P., & Sarkar, S. K. (2010). The impact of switching costs upon the service quality – perceived value – customer satisfaction – service loyalty chain: a study in the context of cellular services in India. Services Marketing Quarterly, 31(2), 151-173.

Erdmann, A., & Ponzoa, J. M. (2021). Digital inbound marketing: Measuring the economic performance of grocery e-commerce in Europe and the USA. Technological forecasting and social change, 162, 120373.

Escher, B. I., & Fenner, K. (2011). Recent advances in environmental risk assessment of transformation products. Environmental science & technology, 45(9), 3835-3847.

Esling, P., & Agon, C. (2012). Time-series data mining. ACM Computing Surveys (CSUR), 45(1), 1-34.

Fahad, A., Alshatri, N., Tari, Z., Alamri, A., Khalil, I., Zomaya, A. Y., ... & Bouras, A. (2014). A survey of clustering algorithms for big data: Taxonomy and empirical analysis. IEEE transactions on emerging topics in computing, 2(3), 267-279.

Faisal, A., Yigitcanlar, T., Kamruzzaman, M., & Paz, A. (2021). Mapping two decades of autonomous vehicle research: A systematic scientometric analysis. Journal of Urban Technology, 28(3-4), 45-74.

Fan, C. Y., Chang, P. C., Lin, J. J., & Hsieh, J. C. (2011). A hybrid model combining case-based reasoning and fuzzy decision tree for medical data classification. Applied Soft Computing, 11(1), 632-644.

Fan, G. F., Yu, M., Dong, S. Q., Yeh, Y. H., & Hong, W. C. (2021). Forecasting short-term electricity load using hybrid support vector regression with grey catastrophe and random forest modeling. Utilities Policy, 73, 101294.

Faris, H. (2018). A hybrid swarm intelligent neural network model for customer churn prediction and identifying the influencing factors. Information, 9(11), 288.

Fathian, M., Hoseinpoor, Y., & Minaei-Bidgoli, B. (2016). Offering a hybrid approach of data mining to predict the customer churn based on bagging and boosting methods. Kybernetes.

Fatt, C. K., Khin, E. W. S., & Heng, T. N. (2010). The impact of organizational justice on employee’ s job satisfaction: The Malaysian companies perspectives. American Journal of Economics and Business Administration, 2(1), 56-63.

Fernάndez, A., Garcí a, S., Galar, M., Prati, R. C., Krawczyk, B., & Herrera, F. (2018). Learning from imbalanced data sets (Vol. 10, pp. 978-3). Berlin: Springer.

Fife, D. A., & D’ Onofrio, J. (2022). Common, uncommon, and novel applications of random forest in psychological research. Behavior Research Methods, 1-20.

Fiore, U., Palmieri, F., Castiglione, A., & De Santis, A. (2013). Network anomaly detection with the restricted Boltzmann machine. Neurocomputing, 122, 13-23.

Foote, J., Gaffney, N., & Evans, J. R. (2010). Corporate social responsibility: Implications for performance excellence. Total Quality Management, 21(8), 799-812.

Franke, M., & John, F. (2011). What comes next after recession? ‑ Airline industry scenarios and potential end games. Journal of Air Transport Management, 17(1), 19-26.

Fuchs, C., & Schreier, M. (2011). Customer empowerment in new product development. Journal of product innovation management, 28(1), 17-32.

Fujo, S. W., Subramanian, S., & Khder, M. A. (2022). Customer Churn Prediction in Telecommunication Industry Using Deep Learning. Information Sciences Letters, 11(1), 24.

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. International journal of information management, 35(2), 137-144.

Ganesh, J., Arnold, M. J., & Reynolds, K. E. (2000). Understanding the customer base of service providers: an examination of the differences between switchers and stayers. Journal of marketing, 64(3), 65-87.

Gao, X., Zhang, J., Wei, Z., & Hakonarson, H. (2018). DeepPolyA: a convolutional neural network approach for polyadenylation site prediction. IEEE Access, 6, 24340-24349.

Geetha, M., & Kumari, J. A. (2012). Analysis of churn behavior of consumers in Indian telecom sector. Journal of Indian Business Research.

Ghorban, Z. S., & Tahernejad, H. (2012). A study on effect of brand credibility on word of mouth: With reference to internet service providers in Malaysia. International Journal of Marketing Studies, 4(1), 26.

Ghose, A., & Ipeirotis, P. G. (2010). Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics. IEEE transactions on knowledge and data engineering, 23(10), 1498-1512.

Gibert, K., Izquierdo, J., Sànchez-Marrè, M., Hamilton, S. H., Rodríguez-Roda, I., & Holmes, G. (2018). Which method to use? An assessment of data mining methods in Environmental Data Science. Environmental modelling & software, 110, 3-27.

Giordani, T., Flamini, F., Pompili, M., Viggianiello, N., Spagnolo, N., Crespi, A., ... & Sciarrino, F. (2018). Experimental statistical signature of many-body quantum interference. Nature Photonics, 12(3), 173-178.

Goeschel, K. (2016, March). Reducing false positives in intrusion detection systems using data-mining techniques utilizing support vector machines, decision trees, and naive Bayes for off-line analysis. In SoutheastCon 2016 (pp. 1-6). IEEE.

Gomber, P., Kauffman, R. J., Parker, C., & Weber, B. W. (2018). On the fintech revolution: Interpreting the forces of innovation, disruption, and transformation in financial services. Journal of management information systems, 35(1), 220-265.

Gopal, P., & MohdNawi, N. B. (2021, December). A Survey on Customer Churn Prediction using Machine Learning and data mining Techniques in E-commerce. In 2021 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE) (pp. 1-8). IEEE.

Gordini, N., & Veglio, V. (2017). Customers churn prediction and marketing retention strategies. An application of support vector machines based on the AUC parameter-selection technique in B2B e-commerce industry. Industrial Marketing Management, 62, 100-107.

Grayson, D., & Hodges, A. (2017). Corporate social opportunity!: Seven steps to make corporate social responsibility work for your business. Routledge.

Greene, J. A., Robertson, J., & Costa, L. J. C. (2011). Assessing self-regulated learning using think-aloud methods. Handbook of self-regulation of learning and performance, 313-328.

Greer, L. L., & van Kleef, G. A. (2010). Equality versus differentiation: The effects of power dispersion on group interaction. Journal of Applied Psychology, 95(6), 1032.

Grover, V., Chiang, R. H., Liang, T. P., & Zhang, D. (2018). Creating strategic business value from big data analytics: A research framework. Journal of management information systems, 35(2), 388-423.

Gubela, R. M., Lessmann, S., & Jaroszewicz, S. (2020). Response transformation and profit decomposition for revenue uplift modeling. European Journal of Operational Research, 283(2), 647-661.

Guelman, L., Guillén, M., & Pérez-Marín, A. M. (2012, May). Random forests for uplift modeling: an insurance customer retention case. In International conference on modeling and simulation in engineering, economics and management (pp. 123-133). Springer, Berlin, Heidelberg.

Gui, C. (2017). Analysis of imbalanced data set problem: The case of churn prediction for telecommunication. Artif. Intell. Res., 6(2), 93.

Gunasekaran, A., Rai, B. K., & Griffin, M. (2011). Resilience and competitiveness of small and medium size enterprises: an empirical research. International journal of production research, 49(18), 5489-5509.

Guzmán-Ponce, A., Valdovinos, R. M., Sánchez, J. S., & Marcial-Romero, J. R. (2020). A new under-sampling method to face class overlap and imbalance. Applied Sciences, 10(15), 5164.

Hadden, J., Tiwari, A., Roy, R., & Ruta, D. (2007). Computer assisted customer churn management: State-of-the-art and future trends. Computers & Operations Research, 34(10), 2902-2917.

Hailu, K. (2021). COLLEGE OF LAW AND GOVERNANCE STUDIES SCHOOL OF LAW (Doctoral dissertation, ADDIS ABABA UNIVERSITY).

Han, J., Pei, J., & Tong, H. (2022). Data mining: concepts and techniques. Morgan kaufmann.

Hanif, I. (2019, August). Implementing extreme gradient boosting (xgboost) classifier to improve customer churn prediction. In ICSA 2019: Proceedings of the 1st International Conference on Statistics and Analytics, ICSA 2019, 2-3 August 2019, Bogor, Indonesia (p. 434). European Alliance for Innovation.

Hansen, H., Samuelsen, B. M., & Sallis, J. E. (2013). The moderating effects of need for cognition on drivers of customer loyalty. European Journal of Marketing, 47(8), 1157-1176.

Harding, J. A., Shahbaz, M., & Kusiak, A. (2006). Data mining in manufacturing: a review.

Harris, P., Brunsdon, C., & Charlton, M. (2011). Geographically weighted principal components analysis. International Journal of Geographical Information Science, 25(10), 1717-1736.

Harzing, A. W. (2013). A preliminary test of Google Scholar as a source for citation data: a longitudinal study of Nobel prize winners. Scientometrics, 94(3), 1057-1075.

Hashem, I. A. T., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A., & Khan, S. U. (2015). The rise of “big data” on cloud computing: Review and open research issues. Information systems, 47, 98-115.

Hassani, H., Huang, X., & Silva, E. (2018). Digitalisation and big data mining in banking. Big Data and Cognitive Computing, 2(3), 18.

Hayes, T., Usami, S., Jacobucci, R., & McArdle, J. J. (2015). Using Classification and Regression Trees (CART) and random forests to analyze attrition: Results from two simulations. Psychology and aging, 30(4), 911.

Helm, V., Humbert, A., & Miller, H. (2014). Elevation and elevation change of Greenland and Antarctica derived from CryoSat-2. The Cryosphere, 8(4), 1539-1559.

Henderson, P., Islam, R., Bachman, P., Pineau, J., Precup, D., & Meger, D. (2018, April). Deep reinforcement learning that matters. In Proceedings of the AAAI conference on artificial intelligence (Vol. 32, No. 1).

Hightower, C. E., Riedel, B. J., Feig, B. W., Morris, G. S., Ensor Jr, J. E., Woodruff, V. D., ... & Sun, X. G. (2010). A pilot study evaluating predictors of postoperative outcomes after major abdominal surgery: physiological capacity compared with the ASA physical status classification system. British journal of anaesthesia, 104(4), 465-471.

Holtrop, N., Wieringa, J. E., Gijsenberg, M. J., & Verhoef, P. C. (2017). No future without the past? Predicting churn in the face of customer privacy. International Journal of Research in Marketing, 34(1), 154-172.

Honaker, J., & King, G. (2010). What to do about missing values in time‐series cross‐section data. American journal of political science, 54(2), 561-581.

Hong, X., Chen, S., & Harris, C. J. (2013). Complex-valued B-spline neural networks for modeling and inverse of Wiener systems. Complex-Valued Neural Networks: Advances and Applications, 209-233.

Hormozi, A. M., & Giles, S. (2004). Data mining: A competitive weapon for banking and retail industries. Information systems management, 21(2), 62-71.

Howland, M., Armeli, S., Feinn, R., & Tennen, H. (2017). Daily emotional stress reactivity in emerging adulthood: Temporal stability and its predictors. Anxiety, Stress, & Coping, 30(2), 121-132.

Huang, B. Q., Kechadi, T. M., Buckley, B., Kiernan, G., Keogh, E., & Rashid, T. (2010). A new feature set with new window techniques for customer churn prediction in land-line telecommunications. Expert Systems with Applications, 37(5), 3657-3665.

Huang, B., Buckley, B., & Kechadi, T. M. (2010). Multi-objective feature selection by using NSGA-II for customer churn prediction in telecommunications. Expert Systems with Applications, 37(5), 3638-3646.

Huang, B., Kechadi, M. T., & Buckley, B. (2012). Customer churn prediction in telecommunications. Expert Systems with Applications, 39(1), 1414-1425.

Huang, J., Cheng, X. Q., Guo, J., Shen, H. W., & Yang, K. (2010). Social recommendation with interpersonal influence. In ECAI 2010 (pp. 601-606). IOS Press.

Huang, S. H., & Yang, J. M. (2012). A Study on the Productivity Review forManagement of Performance Using Bibliometric Methodology. In WHICEB (p. 4).

Huang, Y., & Kechadi, T. (2013). An effective hybrid learning system for telecommunication churn prediction. Expert Systems with Applications, 40(14), 5635-5647.

Huang, Y., Leu, M. C., Mazumder, J., & Donmez, A. (2015). Additive manufacturing: current state, future potential, gaps and needs, and recommendations. Journal of Manufacturing Science and Engineering, 137(1).

Huang, Y., Zhu, F., Yuan, M., Deng, K., Li, Y., Ni, B., ... & Zeng, J. (2015, May). Telco churn prediction with big data. In Proceedings of the 2015 ACM SIGMOD international conference on management of data (pp. 607-618).

Hudaib, A., Dannoun, R., Harfoushi, O., Obiedat, R., & Faris, H. (2015). Hybrid data mining models for predicting customer churn. International Journal of Communications, Network and System Sciences, 8(05), 91.

Hung, J. L., & Zhang, K. (2012). Examining mobile learning trends 2003 – 2008: A categorical meta-trend analysis using text mining techniques. Journal of Computing in Higher education, 24(1), 1-17.

Hung, S. Y., Hung, W. H., Tsai, C. A., & Jiang, S. C. (2010). Critical factors of hospital adoption on CRM system: Organizational and information system perspectives. Decision support systems, 48(4), 592-603.

Hung, S. Y., Yen, D. C., & Wang, H. Y. (2006). Applying data mining to telecom churn management. Expert Systems with Applications, 31(3), 515-524.

Höppner, S., Stripling, E., Baesens, B., vanden Broucke, S., & Verdonck, T. (2020). Profit driven decision trees for churn prediction. European journal of operational research, 284(3), 920-933.

Idris, A., & Iftikhar, A. (2019). Intelligent churn prediction for telecom using GP-AdaBoost learning and PSO undersampling. Cluster Computing, 22(3), 7241-7255.

Idris, A., & Khan, A. (2017). Churn prediction system for telecom using filter – wrapper and ensemble classification. The Computer Journal, 60(3), 410-430.

Idris, A., Khan, A., & Lee, Y. S. (2012, October). Genetic programming and adaboosting based churn prediction for telecom. In 2012 IEEE international conference on Systems, Man, and Cybernetics (SMC) (pp. 1328-1332). IEEE.

Idris, A., Khan, A., & Lee, Y. S. (2013). Intelligent churn prediction in telecom: employing mRMR feature selection and RotBoost based ensemble classification. Applied intelligence, 39(3), 659-672.

Idris, A., Rizwan, M., & Khan, A. (2012). Churn prediction in telecom using Random Forest and PSO based data balancing in combination with various feature selection strategies. Computers & Electrical Engineering, 38(6), 1808-1819.

Ishaq, A., Sadiq, S., Umer, M., Ullah, S., Mirjalili, S., Rupapara, V., & Nappi, M. (2021). Improving the prediction of heart failure patients’ survival using SMOTE and effective data mining techniques. IEEE access, 9, 39707-39716.

Ismail, M. R., Awang, M. K., Rahman, M. N. A., & Makhtar, M. (2015). A multi-layer perceptron approach for customer churn prediction. International Journal of Multimedia and Ubiquitous Engineering, 10(7), 213-222.

Iyengar, R., Jedidi, K., Essegaier, S., & Danaher, P. J. (2011). The impact of tariff structure on customer retention, usage, and profitability of access services. Marketing Science, 30(5), 820-836.

Jacobsson, T. J., Hultqvist, A., García-Fernάndez, A., Anand, A., Al-Ashouri, A., Hagfeldt, A., ... & Unger, E. (2022). An open-access database and analysis tool for perovskite solar cells based on the FAIR data principles. Nature Energy, 7(1), 107-115.

Jadhav, R. J., & Pawar, U. T. (2011). Churn prediction in telecommunication using data mining technology. International Journal of Advanced Computer Science and Applications, 2(2).

Jahanzeb, S., & Jabeen, S. (2007). Churn management in the telecom industry of Pakistan: A comparative study of Ufone and Telenor. Journal of Database Marketing & Customer Strategy Management, 14(2), 120-129.

Jahromi, A. T., Stakhovych, S., & Ewing, M. (2014). Managing B2B customer churn, retention and profitability. Industrial Marketing Management, 43(7), 1258-1268.

Jain, P., Coogan, S. C., Subramanian, S. G., Crowley, M., Taylor, S., & Flannigan, M. D. (2020). A review of machine learning applications in wildfire science and management. Environmental Reviews, 28(4), 478-505.

Jairak, K., & Praneetpolgrang, P. (2013). Applying IT governance balanced scorecard and importance-performance analysis for providing IT governance strategy in university. Information Management & Computer Security.

Jamalian, E., & Foukerdi, R. (2018). A hybrid data mining method for customer churn prediction. Engineering, Technology & Applied Science Research, 8(3), 2991-2997.

Jawaria, F. A., Imran, A., Kashif, U. R., Ayse, K. Y., Nadeem, S., & Hasan, A. (2010). Determinants of consumer retention in cellular industry of Pakistan. African Journal of Business Management, 4(12), 2402-2408.

Jessy, J. (2011). An analysis on the customer loyalty in telecom sector: Special reference to Bharath Sanchar Nigam limited, India. African journal of marketing management, 3(1), 1-5.

Jeyakarthic, M., & Venkatesh, S. (2020). An effective customer churn prediction model using adaptive gain with back propagation neural network in cloud computing environment. Journal of Research on the Lepidoptera, 51(1), 386-399.

Jia, Y., Chao, K., Cheng, X., Xu, L., Zhao, X., & Yao, L. (2019). Telecom big data based precise user classification scheme. In 2019 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI) (pp. 1517-1520). IEEE.

Johannes, M., Korteweg, A., & Polson, N. (2014). Sequential learning, predictability, and optimal portfolio returns. The Journal of Finance, 69(2), 611-644.

Johnson, J. M., & Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance. Journal of Big Data, 6(1), 1-54.

Joo, J. (2012). Asking about and predicting consumer preference: Implications for new product development (Doctoral dissertation, University of Toronto).

Juhaňák, L., Zounek, J., & Rohlíková, L. (2019). Using process mining to analyze students' quiz-taking behavior patterns in a learning management system. Computers in Human Behavior, 92, 496-506.

Jukić, N., Sharma, A., Nestorov, S., & Jukić, B. (2015). Augmenting data warehouses with big data. Information Systems Management, 32(3), 200-209.

Jurisic, B., & Azevedo, A. (2011). Building customer – brand relationships in the mobile communications market: The role of brand tribalism and brand reputation. Journal of Brand Management, 18(4), 349-366.

Kamarudin, M. H., Maple, C., Watson, T., & Safa, N. S. (2017). A logitboost-based algorithm for detecting known and unknown web attacks. IEEE Access, 5, 26190-26200.

Kambatla, K., Kollias, G., Kumar, V., & Grama, A. (2014). Trends in big data analytics. Journal of parallel and distributed computing, 74(7), 2561-2573.

Kamel, M., Nour, M., Awad, M., Essa, M., & Abdelbaki, N. (2021, December). Novel Data Mining Approach Predicting Alerts in The Telecom Industry. In 2021 Tenth International Conference on Intelligent Computing and Information Systems (ICICIS) (pp. 200-206). IEEE.

Kanevski, M., Parkin, R., Pozdnukhov, A., Timonin, V., Maignan, M., Demyanov, V., & Canu, S. (2004). Environmental data mining and modeling based on machine learning algorithms and geostatistics. Environmental Modelling & Software, 19(9), 845-855.

Kantardzic, M. (2011). Data mining: concepts, models, methods, and algorithms. John Wiley & Sons.

Kanwal, S., Rashid, J., Kim, J., Nisar, M. W., Hussain, A., Batool, S., & Kanwal, R. (2021, November). An attribute weight estimation using particle swarm optimization and machine learning approaches for customer churn prediction. In 2021 International Conference on Innovative Computing (ICIC) (pp. 1-6). IEEE.

Karegowda, A. G., Manjunath, A. S., & Jayaram, M. A. (2010). Comparative study of attribute selection using gain ratio and correlation based feature selection. International Journal of Information Technology and Knowledge Management, 2(2), 271-277.

Karimi, S., Wang, C., Metke-Jimenez, A., Gaire, R., & Paris, C. (2015). Text and data mining techniques in adverse drug reaction detection. ACM Computing Surveys (CSUR), 47(4), 1-39.

Karnstedt, M., Hennessy, T., Chan, J., Basuchowdhuri, P., Hayes, C., & Strufe, T. (2010). Churn in social networks. In Handbook of social network technologies and applications (pp. 185-220). Springer, Boston, MA.

Karvana, K. G. M., Yazid, S., Syalim, A., & Mursanto, P. (2019, October). Customer churn analysis and prediction using data mining models in banking industry. In 2019 International Workshop on Big Data and Information Security (IWBIS) (pp. 33-38). IEEE.

Kato, P. M. (2010). Video games in health care: Closing the gap. Review of general psychology, 14(2), 113-121.

Kaur, P., Singh, M., & Josan, G. S. (2015). Classification and prediction based data mining algorithms to predict slow learners in education sector. Procedia Computer Science, 57, 500-508.

Kayaalp, F. (2017). Review of Customer Churn Analysis Studies in Telecommunications Industry. Karaelmas Science & Engineering Journal, 7(2).

Kelly, S., Johnston, P., & Danheiser, S. (2017). Value-ology: Aligning sales and marketing to shape and deliver profitable customer value propositions. Springer.

Kendig, C. E. (2016). What is proof of concept research and how does it generate epistemic and ethical categories for future scientific practice?. Science and Engineering Ethics, 22(3), 735-753.

Keramati, A., & Marandi, R. J. (2015). Addressing churn prediction problem with Meta-heuristic, Machine learning, Neural Network and data mining techniques: a case study of a telecommunication company. International Journal of Future Computer and Communication, 4(5), 350.

Keramati, A., Jafari-Marandi, R., Aliannejadi, M., Ahmadian, I., Mozaffari, M., & Abbasi, U. (2014). Improved churn prediction in telecommunication industry using data mining techniques. Applied Soft Computing, 24, 994-1012.

Khalid, S., Khalil, T., & Nasreen, S. (2014, August). A survey of feature selection and feature extraction techniques in machine learning. In 2014 science and information conference (pp. 372-378). IEEE.

Khan, M. R., Manoj, J., Singh, A., & Blumenstock, J. (2015, June). Behavioral modeling for churn prediction: Early indicators and accurate predictors of custom defection and loyalty. In 2015 IEEE International Congress on Big Data (pp. 677-680). IEEE.

Khan, M. T. (2013). Customers loyalty: Concept & definition (a review). International Journal of Information, Business and Management, 5(3), 168.

Khan, Y., Shafiq, S., Naeem, A., Ahmed, S., Safwan, N., & Hussain, S. (2019). Customers churn prediction using artificial neural networks (ANN) in telecom industry. International Journal of Advanced Computer Science and Applications, 10(9).

Khanna, I., & Sharma, S. (2020). Could the National Capital Region serve as a control region for effective air quality management in Delhi. Policy Brief, Collaborative Clean Air Policy Centre, May.

Khayyat, N. T. (2017). A study of telecommunication policies and broadband penetration for Sweden and South Korea. UKH Journal of Science and Engineering, 1(1), 26-38.

Khemakhem, S., Said, F. B., & Boujelbene, Y. (2018). Credit risk assessment for unbalanced datasets based on data mining, artificial neural network and support vector machines. Journal of Modelling in Management.

Khudhair, Z. S., Zubaidi, S. L., Ortega-Martorell, S., Al-Ansari, N., Ethaib, S., & Hashim, K. (2022). A Review of Hybrid Soft Computing and Data Pre-Processing Techniques to Forecast Freshwater Quality’s Parameters: Current Trends and Future Directions. Environments, 9(7), 85.

Kian, T. S., & Yusoff, W. F. W. (2012, December). Generation x and y and their work motivation. In Proceedings International Conference of Technology Management, Business and Entrepreneurship (Vol. 396, p. 408).

Kiguchi, M., Saeed, W., & Medi, I. (2022). Churn prediction in digital game-based learning using data mining techniques: Logistic regression, decision tree, and random forest. Applied Soft Computing, 118, 108491.

Kim, A. J., & Ko, E. (2012). Do social media marketing activities enhance customer equity? An empirical study of luxury fashion brand. Journal of Business research, 65(10), 1480-1486.

Kim, H. S., & Yoon, C. H. (2004). Determinants of subscriber churn and customer loyalty in the Korean mobile telephony market. Telecommunications policy, 28(9-10), 751-765.

Kim, K., Jun, C. H., & Lee, J. (2014). Improved churn prediction in telecommunication industry by analyzing a large network. Expert Systems with Applications, 41(15), 6575-6584.

Kim, M. K., Park, M. C., & Jeong, D. H. (2004). The effects of customer satisfaction and switching barrier on customer loyalty in Korean mobile telecommunication services. Telecommunications policy, 28(2), 145-159.

Kim, S., Chang, Y., Wong, S. F., & Park, M. C. (2020). Customer resistance to churn in a mature mobile telecommunications market. International Journal of Mobile Communications, 18(1), 41-66.

Kim, S., Choi, D., Lee, E., & Rhee, W. (2017). Churn prediction of mobile and online casual games using play log data. PloS one, 12(7), e0180735.

Kim, W. C., & Mauborgne, R. (2014). Blue ocean strategy, expanded edition: How to create uncontested market space and make the competition irrelevant. Harvard business review Press.

Kiragu, S. M. (2014). Assessment of challenges facing insurance companies in building competitive advantage in Kenya: A survey of insurance firms. International journal of social sciences and entrepreneurship, 1(11), 467-490.

Kirui, C., Hong, L., Cheruiyot, W., & Kirui, H. (2013). Predicting customer churn in mobile telephony industry using probabilistic classifiers in data mining. International Journal of Computer Science Issues (IJCSI), 10(2 Part 1), 165.

Kisioglu, P., & Topcu, Y. I. (2011). Applying Bayesian Belief Network approach to customer churn analysis: A case study on the telecom industry of Turkey. Expert Systems with Applications, 38(6), 7151-7157.

Kleissner, C. (1998, January). Data mining for the enterprise. In Proceedings of the Thirty-First Hawaii International Conference on System Sciences (Vol. 7, pp. 295-304). IEEE.

Koscher, K., Czeskis, A., Roesner, F., Patel, S., Kohno, T., Checkoway, S., ... & Savage, S. (2010, May). Experimental security analysis of a modern automobile. In 2010 IEEE symposium on security and privacy (pp. 447-462). IEEE.

Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. Proceedings of the national academy of sciences, 110(15), 5802-5805.

Kostić, S. M., Simić, M. I., & Kostić, M. V. (2020). Social network analysis and churn prediction in telecommunications using graph theory. Entropy, 22(7), 753.

Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. Computational and structural biotechnology journal, 13, 8-17.

Krajewski, A. M., Siegel, J. W., Xu, J., & Liu, Z. K. (2022). Extensible structure-informed prediction of formation energy with improved accuracy and usability employing neural networks. Computational Materials Science, 208, 111254.

Kriegel, H. P., Borgwardt, K. M., Kröger, P., Pryakhin, A., Schubert, M., & Zimek, A. (2007). Future trends in data mining. Data Mining and Knowledge Discovery, 15(1), 87-97.

Kumar, A., Sangwan, S. R., & Nayyar, A. (2020). Multimedia social big data: Mining. In Multimedia big data computing for IoT applications (pp. 289-321). Springer, Singapore.

Kumar, R., & Indrayan, A. (2011). Receiver operating characteristic (ROC) curve for medical researchers. Indian pediatrics, 48(4), 277-287.

Lakshmi, B. N., & Raghunandhan, G. H. (2011, February). A conceptual overview of data mining. In 2011 National Conference on Innovations in Emerging Technology (pp. 27-32). IEEE.

Lalazaryan, A., & Zare-Farashbandi, F. (2014). A review of models and theories of health information seeking behavior. International Journal of Health System and Disaster Management, 2(4), 193.

Lalwani, P., Mishra, M. K., Chadha, J. S., & Sethi, P. (2022). Customer churn prediction system: a machine learning approach. Computing, 104(2), 271-294.

Lan, K., Wang, D. T., Fong, S., Liu, L. S., Wong, K. K., & Dey, N. (2018). A survey of data mining and deep learning in bioinformatics. Journal of medical systems, 42(8), 1-20.

Langarizadeh, M., & Moghbeli, F. (2016). Applying naive bayesian networks to disease prediction: a systematic review. Acta Informatica Medica, 24(5), 364.

Le, V. P. M., Meenagh, D., Minford, P., & Wickens, M. (2011). How much nominal rigidity is there in the US economy? Testing a New Keynesian DSGE Model using indirect inference. Journal of Economic Dynamics and Control, 35(12), 2078-2104.

Lee, C. S., & Ma, L. (2012). News sharing in social media: The effect of gratifications and prior experience. Computers in human behavior, 28(2), 331-339.

Lee, E. B., Kim, J., & Lee, S. G. (2017). Predicting customer churn in mobile industry using data mining technology. Industrial Management & Data Systems.

Lee, H., Lee, Y., Cho, H., Im, K., & Kim, Y. S. (2011). Mining churning behaviors and developing retention strategies based on a partial least squares (PLS) model. Decision Support Systems, 52(1), 207-216.

Lee, I., & Shin, Y. J. (2018). Fintech: Ecosystem, business models, investment decisions, and challenges. Business horizons, 61(1), 35-46.

Lee, J. S., & Hsiang, J. (2020). Patent classification by fine-tuning BERT language model. World Patent Information, 61, 101965.

Lee, K. C., & Jo, N. Y. (2010, December). Bayesian network approach to predict mobile churn motivations: Emphasis on general Bayesian network, Markov blanket, and what-if simulation. In International Conference on Future Generation Information Technology (pp. 304-313). Springer, Berlin, Heidelberg.

Lee, M., Yun, J. J., Pyka, A., Won, D., Kodama, F., Schiuma, G., ... & Zhao, X. (2018). How to respond to the fourth industrial revolution, or the second information technology revolution? Dynamic new combinations between technology, market, and society through open innovation. Journal of Open Innovation: Technology, Market, and Complexity, 4(3), 21.

Lee, S., & Jun, C. H. (2018). Fast incremental learning of logistic model tree using least angle regression. Expert Systems with Applications, 97, 137-145.

Leevy, J. L., Khoshgoftaar, T. M., Bauder, R. A., & Seliya, N. (2018). A survey on addressing high-class imbalance in big data. Journal of Big Data, 5(1), 1-30.

Leiria, M., Matos, N., & Rebelo, E. (2021). Non-life insurance cancellation: a systematic quantitative literature review. The Geneva Papers on Risk and Insurance-Issues and Practice, 46(4), 593-613.

Lemmens, A., & Gupta, S. (2020). Managing churn to maximize profits. Marketing Science, 39(5), 956-973.

Lemos, M. C., & Rood, R. B. (2010). Climate projections and their impact on policy and practice. Wiley interdisciplinary reviews: climate change, 1(5), 670-682.

Lepot, M., Aubin, J. B., & Clemens, F. H. (2017). Interpolation in time series: An introductive overview of existing methods, their performance criteria and uncertainty assessment. Water, 9(10), 796.

Li, D., Liang, W., Feng, X., Ruan, T., & Jiang, G. (2021). Recent advances in data-mining techniques for measuring transformation products by high-resolution mass spectrometry. TrAC Trends in Analytical Chemistry, 143, 116409.

Li, D., Wang, S., & Li, D. (2015). Spatial data mining. Berlin, Heidelberg: Springer Berlin Heidelberg.

Li, H., Yang, D., Yang, L., Lu, Y., & Lin, X. (2016, October). Supervised massive data analysis for telecommunication customer churn prediction. In 2016 IEEE International Conferences on Big Data and Cloud Computing (BDCloud), Social Computing and Networking (SocialCom), Sustainable Computing and Communications (SustainCom)(BDCloud-SocialCom-SustainCom) (pp. 163-169). IEEE.

Li, H., Zhang, Z., & Zhao, Z. Z. (2019). Data-mining for processes in chemistry, materials, and engineering. Processes, 7(3), 151.

Li, S., Li, Y., Zhao, W. X., Ding, B., & Wen, J. R. (2021). Interpretable Aspect-Aware Capsule Network for Peer Review Based Citation Count Prediction. ACM Transactions on Information Systems (TOIS), 40(1), 1-29.

Li, X., Xie, Q., Jiang, J., Zhou, Y., & Huang, L. (2019). Identifying and monitoring the development trends of emerging technologies using patent analysis and Twitter data mining: The case of perovskite solar cell technology. Technological Forecasting and Social Change, 146, 687-705.

Li, X., Zhang, S. Q., Xu, L. C., & Hong, X. (2020). Predicting regioselectivity in radical C─H functionalization of heterocycles through machine learning. Angewandte Chemie International Edition, 59(32), 13253-13259.

Lima, E., Mues, C., & Baesens, B. (2011). Monitoring and backtesting churn models. Expert Systems with Applications, 38(1), 975-982.

Lin, T. C., Wu, S., Hsu, J. S. C., & Chou, Y. C. (2012). The integration of value-based adoption and expectation ‐ confirmation models: An example of IPTV continuance intention. Decision Support Systems, 54(1), 63-75.

Linoff, G. S., & Berry, M. J. (2011). Data mining techniques: for marketing, sales, and customer relationship management. John Wiley & Sons.

Liu, B., Zhou, X., Wang, Y., Hu, J., He, L., Zhang, R., ... & Guo, Y. (2012). Data processing and analysis in real ‐ world traditional Chinese medicine clinical data: challenges and approaches. Statistics in medicine, 31(7), 653-660.

Liu, C. T., Guo, Y. M., & Lee, C. H. (2011). The effects of relationship quality and switching barriers on customer loyalty. International Journal of Information Management, 31(1), 71-79.

Liu, D. S., & Fan, S. J. (2014). A modified decision tree algorithm based on genetic algorithm for mobile user classification problem. The Scientific World Journal, 2014.

Liu, H., & Motoda, H. (2012). Feature selection for knowledge discovery and data mining (Vol. 454). Springer Science & Business Media.

Long, X., Yin, W., An, L., Ni, H., Huang, L., Luo, Q., & Chen, Y. (2012, March). Churn analysis of online social network users using data mining techniques. In Proceedings of the international MultiConference of Engineers and Conputer Scientists (Vol. 1).

Long, X., Yin, W., An, L., Ni, H., Huang, L., Luo, Q., & Chen, Y. (2012). Churn analysis of online social network users using data mining techniques. In Proceedings of the international MultiConference of Engineers and Conputer Scientists (Vol. 1).

Lu, G., & Tian, X. (2021). An efficient communication intrusion detection scheme in ami combining feature dimensionality reduction and improved LSTM. Security and Communication Networks, 2021.

Lu, N., Lin, H., Lu, J., & Zhang, G. (2012). A customer churn prediction model in telecom industry using boosting. IEEE Transactions on Industrial Informatics, 10(2), 1659-1665.

Luppa, M., Luck, T., Weyerer, S., König, H. H., Brähler, E., & Riedel-Heller, S. G. (2010). Prediction of institutionalization in the elderly. A systematic review. Age and ageing, 39(1), 31-38.

Ma, J., Sheridan, R. P., Liaw, A., Dahl, G. E., & Svetnik, V. (2015). Deep neural nets as a method for quantitative structure − activity relationships. Journal of chemical information and modeling, 55(2), 263-274.

Machado, M. R., Karray, S., & de Sousa, I. T. (2019, August). LightGBM: An effective decision tree gradient boosting method to predict customer loyalty in the finance industry. In 2019 14th International Conference on Computer Science & Education (ICCSE) (pp. 1111-1116). IEEE.

Mahajan, V., Misra, R., & Mahajan, R. (2017). Review on factors affecting customer churn in telecom sector. International Journal of Data Analysis Techniques and Strategies, 9(2), 122-144.

Mahapatra, P., & Singh, S. K. (2021). Artificial Intelligence and Machine Learning: Discovering New Ways of Doing Banking Business. In Artificial Intelligence and Machine Learning in Business Management (pp. 53-80). CRC Press.

Maheswari, K., & Priya, P. P. A. (2017, March). Predicting customer behavior in online shopping using SVM classifier. In 2017 IEEE international conference on intelligent techniques in control, optimization and signal processing (INCOS) (pp. 1-5). IEEE.

Makhtar, M., Nafis, S., Mohamed, M. A., Awang, M. K., Rahman, M. N. A., & Deris, M. M. (2017). Churn classification model for local telecommunication company based on rough set theory. Journal of Fundamental and Applied Sciences, 9(6S), 854-868.

Malik, M. B., Ghazi, M. A., & Ali, R. (2012, November). Privacy preserving data mining techniques: current scenario and future prospects. In 2012 third international conference on computer and communication technology (pp. 26-32). IEEE.

Malik, S., Tyagi, A. K., & Mahajan, S. (2022). Architecture, Generative Model, and Deep Reinforcement Learning for IoT Applications: Deep Learning Perspective. In Artificial Intelligence-based Internet of Things Systems (pp. 243-265). Springer, Cham.

Mammeri, A., Boukerche, A., & Fang, Z. (2015). Video streaming over vehicular ad hoc networks using erasure coding. IEEE Systems Journal, 10(2), 785-796.

Mariscal, G., Marban, O., & Fernandez, C. (2010). A survey of data mining and knowledge discovery process models and methodologies. The Knowledge Engineering Review, 25(2), 137-166.

Marra, G., & Wood, S. N. (2012). Coverage properties of confidence intervals for generalized additive model components. Scandinavian Journal of Statistics, 39(1), 53-74.

Martínez-Álvarez, F., Troncoso, A., Asencio-Cortés, G., & Riquelme, J. C. (2015). A survey on data mining techniques applied to electricity-related time series forecasting. Energies, 8(11), 13162-13193.

Marwat, M. I., Khan, J. A., Alshehri, D. M. D., Ali, M. A., Ali, H., & Assam, M. (2022). Sentiment Analysis of Product Reviews to Identify Deceptive Rating Information in Social Media: A SentiDeceptive Approach. KSII Transactions on Internet and Information Systems (TIIS), 16(3), 830-860.

Massaoudi, M., Refaat, S. S., Chihi, I., Trabelsi, M., Oueslati, F. S., & Abu-Rub, H. (2021). A novel stacked generalization ensemble-based hybrid LGBM-XGB-MLP model for short-term load forecasting. Energy, 214, 118874.

Matsen, F. A., Kodner, R. B., & Armbrust, E. (2010). pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. BMC bioinformatics, 11(1), 1-16.

Matthews, J. R. (2017). The evaluation and measurement of library services. ABC-CLIO.

McIlroy, A., & Barnett, S. (2000). Building customer relationships: do discount cards work?. Managing Service Quality: An International Journal.

Melian, D. M., Dumitrache, A., Stancu, S., & Nastu, A. (2022). Customer Churn Prediction in Telecommunication Industry. A Data Analysis Techniques Approach. Postmodern Openings, 13(1 Sup1), 78-104.

Mention, A. L., & Bontis, N. (2013). Intellectual capital and performance within the banking sector of Luxembourg and Belgium. Journal of Intellectual capital.

Miguéis, V. L., Freitas, A., Garcia, P. J., & Silva, A. (2018). Early segmentation of students according to their academic performance: A predictive modelling approach. Decision Support Systems, 115, 36-51.

Mikalef, P., Boura, M., Lekakos, G., & Krogstie, J. (2019). Big data analytics and firm performance: Findings from a mixed-method approach. Journal of Business Research, 98, 261-276.

Milovic, B., & Milovic, M. (2012). Prediction and decision making in health care using data mining. Kuwait Chapter of the Arabian Journal of Business and Management Review, 1(12), 126.

Minor, K. I., Wells, J. B., Angel, E., & Matz, A. K. (2011). Predictors of early job turnover among juvenile correctional facility staff. Criminal Justice Review, 36(1), 58-75.

Mirkovski, K., Lowry, P. B., & Feng, B. (2016). Factors that influence interorganizational use of information and communications technology in relationship-based supply chains: evidence from the Macedonian and American wine industries. Supply Chain Management: An International Journal.

Mishra, A., & Reddy, U. S. (2017, November). A comparative study of customer churn prediction in telecom industry using ensemble based classifiers. In 2017 International Conference on Inventive Computing and Informatics (ICICI) (pp. 721-725). IEEE.

Mishra, K., & Rani, R. (2017, August). Churn prediction in telecommunication using machine learning. In 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS) (pp. 2252-2257). IEEE.

Mitchell, T. M. (1999). Machine learning and data mining. Communications of the ACM, 42(11), 30-36.

Mitrović, S., Baesens, B., Lemahieu, W., & De Weerdt, J. (2018). On the operational efficiency of different feature types for telco Churn prediction. European Journal of Operational Research, 267(3), 1141-1155.

Moed, H. F., Markusova, V., & Akoev, M. (2018). Trends in Russian research output indexed in Scopus and Web of Science. Scientometrics, 116(2), 1153-1180.

Moeyersoms, J., & Martens, D. (2015). Including high-cardinality attributes in predictive models: A case study in churn prediction in the energy sector. Decision support systems, 72, 72-81.

Mohammadi, E., & Karami, A. (2022). Exploring research trends in big data across disciplines: A text mining analysis. Journal of Information Science, 48(1), 44-56.

Mohanty, S., Jagadeesh, M., & Srivatsa, H. (2013). Big data imperatives: Enterprise ‘Big Data’ warehouse, ‘BI’ implementations and analytics. Apress.

Mohapatra, S. K., Mallick, P. K., & Mohanty, M. N. (2021). Big Data Application in Health Care: A Study. In Technical Advancements of Machine Learning in Healthcare (pp. 31-58). Springer, Singapore.

Momin, S., Bohra, T., & Raut, P. (2020). Prediction of customer churn using machine learning. In EAI International Conference on Big Data Innovation for Sustainable Cognitive Computing (pp. 203-212). Springer, Cham.

Moreira, M., Pelissari, P. I. B. G. B., Parr, C., Wohrmeyer, C., & Pandolfelli, V. C. (2017). Data mining on technical trends and international collaborations in the refractory ceramic area. Ceramics international, 43(9), 6876-6884.

Moro, S., Esmerado, J., Ramos, P., & Alturas, B. (2019). Evaluating a guest satisfaction model through data mining. International Journal of Contemporary Hospitality Management, 32(4), 1523-1538.

Morris, M. R., Teevan, J., & Panovich, K. (2010, April). What do people ask their social networks, and why? A survey study of status message Q&A behavior. In Proceedings of the SIGCHI conference on Human factors in computing systems (pp. 1739-1748).

Mosavi, A., Ozturk, P., & Chau, K. W. (2018). Flood prediction using machine learning models: Literature review. Water, 10(11), 1536.

Moslehi, F., & Haeri, A. (2020). A novel hybrid wrapper – filter approach based on genetic algorithm, particle swarm optimization for feature subset selection. Journal of Ambient Intelligence and Humanized Computing, 11(3), 1105-1127.

Mougan, C., Masip, D., Nin, J., & Pujol, O. (2021, September). Quantile encoder: Tackling high cardinality categorical features in regression problems. In International Conference on Modeling Decisions for Artificial Intelligence (pp. 168-180). Springer, Cham.

Mould, D. R., & Upton, R. N. (2012). Basic concepts in population modeling, simulation, and model ‐ based drug development. CPT: pharmacometrics & systems pharmacology, 1(9), 1-14.

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning. Proceedings of the National Academy of Sciences, 116(44), 22071-22080.

Mustafa, N., Ling, L. S., & Razak, S. F. A. (2021). Customer churn prediction for telecommunication industry: A Malaysian Case Study. F1000Research, 10.

Mäntylä, M. V., Graziotin, D., & Kuutila, M. (2018). The evolution of sentiment analysis—A review of research topics, venues, and top cited papers. Computer Science Review, 27, 16-32.

Nardi, P. M. (2018). Doing survey research: A guide to quantitative methods. Routledge.

Narteni, S., Orani, V., Cambiaso, E., Rucco, M., & Mongelli, M. (2022). On the Intersection of Explainable and Reliable AI for physical fatigue prediction. IEEE Access, 10, 76243-76260.

Natarajan, S., Khot, T., Kersting, K., Gutmann, B., & Shavlik, J. (2012). Gradient-based boosting for statistical relational learning: The relational dependency network case. Machine Learning, 86(1), 25-56.

Nazir, S., Asif, M., & Ahmad, S. (2019, February). The Evolution of Trends and Techniques used for Data Mining. In 2019 2nd International Conference on Advancements in Computational Sciences (ICACS) (pp. 1-6). IEEE.

Neto, P. A. D. M. S., do Carmo Machado, I., McGregor, J. D., De Almeida, E. S., & de Lemos Meira, S. R. (2011). A systematic mapping study of software product lines testing. Information and Software Technology, 53(5), 407-423.

Ng, C. K., Wu, C. H., Yung, K. L., Ip, W. H., & Cheung, T. (2018). A semantic similarity analysis of Internet of Things. Enterprise Information Systems, 12(7), 820-855.

Nikolaou, S., Van Renesse, R., & Schiper, N. (2015). Proactive cache placement on cooperative client caches for online social networks. IEEE Transactions on Parallel and Distributed Systems, 27(4), 1174-1186.

Nitzan, I., & Libai, B. (2011). Social effects on customer retention. Journal of Marketing, 75(6), 24-38.

Oghojafor, B., Mesike, G., Bakarea, R., Omoera, C., & Adeleke, I. (2012). Discriminant analysis of factors affecting telecoms customer churn. International Journal of Business Administration, 3(2), 59.

Olle, G. D. O., & Cai, S. (2014). A hybrid churn prediction model in mobile telecommunication industry. International Journal of e-Education, e-Business, e-Management and e-Learning, 4(1), 55.

Olszak, C. M. (2015). Business intelligence and analytics in organizations. In Advances in ICT for Business, Industry and Public Sector (pp. 89-109). Springer, Cham.

Ominike Akpovi, A. (2016). Mobile number portability (MNP) in Nigeria. European Journal of Computer Science and Information Technology, 4(4), 41-52.

Orozco, J., Tarhini, A., & Tarhini, T. (2015). A framework of IS/business alignment management practices to improve the design of IT Governance architectures. International Journal of Business and Management, 10(4), 1.

Osterwalder, A., & Pigneur, Y. (2010). Business model generation: a handbook for visionaries, game changers, and challengers (Vol. 1). John Wiley & Sons.

Owczarczuk, M. (2010). Churn models for prepaid customers in the cellular telecommunication industry using large data marts. Expert Systems with Applications, 37(6), 4710-4712.

Oztemel, E., & Gursev, S. (2020). Literature review of Industry 4.0 and related technologies. Journal of Intelligent Manufacturing, 31(1), 127-182.

Padhy, N., Mishra, D., & Panigrahi, R. (2012). The survey of data mining applications and feature scope. arXiv preprint arXiv:1211.5723.

Pagani, R. N., Kovaleski, J. L., & Resende, L. M. (2015). Methodi Ordinatio: a proposed methodology to select and rank relevant scientific papers encompassing the impact factor, number of citation, and year of publication. Scientometrics, 105(3), 2109-2135.

Paidi, A. N. (2012). Data mining: Future trends and applications. International Journal of Modern Engineering Research, 2(6), 4657-4663.

Pamina, J., Beschi Raja, J., Sam Peter, S., Soundarya, S., Sathya Bama, S., & Sruthi, M. S. (2019, September). Inferring machine learning based parameter estimation for telecom churn prediction. In International Conference On Computational Vision and Bio Inspired Computing (pp. 257-267). Springer, Cham.

Papamitsiou, Z., & Economides, A. A. (2014). Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence. Journal of Educational Technology & Society, 17(4), 49-64.

Parack, S., Zahid, Z., & Merchant, F. (2012, January). Application of data mining in educational databases for predicting academic trends and patterns. In 2012 IEEE international conference on technology enhanced education (ICTEE) (pp. 1-4). IEEE.

Park, G., Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Kosinski, M., Stillwell, D. J., ... & Seligman, M. E. (2015). Automatic personality assessment through social media language. Journal of personality and social psychology, 108(6), 934.

Park, N. G., & Zhu, K. (2020). Scalable fabrication and coating methods for perovskite solar cells and solar modules. Nature Reviews Materials, 5(5), 333-350.

Parry, E., & Urwin, P. (2011). Generational differences in work values: A review of theory and evidence. International journal of management reviews, 13(1), 79-96.

Patel, P. S., & Desai, S. (2015). A comparative study on data mining tools. International Journal of Advanced Trends in Computer Science and Engineering, 4(2).

Pathak, S. (2014). Social implications of data mining techniques. J. Basic Appl. Eng. Res.(JBAER), 1, 860.

Paulrajan, R., & Rajkumar, H. (2011). Service quality and customers preference of cellular mobile service providers. Journal of technology management & innovation, 6(1), 38-45.

Pejić Bach, M., Pivar, J., & Jaković, B. (2021). Churn management in telecommunications: hybrid approach using cluster analysis and decision trees. Journal of Risk and Financial Management, 14(11), 544.

Peral, J., Maté, A., & Marco, M. (2017). Application of data mining techniques to identify relevant key performance indicators. Computer Standards & Interfaces, 54, 76-85.

Perera, K. D., Weragoda, G. K., Haputhanthri, R., & Rodrigo, S. K. (2021). Study of concentration dependent curcumin interaction with serum biomolecules using ATR-FTIR spectroscopy combined with Principal Component Analysis (PCA) and Partial Least Square Regression (PLS-R). Vibrational Spectroscopy, 116, 103288.

Perols, J. (2011). Financial statement fraud detection: An analysis of statistical and machine learning algorithms. Auditing: A Journal of Practice & Theory, 30(2), 19-50.

Perumal, M., Velumani, B., Sadhasivam, A., & Ramaswamy, K. (2015). Spatial data mining approaches for gis – a brief review. In Emerging ICT for Bridging the Future-Proceedings of the 49th Annual Convention of the Computer Society of India CSI Volume 2 (pp. 579-592). Springer, Cham.

Petkovski, A. J., Stojkoska, B. L. R., Trivodaliev, K. V., & Kalajdziski, S. A. (2016). Analysis of churn prediction: a case study on telecommunication services in Macedonia. In 2016 24th Telecommunications Forum (TELFOR) (pp. 1-4). IEEE.

Petre, R. S. (2012). Data mining in cloud computing. Database Systems Journal, 3(3), 67-71.

Phua, C., Lee, V., Smith, K., & Gayler, R. (2010). A comprehensive survey of data mining-based fraud detection research. arXiv preprint arXiv:1009.6119.

PK, S. K. (2018). Author productivity and the application of Lotka’s law in LIS publications. Annals of Library and Information Studies (ALIS), 64(4), 234-241.

Pouriyeh, S., Vahid, S., Sannino, G., De Pietro, G., Arabnia, H., & Gutierrez, J. (2017, July). A comprehensive investigation and comparison of machine learning techniques in the domain of heart disease. In 2017 IEEE symposium on computers and communications (ISCC) (pp. 204-207). IEEE.

Pradhan, B. (2013). A comparative study on the predictive ability of the decision tree, support vector machine and neuro-fuzzy models in landslide susceptibility mapping using GIS. Computers & Geosciences, 51, 350-365.

Prasad, U. D., & Madhavi, S. (2012). Prediction of churn behavior of bank customers using data mining tools. Business Intelligence Journal, 5(1), 96-101.

Praseeda, C. K., & Shivakumar, B. L. (2021). Fuzzy particle swarm optimization (FPSO) based feature selection and hybrid kernel distance based possibilistic fuzzy local information C-means (HKD-PFLICM) clustering for churn prediction in telecom industry. SN Applied Sciences, 3(6), 1-18.

Price, W. N., & Cohen, I. G. (2019). Privacy in the age of medical big data. Nature medicine, 25(1), 37-43.

Provost, F., & Fawcett, T. (2013). Data science and its relationship to big data and data-driven decision making. Big data, 1(1), 51-59.

Pustokhina, I. V., Pustokhin, D. A., Aswathy, R. H., Jayasankar, T., Jeyalakshmi, C., Díaz, V. G., & Shankar, K. (2021). Dynamic customer churn prediction strategy for business intelligence using text analytics with evolutionary optimization algorithms. Information Processing & Management, 58(6), 102706.

Pustokhina, I. V., Pustokhin, D. A., Nguyen, P. T., Elhoseny, M., & Shankar, K. (2021). Multi-objective rain optimization algorithm with WELM model for customer churn prediction in telecommunication sector. Complex & Intelligent Systems, 1-13.

Qureshi, S. A., Rehman, A. S., Qamar, A. M., Kamal, A., & Rehman, A. (2013). Telecommunication subscribers' churn prediction model using machine learning. In Eighth international conference on digital information management (ICDIM 2013) (pp. 131-136). IEEE.

Rachid, A. D., Abdellah, A., Belaid, B., & Rachid, L. (2018). Clustering prediction techniques in defining and predicting customers defection: The case of e-commerce context. International Journal of Electrical and Computer Engineering, 8(4), 2367.

Raghupathi, W., & Raghupathi, V. (2014). Big data analytics in healthcare: promise and potential. Health information science and systems, 2(1), 1-10.

Ramesh, P., Jeba Emilyn, J., & Vijayakumar, V. (2022). Hybrid Artificial Neural Networks Using Customer Churn Prediction. Wireless Personal Communications, 124(2), 1695-1709.

Rayhan, M., Sultana, S., & Majid, A. (2019). Financial factors analysis for acquisition premium and anticipation using extreme gradient boosting and deep recurrent neural network (Doctoral dissertation, Brac University).

Reddy, D. L. C. (2011). A Review on Data mining from Past to the Future. International Journal of Computer Applications, 975(2011), 8887.

Ref, R., & Guan, L. (2012). A New Look at an Old Problem Selling to Small‐and Medium‐Size Businesses. Selling Through Someone Else: How to use Agile Sales Networks and Partners to Sell More, 217-233.

Reicher, R., & Szeghegyi, Á. (2015). Factors affecting the selection and implementation of a customer relationship management (CRM) process. Acta Polytechnica Hungarica, 12(4), 183-200.

Reinartz, W. J., & Kumar, V. (2003). The impact of customer relationship characteristics on profitable lifetime duration. Journal of marketing, 67(1), 77-99.

Rhoudri, S., & Benazzou, L. (2021). Predictive Factors of Withdrawal Behavior among Profit-Sharing Investment Depositors in Morocco. International Journal of Accounting, Finance, Auditing, Management and Economics, 2(4), 498-516.

Richter, Y., Yom-Tov, E., & Slonim, N. (2010, April). Predicting customer churn in mobile networks through analysis of social groups. In Proceedings of the 2010 SIAM international conference on data mining (pp. 732-741). Society for Industrial and Applied Mathematics.

Ristoski, P., & Paulheim, H. (2016). Semantic Web in data mining and knowledge discovery: A comprehensive survey. Journal of Web Semantics, 36, 1-22.

Ritchie, M. D., Holzinger, E. R., Li, R., Pendergrass, S. A., & Kim, D. (2015). Methods of integrating data to uncover genotype – phenotype interactions. Nature Reviews Genetics, 16(2), 85-97.

Rodan, A., & Faris, H. (2015, November). Echo state network with SVM-readout for customer churn prediction. In 2015 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT) (pp. 1-5). IEEE.

Rodan, A., Fayyoumi, A., Faris, H., Alsakran, J., & Al-Kadi, O. (2015). Negative correlation learning for customer churn prediction: A comparison study. The Scientific World Journal, 2015.

Rodrigues, M. W., Isotani, S., & Zarate, L. E. (2018). Educational Data Mining: A review of evaluation process in the e-learning. Telematics and Informatics, 35(6), 1701-1717.

Rogers, D. L. (2016). The digital transformation playbook: Rethink your business for the digital age. Columbia University Press.

Romero, C., & Ventura, S. (2013). Data mining in education. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 3(1), 12-27.

Rothmeier, K., Pflanzl, N., Hüllmann, J. A., & Preuss, M. (2020). Prediction of player churn and disengagement based on user activity data of a freemium online strategy game. IEEE Transactions on Games, 13(1), 78-88.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206-215.

Rygielski, C., Wang, J. C., & Yen, D. C. (2002). Data mining techniques for customer relationship management. Technology in society, 24(4), 483-502.

Rzepakowski, P., & Jaroszewicz, S. (2012). Decision trees for uplift modeling with single and multiple treatments. Knowledge and Information Systems, 32(2), 303-327.

Sadhasivam, J., Muthukumaran, V., Raja, J. T., Vinothkumar, V., Deepa, R., & Nivedita, V. (2021, July). Applying data mining technique to predict trends in air pollution in Mumbai. In Journal of Physics: Conference Series (Vol. 1964, No. 4, p. 042055). IOP Publishing.

Safara, F. (2020). A computational model to predict consumer behaviour during COVID-19 pandemic. Computational Economics, 1-14.

Salathé, M., & Khandelwal, S. (2011). Assessing vaccination sentiments with online social media: implications for infectious disease dynamics and control. PLoS computational biology, 7(10), e1002199.

Salloum, S. A., Al-Emran, M., Monem, A. A., & Shaalan, K. (2017). A survey of text mining in social media: facebook and twitter perspectives. Adv. Sci. Technol. Eng. Syst. J, 2(1), 127-133.

Salloum, S. A., Alshurideh, M., Elnagar, A., & Shaalan, K. (2020, April). Mining in educational data: review and future directions. In The International Conference on Artificial Intelligence and Computer Vision (pp. 92-102). Springer, Cham.

Salo, F., Nassif, A. B., & Essex, A. (2019). Dimensionality reduction with IG-PCA and ensemble classifier for network intrusion detection. Computer Networks, 148, 164-175.

Saradhi, V. V., & Palshikar, G. K. (2011). Employee churn prediction. Expert Systems with Applications, 38(3), 1999-2006.

Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. SN Computer Science, 2(3), 1-21.

Saroja Thota, L., & Appa Rao, A. (2013). Overview of Emperical Data Mining Research. International journal of advanced research in computer science, 4(10).

Schmidt, C., & Sun, W. N. (2018). Synthesizing agile and knowledge discovery: case study results. Journal of Computer Information Systems, 58(2), 142-150.

Schramm-Klein, H., Morschett, D., & Swoboda, B. (2015). Retailer corporate social responsibility: Shedding light on CSR’s impact on profit of intermediaries in marketing channels. International Journal of Retail & Distribution Management.

Seera, M., & Lim, C. P. (2014). A hybrid intelligent system for medical data classification. Expert systems with applications, 41(5), 2239-2249.

Seo, D., Ranganathan, C., & Babad, Y. (2008). Two-level model of customer retention in the US mobile telecommunications service market. Telecommunications policy, 32(3-4), 182-196.

Sergeev, A. P., Buevich, A. G., Baglaeva, E. M., & Shichkin, A. V. (2019). Combining spatial autocorrelation with machine learning increases prediction accuracy of soil heavy metals. Catena, 174, 425-435.

Serrano, Y., Rahn, M., Crump, E., Venkatramanan, S., & Haas, J. D. (2013). Iron deficiency and physical activity after a dietary iron intervention in female Indian tea pickers.

Shaaban, E., Helmy, Y., Khedr, A., & Nasr, M. (2012). A proposed churn prediction model. International Journal of Engineering Research and Applications, 2(4), 693-697.

Shanahan, J. G. (2012). Soft computing for knowledge discovery: introducing Cartesian granule features (Vol. 570). Springer Science & Business Media.

Sharma., & Kumar. (2011). A neural network based approach for Predicting Customer churn in cellular network services,International Journal of Computer Applications. 27 (11), 26 – 31.

Shaw, M. J., Subramaniam, C., Tan, G. W., & Welge, M. E. (2001). Knowledge management and data mining for marketing. Decision support systems, 31(1), 127-137.

Shekhar, S., Evans, M. R., Kang, J. M., & Mohan, P. (2011). Identifying patterns in spatial information: A survey of methods. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 1(3), 193-214.

Shen, Q., Li, H., Liao, Q., Zhang, W., & Kalilou, K. (2014, May). Improving churn prediction in telecommunications using complementary fusion of multilayer features based on factorization and construction. In The 26th Chinese Control and Decision Conference (2014 CCDC) (pp. 2250-2255). IEEE.

Shi, C., & Wang, Y. (2021). Non-parametric machine learning methods for interpolation of spatially varying non-stationary and non-Gaussian geotechnical properties. Geoscience Frontiers, 12(1), 339-350.

Shin, D., & Shim, J. (2021). A systematic review on data mining for mathematics and science education. International Journal of Science and Mathematics Education, 19(4), 639-659.

Shukla, A. K., Muhuri, P. K., & Abraham, A. (2020). A bibliometric analysis and cutting-edge overview on fuzzy techniques in Big Data. Engineering Applications of Artificial Intelligence, 92, 103625.

Shukla, V., Prashar, S., & Pandiya, B. (2021). Is price a significant predictor of the churn behavior during the global pandemic? A predictive modeling on the telecom industry. Journal of Revenue and Pricing Management, 1-14.

Sigloch, S. (2018). Mobile Internet connectivity, exploring structural bottlenecks in Tamil Nadu using active Internet periphery measurements (Doctoral dissertation, Anglia Ruskin University).

Siguenza-Guzman, L., Saquicela, V., Avila-Ordóñez, E., Vandewalle, J., & Cattrysse, D. (2015). Literature review of data mining applications in academic libraries. The Journal of Academic Librarianship, 41(4), 499-510.

Simon, A. R., & Shaffer, S. L. (2001). Data warehousing and business intelligence for e-commerce. Elsevier.

Singh, N., & Singh, P. (2020). Stacking-based multi-objective evolutionary ensemble framework for prediction of diabetes mellitus. Biocybernetics and Biomedical Engineering, 40(1), 1-22.

Singh, P., Patil, Y., & Rale, V. (2019). Biosurfactant production: emerging trends and promising strategies. Journal of applied microbiology, 126(1), 2-13.

Sirichanya, C., & Kraisak, K. (2021). Semantic data mining in the information age: A systematic review. International Journal of Intelligent Systems, 36(8), 3880-3916.

Sivarajah, U., Kamal, M. M., Irani, Z., & Weerakkody, V. (2017). Critical analysis of Big Data challenges and analytical methods. Journal of business research, 70, 263-286.

Sivasankar, E., & Vijaya, J. (2019). Hybrid PPFCM-ANN model: an efficient system for customer churn prediction through probabilistic possibilistic fuzzy clustering and artificial neural network. Neural Computing and Applications, 31(11), 7181-7200.

Sjarif, N. N. A., Azmi, N. F., Sarkan, H. M., Sam, S. M., & Osman, M. Z. (2020, May). Predicting Churn: How Multilayer Perceptron Method Can Help with Customer Retention in Telecom Industry. In IOP Conference Series: Materials Science and Engineering (Vol. 864, No. 1, p. 012076). IOP Publishing.

Slof, D., Frasincar, F., & Matsiiako, V. (2021). A competing risks model based on latent Dirichlet Allocation for predicting churn reasons. Decision Support Systems, 146, 113541.

Sobol-Shikler, T. (2012). Inference of Co-occurring Classes: Multi-class and Multi-label Classification. Computational Intelligence Paradigms in Advanced Pattern Classification, 171-197.

Soni, J., Ansari, U., Sharma, D., & Soni, S. (2011). Predictive data mining for medical diagnosis: An overview of heart disease prediction. International Journal of Computer Applications, 17(8), 43-48.

Sowmya, R., & Suneetha, K. R. (2017, January). Data mining with big data. In 2017 11th International Conference on Intelligent Systems and Control (ISCO) (pp. 246-250). IEEE.

Sriramoju, S. B. (2017). Opportunities and security implications of big data mining. International Journal of Research in Science and Engineering, 3(6), 44-58.

Stone, M. D., & Woodcock, N. D. (2014). Interactive, direct and digital marketing: A future that depends on better use of business intelligence. Journal of research in interactive marketing.

Stripling, E., vanden Broucke, S., Antonio, K., Baesens, B., & Snoeck, M. (2018). Profit maximizing logistic model for customer churn prediction using genetic algorithms. Swarm and Evolutionary Computation, 40, 116-130.

Subhash, S., & Cudney, E. A. (2018). Gamified learning in higher education: A systematic review of the literature. Computers in human behavior, 87, 192-206.

Subotnik, R. F., Olszewski-Kubilius, P., & Worrell, F. C. (2011). Rethinking giftedness and gifted education: A proposed direction forward based on psychological science. Psychological science in the public interest, 12(1), 3-54.

Subramanian, P., & Palaniappan, S. (2016). Determinants of customer experience in the telecom industry using confirmatory factor analysis: An empirical study. International Journal of Conceptions on Computing and Information Technology, 4(4), 1-6.

Sudharsan, R., & Ganesh, E. N. (2022). A Swish RNN based customer churn prediction for the telecom industry with a novel feature selection strategy. Connection Science, 34(1), 1855-1876.

Suh, E., & Alhaery, M. (2016). Customer retention: Reducing online casino player churn through the application of predictive modeling. UNLV Gaming Research & Review Journal, 20(2), 6.

Sulaimon, O. S., Emmanuel, O. E., & Bilqis, B. B. (2016). Relevant drivers for customerschurn and retention decision in the Nigerian mobile telecommunication industry. Journal of Competitiveness, 8(3).

Sun, N., Zhang, J., Rimba, P., Gao, S., Zhang, L. Y., & Xiang, Y. (2018). Data-driven cybersecurity incident prediction: A survey. IEEE communications surveys & tutorials, 21(2), 1744-1772.

Sun, Y. (2017). The reasons why China’ s OBOR initiative goes digital. Aalborg University and University of International Relations, Denmark.

Sundaram, S., Kellnhofer, P., Li, Y., Zhu, J. Y., Torralba, A., & Matusik, W. (2019). Learning the signatures of the human grasp using a scalable tactile glove. Nature, 569(7758), 698-702.

Sundarkumar, G. G., Ravi, V., & Siddeshwar, V. (2015, December). One-class support vector machine based undersampling: Application to churn prediction and insurance fraud detection. In 2015 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC) (pp. 1-7). IEEE.

Sάnchez, B. U., & Asimakopoulos, G. (2012). Regulation and competition in the European mobile communications industry: An examination of the implementation of mobile number portability. Telecommunications Policy, 36(3), 187-196.

Talasila, V., Madhubabu, K., Madhubabu, K., Mahadasyam, M., Atchala, N., & Kande, L. (2020). The prediction of diseases using rough set theory with recurrent neural network in big data analytics. International Journal of Intelligent Engineering and Systems, 13(5), 10-18.

Tamaddoni Jahromi, A., Sepehri, M. M., Teimourpour, B., & Choobdar, S. (2010). Modeling customer churn in a non-contractual setting: the case of telecommunications service providers. Journal of Strategic Marketing, 18(7), 587-598.

Tamaddoni, A., Stakhovych, S., & Ewing, M. (2016). Comparing churn prediction techniques and assessing their performance: A contingent perspective. Journal of service research, 19(2), 123-141.

Taneja, S., Sharma, N., Oberoi, K., & Navoria, Y. (2016, August). Predicting trends in air pollution in Delhi using data mining. In 2016 1st India international conference on information processing (IICIP) (pp. 1-6). IEEE.

Tang, L., Thomas, L., Fletcher, M., Pan, J., & Marshall, A. (2014). Assessing the impact of derived behavior information on customer attrition in the financial service industry. European Journal of Operational Research, 236(2), 624-633.

Tang, Q., Xia, G., & Zhang, X. (2020). A hybrid classification model for churn prediction based on customer clustering. Journal of Intelligent & Fuzzy Systems, 39(1), 69-80.

Thakkar, A., & Chaudhari, K. (2021). Fusion in stock market prediction: a decade survey on the necessity, recent developments, and potential future directions. Information Fusion, 65, 95-107.

Thanuja, V., Venkateswarlu, B., & Anjaneyulu, G. S. G. N. (2011). Applications of data mining in customer relationship management. Journal of Computer and Mathematical Sciences, 2(3), 399-580.

Thomas, M. C., Zhu, W., & Romagnoli, J. A. (2018). Data mining and clustering in chemical process databases for monitoring and knowledge discovery. Journal of Process Control, 67, 160-175.

Thuraisingham, B. (2000). A primer for understanding and applying data mining. It Professional, 2(1), 28-31.

Tianyuan, Z. (2018). Telecom customer segmentation and precise package design by using data mining (Doctoral dissertation).

Tien, J. M. (2013). Big data: Unleashing information. Journal of Systems Science and Systems Engineering, 22(2), 127-151.

Tiwari, D., Kumar, A., & Tripathi, A. (2019). Virtual doctor.

Tolstoy, D., Nordman, E. R., Hånell, S. M., & Özbek, N. (2021). The development of international e-commerce in retail SMEs: An effectuation perspective. Journal of World Business, 56(3), 101165.

Torkzadeh, G., Chang, J. C. J., & Hansen, G. W. (2006). Identifying issues in customer relationship management at Merck-Medco. Decision Support Systems, 42(2), 1116-1130.

Tosun, A., Bener, A., Turhan, B., & Menzies, T. (2010). Practical considerations in deploying statistical methods for defect prediction: A case study within the Turkish telecommunications industry. Information and Software Technology, 52(11), 1242-1257.

Tsai, C. F., & Chen, M. Y. (2010). Variable selection by association rules for customer churn prediction of multimedia on demand. Expert Systems with Applications, 37(3), 2006-2015.

Tsai, C. F., & Lu, Y. H. (2010). Data mining techniques in customer churn prediction. Recent Patents on Computer Science, 3(1), 28-32.

Tsai, H. H. (2011). Research trends analysis by comparing data mining and customer relationship management through bibliometric methodology. Scientometrics, 87(3), 425-450.

Tsai, H. H. (2012). Global data mining: An empirical study of current trends, future forecasts and technology diffusions. Expert systems with applications, 39(9), 8172-8181.

Tsai, H. H. (2013). Knowledge management vs. data mining: Research trend, forecast and citation approach. Expert Systems with Applications, 40(8), 3160-3173.

Tufféry, S. (2011). Data mining and statistics for decision making. John Wiley & Sons.

Ucbasaran, D., Shepherd, D. A., Lockett, A., & Lyon, S. J. (2013). Life after business failure: The process and consequences of business failure for entrepreneurs. Journal of management, 39(1), 163-202.

Uddin, M., & Rahman, A. A. (2012). Energy efficiency and low carbon enabler green IT framework for data centers considering green metrics. Renewable and Sustainable Energy Reviews, 16(6), 4078-4094.

Udugama, I. A., Gargalo, C. L., Yamashita, Y., Taube, M. A., Palazoglu, A., Young, B. R., ... & Bayer, C. (2020). The role of big data in industrial (bio) chemical process operations. Industrial & Engineering Chemistry Research, 59(34), 15283-15297.

Ulas, B. T., Imer, S., Kalayci, T. A., & Asan, U. (2023). Modeling Customer Churn Behavior in E-commerce Using Bayesian Networks. In Global Joint Conference on Industrial Engineering and Its Application Areas (pp. 283-300). Springer, Cham.

Ullah, I., Liu, K., Yamamoto, T., Zahid, M., & Jamal, A. (2022). Prediction of electric vehicle charging duration time using ensemble machine learning algorithm and Shapley additive explanations. International Journal of Energy Research, 46(11), 15211-15230.

Ullah, I., Raza, B., Malik, A. K., Imran, M., Islam, S. U., & Kim, S. W. (2019). A churn prediction model using random forest: analysis of machine learning techniques for churn prediction and factor identification in telecom sector. IEEE access, 7, 60134-60149.

Umayaparvathi, V., & Iyakutti, K. (2012). Applications of data mining techniques in telecom churn prediction. International Journal of Computer Applications, 42(20), 5-9.

Vafeiadis, T., Diamantaras, K. I., Sarigiannidis, G., & Chatzisavvas, K. C. (2015). A comparison of machine learning techniques for customer churn prediction. Simulation Modelling Practice and Theory, 55, 1-9.

Valluri, C., Raju, S., & Patil, V. H. (2022). Customer determinants of used auto loan churn: comparing predictive performance using machine learning techniques. Journal of Marketing Analytics, 10(3), 279-296.

Van Dijck, J., & Poell, T. (2013). Understanding social media logic. Media and communication, 1(1), 2-14.

Van Iwaarden, J., Van der Wiele, T., Ball, L., & Millen, R. (2003). Applying SERVQUAL to web sites: An exploratory study. International Journal of Quality & Reliability Management.

Van Nguyen, T., Zhou, L., Chong, A. Y. L., Li, B., & Pu, X. (2020). Predicting customer demand for remanufactured products: A data-mining approach. European Journal of Operational Research, 281(3), 543-558.

Vassakis, K., Petrakis, E., & Kopanakis, I. (2018). Big data analytics: applications, prospects and challenges. Mobile big data, 3-20.

Vaughan, L., & Chen, Y. (2015). Data mining from web search queries: A comparison of google trends and baidu index. Journal of the Association for Information Science and Technology, 66(1), 13-22

Veningston, K., Rao, P. V., Selvan, C., & Ronalda, M. (2022). Investigation on Customer Churn Prediction Using Machine Learning Techniques. In Proceedings of International Conference on Data Science and Applications (pp. 109-119). Springer, Singapore.

Verbeke, W., Dejaeger, K., Martens, D., Hur, J., & Baesens, B. (2012). New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. European journal of operational research, 218(1), 211-229.

Verbeke, W., Martens, D., & Baesens, B. (2014). Social network analysis for customer churn prediction. Applied Soft Computing, 14, 431-446.

Verbeke, W., Martens, D., Mues, C., & Baesens, B. (2011). Building comprehensible customer churn prediction models with advanced rule induction techniques. Expert systems with applications, 38(3), 2354-2364.

Verhoef, P. C., Venkatesan, R., McAlister, L., Malthouse, E. C., Krafft, M., & Ganesan, S. (2010). CRM in data-rich multichannel retailing environments: a review and future research directions. Journal of interactive marketing, 24(2), 121-137.

Vijaya, J., Srimathi, S., Karthikeyan, S., & Siddarth, S. (2020). An Improved Telecommunication Churn Prediction System by PPFCM Clustering Hybrid Model. In Recent Advances in Computer Based Systems, Processes and Applications (pp. 169-175). CRC Press.

Vijayaraman, B., & Chellappa, S. (2016). Analysing Customer Churn and Customer Attitude in Telecom Market. Asian Journal of Research in Social Sciences and Humanities, 6(6), 362-374.

Visser, J., Nemoto, T., & Browne, M. (2014). Home delivery and the impacts on urban freight transport: A review. Procedia-social and behavioral sciences, 125, 15-27.

Vo, N. N., Liu, S., Li, X., & Xu, G. (2021). Leveraging unstructured call log data for customer churn prediction. Knowledge-Based Systems, 212, 106586.

Vu, K. M. (2013). Information and communication technology (ICT) and Singapore’s economic growth. Information Economics and policy, 25(4), 284-300.

Walter, L., Denter, N. M., & Kebel, J. (2022). A review on digitalization trends in patent information databases and interrogation tools. World Patent Information, 69, 102107.

Wang, C. H., & Fong, H. Y. (2016). Integrating fuzzy Kano model with importance-performance analysis to identify the key determinants of customer retention for airline services. Journal of Industrial and Production Engineering, 33(7), 450-458.

Wang, G., Hao, J., Ma, J., & Huang, L. (2010). A new approach to intrusion detection using Artificial Neural Networks and fuzzy clustering. Expert systems with applications, 37(9), 6225-6232.

Wang, L., Liu, Y., & Wu, J. (2018). Research on financial advertisement personalised recommendation method based on customer segmentation. International Journal of Wireless and Mobile Computing, 14(1), 97-101.

Wang, Q. F., Xu, M., & Hussain, A. (2019). Large-scale ensemble model for customer churn prediction in search ads. Cognitive Computation, 11(2), 262-270.

Wang, Q., & Sawhney, S. (2014, October). VeCure: A practical security framework to protect the CAN bus of vehicles. In 2014 International Conference on the Internet of Things (IOT) (pp. 13-18). IEEE.

Wang, Y., Kung, L., & Byrd, T. A. (2018). Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations. Technological forecasting and social change, 126, 3-13.

Wang, Z., Su, Y., Jin, S., Shen, W., Ren, J., Zhang, X., & Clark, J. H. (2020). A novel unambiguous strategy of molecular feature extraction in machine learning assisted predictive models for environmental properties. Green Chemistry, 22(12), 3867-3876.

Wassouf, W. N., Alkhatib, R., Salloum, K., & Balloul, S. (2020). Predictive analytics using big data for increased customer loyalty: Syriatel Telecom Company case study. Journal of Big Data, 7(1), 1-24.

Weiss, G. (2010). Data mining in the telecommunications industry. In Networking and Telecommunications: Concepts, Methodologies, Tools, and Applications (pp. 194-201). IGI Global.

Wiemer, H., Drowatzky, L., & Ihlenfeldt, S. (2019). Data mining methodology for engineering applications (DMME)—A holistic extension to the CRISP-DM model. Applied Sciences, 9(12), 2407.

Williams, T., Klakegg, O. J., Walker, D. H., Andersen, B., & Magnussen, O. M. (2012). Identifying and acting on early warning signs in complex projects. Project Management Journal, 43(2), 37-53.

Williamson, P. J. (2010). Cost innovation: preparing for a ‘value-for-money’ revolution. Long Range Planning, 43(2-3), 343-353.

Wilson, K. B., Bhakoo, V., & Samson, D. (2018). Crowdsourcing: A contemporary form of project management with linkages to open innovation and novel operations. International Journal of Operations & Production Management.

Winer, R. S. (2001). A framework for customer relationship management. California management review, 43(4), 89-105.

Wlodarczak, P., Soar, J., & Ally, M. (2015, October). Multimedia data mining using deep learning. In 2015 Fifth International Conference on Digital Information Processing and Communications (ICDIPC) (pp. 190-196). IEEE.

Wood, D. A. (2021). Prediction and data mining of burned areas of forest fires: Optimized data matching and mining algorithm provides valuable insight. Artificial Intelligence in Agriculture, 5, 24-42.

Wood, D. E., & Salzberg, S. L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome biology, 15(3), 1-12.

Wu, X., & Meng, S. (2016, June). E-commerce customer churn prediction based on improved SMOTE and AdaBoost. In 2016 13th International conference on service systems and service management (ICSSSM) (pp. 1-5). IEEE.

Wu, X., Zhu, X., Wu, G. Q., & Ding, W. (2013). Data mining with big data. IEEE transactions on knowledge and data engineering, 26(1), 97-107.

Xevelonakis, E., & Som, P. (2012). The impact of social network-based segmentation on customer loyalty in the telecommunication industry. Journal of Database Marketing & Customer Strategy Management, 19(2), 98-106.

Xia, Y., Liu, C., Li, Y., & Liu, N. (2017). A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring. Expert systems with applications, 78, 225-241.

Xie, Y., Li, X., Ngai, E. W. T., & Ying, W. (2009). Customer churn prediction using improved balanced random forests. Expert Systems with Applications, 36(3), 5445-5449.

Xu, J. D., Benbasat, I., & Cenfetelli, R. (2011). The effects of service and consumer product knowledge on online customer loyalty. Journal of the Association for Information Systems, 12(11), 1.

Xu, T., Ma, Y., & Kim, K. (2021). Telecom churn prediction system based on ensemble learning using feature grouping. Applied Sciences, 11(11), 4742.

Xue, S., & Hong, Y. (2016). Earnings management, corporate governance and expense stickiness. China Journal of Accounting Research, 9(1), 41-58.

Yadav, S., Jain, A., & Singh, D. (2018, December). Early prediction of employee attrition using data mining techniques. In 2018 IEEE 8th International Advance Computing Conference (IACC) (pp. 349-354). IEEE.

Yan, H., Yang, N., Peng, Y., & Ren, Y. (2020). Data mining in the construction industry: Present status, opportunities, and future trends. Automation in Construction, 119, 103331.

Yang J T, Wan C S, Fu Y J. Qualitative examination of employee turnover and retention strategies in international tourist hotels in Taiwan[J]. International journal of hospitality management, 2012, 31(3): 837-848.

Yao, Y., Wang, J., Xie, M., Hu, L., & Wang, J. (2020). A new approach for fault diagnosis with full-scope simulator based on state information imaging in nuclear power plant. Annals of Nuclear Energy, 141, 107274.

Yazdi, M. F., Kamel, S. R., Chabok, S. J. M., & Kheirabadi, M. (2020). Flight delay prediction based on deep learning and Levenberg-Marquart algorithm. Journal of Big Data, 7(1), 1-28.

Yoo, M., & Bai, B. (2013). Customer loyalty marketing research: A comparative approach between hospitality and business journals. International Journal of Hospitality Management, 33, 166-177.

Yu, Q., Yen, D. A., Barnes, B. R., & Huang, Y. A. (2019). Enhancing firm performance through internal market orientation and employee organizational commitment. The International Journal of Human Resource Management, 30(6), 964-987.

Yu, Z. J., Haghighat, F., & Fung, B. C. (2016). Advances and challenges in building engineering and data mining applications for energy-efficient communities. Sustainable Cities and Society, 25, 33-38.

Zahid, H., Mahmood, T., Morshed, A., & Sellis, T. (2019). Big data analytics in telecommunications: literature review and architecture recommendations. IEEE/CAA Journal of Automatica Sinica, 7(1), 18-38.

Zaki, M. J., Ogihara, M., Parthasarathy, S., & Li, W. (1996, January). Parallel data mining for association rules on shared-memory multi-processors. In Supercomputing'96: Proceedings of the 1996 ACM/IEEE Conference on Supercomputing (pp. 43-43). IEEE.

Zaslavsky, A., Perera, C., & Georgakopoulos, D. (2013). Sensing as a service and big data. arXiv preprint arXiv:1301.0159.

Zeithaml, V. A. (2000). Service quality, profitability, and the economic worth of customers: what we know and what we need to learn. Journal of the academy of marketing science, 28(1), 67-85.

Zeng, D. (2015). Crystal Balls, Statistics, Big Data, and Psychohistory: Predictive Analytics and Beyond. IEEE Intelligent Systems, 30(02), 2-4.

Zhang, J. Z., & Chang, C. W. (2021). Consumer dynamics: Theories, methods, and emerging directions. Journal of the Academy of Marketing Science, 49(1), 166-196.

Zhang, J., & Liang, X. J. (2011). Business ecosystem strategies of mobile network operators in the 3G era: The case of China Mobile. Telecommunications policy, 35(2), 156-171.

Zhang, N., Li, M., & Lou, W. (2011, June). Distributed data mining with differential privacy. In 2011 IEEE international conference on Communications (ICC) (pp. 1-5). IEEE.

Zhang, T. J., Huang, X. H., Tang, J. F., & Luo, X. G. (2011). Case study on cluster analysis of the telecom customers based on consumers' behavior. In 2011 IEEE 18th International Conference on Industrial Engineering and Engineering Management (pp. 1358-1362). IEEE.

Zhang, X., Zhu, J., Xu, S., & Wan, Y. (2012). Predicting customer churn through interpersonal influence. Knowledge-Based Systems, 28, 97-104.

Zhao, L., Gao, Q., Dong, X., Dong, A., & Dong, X. (2017). K-local maximum margin feature extraction algorithm for churn prediction in telecom. Cluster Computing, 20(2), 1401-1409.

Zhao, M., Zeng, Q., Chang, M., Tong, Q., & Su, J. (2021). A prediction model of customer churn considering customer value: an empirical research of telecom industry in China. Discrete Dynamics in Nature and Society, 2021.

Zhao, T., Zhang, X., & Wang, S. (2021, March). Graphsmote: Imbalanced node classification on graphs with graph neural networks. In Proceedings of the 14th ACM international conference on web search and data mining (pp. 833-841).

Zhao, Y., Hryniewicki, M. K., Cheng, F., Fu, B., & Zhu, X. (2018, September). Employee turnover prediction with machine learning: A reliable approach. In Proceedings of SAI intelligent systems conference (pp. 737-758). Springer, Cham.

Ziegler, A., & König, I. R. (2014). Mining data with random forests: current options for real‐world applications. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 4(1), 55-63.

Zolfaghar, K., & Aghaie, A. (2012). A syntactical approach for interpersonal trust prediction in social web applications: Combining contextual and structural data. Knowledge-Based Systems, 26, 93-102.

Óskarsdóttir, M., Bravo, C., Verbeke, W., Sarraute, C., Baesens, B., & Vanthienen, J. (2016, August). A comparative study of social network classifiers for predicting churn in the telecommunication industry. In 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM) (pp. 1151-1158). IEEE.

Óskarsdóttir, M., Bravo, C., Verbeke, W., Sarraute, C., Baesens, B., & Vanthienen, J. (2017). Social network analytics for churn prediction in telco: Model building, evaluation and network architecture. Expert Systems with Applications, 85, 204-220