INSTITUTO
UNIVERSITÁRIO
DE LISBOA

Department of Quantitative Methods for Management and Economics/ Department of Information Science and Technology

Addressing Data Imbalance in Customer Churn Prediction: A Novel Approach for Telecommunications Companies

Gonçalo Xavier Cota Oliveira

Master in Data Science

Supervisor:
PhD, José Manuel Gonçalves Dias, Full Professor
ISCTE-IUL

November, 2023

BUSINESS SCHOOL

TECNOLOGIAS E ARQUITETURA

Department of Quantitative Methods for Management and Economics/ Department of Information Science and Technology

Addressing Data Imbalance in Customer Churn Prediction: A Novel Approach for Telecommunications Companies

Gonçalo Xavier Cota Oliveira

Master in Data Science

Supervisor:
PhD, José Manuel Gonçalves Dias, Full Professor
ISCTE-IUL

November, 2023

Acknowledgments

To my advisor, professor José Dias, I am grateful for your patient and dedicated guidance. Your support and guidance were fundamental to the development of my work and to my growth as a researcher.

To NOWO and all my colleagues, you were pivotal in every phase of this process. The collaboration and support I found among all of you were essential to the success of this journey and to the enrichment of my professional experience.

To all the teachers who have crossed my academic path, I am grateful for shaping and nurturing my intellectual curiosity in crucial moments.

To my friends and family, I appreciate your support and encouragement throughout this journey. And to Vitor, I am thankful for the love, companionship, and constant support.

Resumo

A indústria de telecomunicações tem conhecido um rápido crescimento e transformação, o que resultou numa intensificação da concorrência e na necessidade das empresas de se adaptarem às constantes mudanças nas preferências dos consumidores. Como tal, as empresas intensificaram estratégias de compreensão do ciclo de vida do cliente.

A gestão eficiente da taxa de *churn* é reconhecida como uma vantagem competitiva crucial no setor das telecomunicações.

Através da revisão de literatura, observou-se que a investigação existente relativamente à predição do *churn* (métrica que mede a percentagem de clientes que cancelam os seus serviços em um determinado período de tempo) abrange principalmente cliente com serviços móveis, deixando uma lacuna em relação aos clientes fixos. Este estudo tem como objetivo preencher essa lacuna através da construção de um modelo preditivo de *churn* dos clientes numa empresa de telecomunicações portuguesa, com foco nos serviços quadruple-play – que contemplam televisão, internet, telefone e telemóvel.

Tendo em conta a natureza da base de dados, o estudo explora estratégias para abordar desequilíbrios de classes na modelação preditiva através da introdução de técnicas *Tomek's Links* e da *Synthetic Minority Oversampling Technique* (SMOTE), que melhoraram as métricas utilizadas para aferir a qualidade do modelo. O algoritmo CatBoost com a utilização de SMOTE obteve os melhores resultados dos critérios propostos neste conjunto de dados de clientes. A afinação dos hiperparâmetros resultou numa melhoria no desempenho deste algoritmo.

A análise SHAP revelou que as visitas a lojas físicas reduzem a taxa de *churn*, enquanto que o contacto com o *call center* contribui para a taxa de *churn* destes clientes.

Palavras-chave: *Churn*; telecomunicações; previsão; modelação; desequilíbrio de classes

Abstract

The telecommunications industry has experienced rapid growth and transformation, resulting in intensified competition and the need for companies to adapt to constant changes in consumer preferences. As such, companies have intensified their strategies to understand the customer lifecycle.

Efficient churn management is recognized as a crucial competitive advantage in the telecommunications sector.

Through literature review, it was observed that existing research on churn prediction (a metric that measures the percentage of customers who cancel their services within a certain period of time) mainly covers customers with mobile services, leaving a gap regarding fixed-line customers. This study aims to fill this gap by constructing a predictive churn model for customers in a Portuguese telecommunications company, focusing on quadruple-play services - which include television, internet, telephone, and mobile.

Considering the nature of the database, the study explores strategies to address class imbalances in predictive modeling through the introduction of Tomek's Links and Synthetic Minority Oversampling Technique (SMOTE) techniques, which improved the metrics used to assess model quality. The CatBoost algorithm using SMOTE achieved the best results in the proposed criteria for this customer dataset. The fine-tuning of the hyperparameters resulted in an improvement in the performance of this algorithm.

The SHAP analysis revealed that visits to physical stores reduce churn rate, while contact with the call center contributes to the churn rate of these customers.

Keywords: Churn; telecommunications; prediction; modeling; class imbalance

Index

List of Tables

List of Figures

Abbreviations and Acronyms

5G: Fifth Generation of wireless technology

ADA: Adaboost

ANACOM: The national regulatory authority for communications in Portugal

AUC: Area Under the Curve

AW: Activity Window

CRM: Customer Relationship Management

CRISP-DM: Cross Industry Standard Process for Data Mining

CW: Churn Window

DT: Decision Tree

ET: Extra Trees

GBC: Gradient Boosting Classifier

LIGHT GBM: Light Gradient Boosting Machine

GDPR: General Data Protection Regulation

KNN: K-Nearest Neighbors

LDA: Linear Discriminant Analysis

LR: Logistic Regression

NB: Naïve Bayes

OW: Observation Window

PCA: Principal Component Analysis

QDA: Quadratic Discriminant Analysis

RF: Random Forest

SHaP: SHapley Additive exPlanations

SMOTE: Synthetic Minority Oversampling Technique

SVM: Support Vector Machine

TCI: Telecommunications Industry

TT: Training Time

VOC: Voice of the Customer

CHAPTER 1

Introduction

The telecommunications industry is rapidly evolving, driven by innovations and cutting-edge technologies. According to theBuddeComm Intelligence Report (Wansink, 2020), active mobile subscriptions reached 7.7 billion in 2020, a remarkable increase from 3.3 billion in 2015. This growth showcases the sector's transformative impact, connecting billions worldwide.

Intense competition between telecom companies is fueled by innovation and technology integration, influencing consumer preferences. To maintain competitiveness and meet market demands, telecom companies must adeptly navigate these dynamics, enhance their services, and innovate to satisfy consumer needs (Moussaoui et al., 2022).

Customer loyalty is essential for maximizing profitability in every industry, because acquiring new customers can be costly, while retaining existing ones is less expensive over time (Fang, 2021; Ullah et al., 2019).

Customers go through different phases in their relationship with a company, which leads to changes in their behavior (Voorhees et al., 2017). In addition, businesses must prioritize the cultivation of customer loyalty to increase engagement. This objective can be met by implementing a strategic care plan that successfully tackles perceived problems in service and effectively retains dissatisfied customers.

Customer churn, also referred to as customer attrition or turnover, is a key metric used to measure the customer loss or disengagement from a business or service. It is calculated by dividing the number of customers who terminate their association with the company within a specific timeframe by the total count of active customers during that same period.

Customer churn, in simpler terms, refers to the defection of customers who stop subscribing a company's products or services. It is a crucial measure for businesses to monitor as it directly affects revenue and profitability. By actively tracking churn rates, companies can identify potential problems and implement proactive strategies to retain customers.

Churn can occur in two ways: voluntarily, when a customer decides to switch to a competing brand; and involuntarily, when a customer gets disconnected typically for reasons that go beyond the control of the customer or/and the company. Involuntary churn can be, for example, when customers fail to make timely payments for their services, this could be due to financial difficulties or other billing-related problems.

Consumer loyalty to a company is influenced by numerous factors, both internal and external. Internal elements include pricing, quality of customer service, overall service experience (Muneeb et al., 2019), lack of desired features, and customer characteristics (Sharma & Rajan, 2017). In the telecommunications industry specifically, reasons for customers switching providers may include poor network quality, insufficient customer care, high fees, and attractive features offered by competitors.

External factors are diverse and impact the decision-making process for customers switching service providers. These factors include aggressive competition with comparable offerings with more competitive prices, new companies that more quickly adapt to market trends (predominantly the rapid technological developments) and geographical considerations like coverage and network quality in specific areas. Additionally, economic conditions, regulatory changes, peer-pressure, and demographics also play a role in influencing customer decisions to switch providers.

Churn can also be influenced also by both psychological and physical switching costs. Psychological switching costs refer to the perceived challenges of moving to a new provider, such as dealing with the inconvenience of transitioning or worries about making mistakes. On the other hand, physical switching costs involve actual expenses incurred during provider switches like termination fees or having to buy new devices.

Competition within a company can significantly impact customer satisfaction and loyalty. The wide range of products and services offered may increase the risk of customer churn, as customers may perceive differences in quality and prefer alternatives provided by the company itself. This awareness of better options further fuels dissatisfaction with current offerings. Ultimately, internal competition plays a crucial role in shaping customer perceptions and their likelihood of continuing to do business with the company (Feick & Price, 1987).

In 2017, Aditya Kapoor conducted a study on customer retention and churn rate in the US telecommunications sector. The report reveals a churn rate of 1.9% among the four major carriers, serving a total of 100 million customers. On average, customers switch after approximately 19 months, considering a typical lifespan of 52 months. Notably, each churn case results in a revenue loss of over $1,100, calculated by multiplying $34 by 33 months.

According to Bhikha (2019), customer churn is a major issue for telecommunications companies in the UK, with approximately 75% of new customers being individuals who have switched from competitors. This underscores the high level of customer sensitivity within this industry.

Portuguese telecommunications companies are still struggling with a high annual churn rate of 18% (ANACOM, 2017), making the telecommunications market in Portugal unstable. Portuguese telecom companies use loyalty programs to keep customers and reduce costs. These programs typically last for up to 24 months, the lock-in period, during which clients are required to pay a fee if they cancel their contracts early. To appeal consumers, companies offer incentives like free installation fees and/or subscriptions to certain services.

Accurately predicting customer churn is crucial for improving retention strategies. By understanding customer data, organizations can anticipate behavior and enhance the customer journey. Comprehending customer churn is a multifaceted task that demands a thoughtful and comprehensive approach, taking into account the unique circumstances and industry of each company. While certain studies aim to identify common churn triggers, others underscore the significance of contextual factors. Employing a customized strategy is paramount to effectively tackling customer churn.

Customer interactions generate data that businesses can leverage to improve marketing and financial results, as well as customer satisfaction and engagement with products and services (Kotler & Keller, 2012; Holmlund, 2020). Telecom companies have access to extensive customer data, including consumption records, contact information, socio-economic profiles, financial details, and other important factors. This data helps them identify customers who may be at risk of churn.

Reactive churn management focuses on re-engaging customers who have already churned, while proactive churn management aims to prevent at-risk customers from leaving. Research suggests that proactive churn management is a superior strategy, allowing operators to implement retention measures before customers migrate to competitors. By proactively identifying and addressing potential issues, operators can develop effective retention strategies and reduce churn (Burez & Van den Poel, 2009; Blattberg, Kim, & Neslin, 2008).

To effectively monetize efforts to retain customers, organizations should focus on targeting the right consumers and addressing the root causes of churn through a holistic understanding. Prioritizing model simplicity enables the application of personalized marketing campaigns at small segments or even individual levels. Applying micro-targeting and geo-marketing techniques can lift customer satisfaction and drive business success (Lies, 2023).

By the end of 2022, the Portuguese electronic communications market was dominated by four major providers: MEO, NOS, Vodafone, and NOWO. These providers offer a range of services bundled under different categories, such as single play (individual services), double play (2P), triple play (3P), quadruple play (4P), or quintuple play (5P), depending on the number

of integrated services included. These packages offer varying combinations of television, telephone, internet, mobile, mobile internet, home security, and energy management services.

In this dynamic telecommunications market, telecom companies are now prioritizing customer retention over acquisition. The opening up of fixed and mobile networks, along with the emergence of new telecom operators in Portugal, has intensified competition, leading to improved service quality and reduced prices.

NOWO emerged in September 2016 through the rebranding of the Cabovisão brand, which was established in September 1993 and began operations in the same year. NOWO claims an extensive fiber optic network, crossing over 70 municipalities and 200 parishes, with a total length of approximately 14,000 km. It can supply with services of over 900 thousand households. The operator has wide coverage in central, interior, and southern regions of Portugal, with stores located countrywide and a dedicated call center. NOWO is well-known for offering the most competitively priced packages in the Portuguese market and stands as the fourth-largest operator in Portugal.

In early 2022, Vodafone Portugal entered into an agreement to acquire Cabonitel, the parent company of NOWO. The transaction is expected to be finalized in 2023/2024, pending approval from the Portuguese Competition Authority (AdC).

Figure 1, based on data from ANACOM (2022), illustrates a consistent deterioration in the market share of NOWO customers, the company that provided the data for this churn analysis. This highlights the vital need to strengthen the company's customer retention strategy, particularly in light of these new challenges.

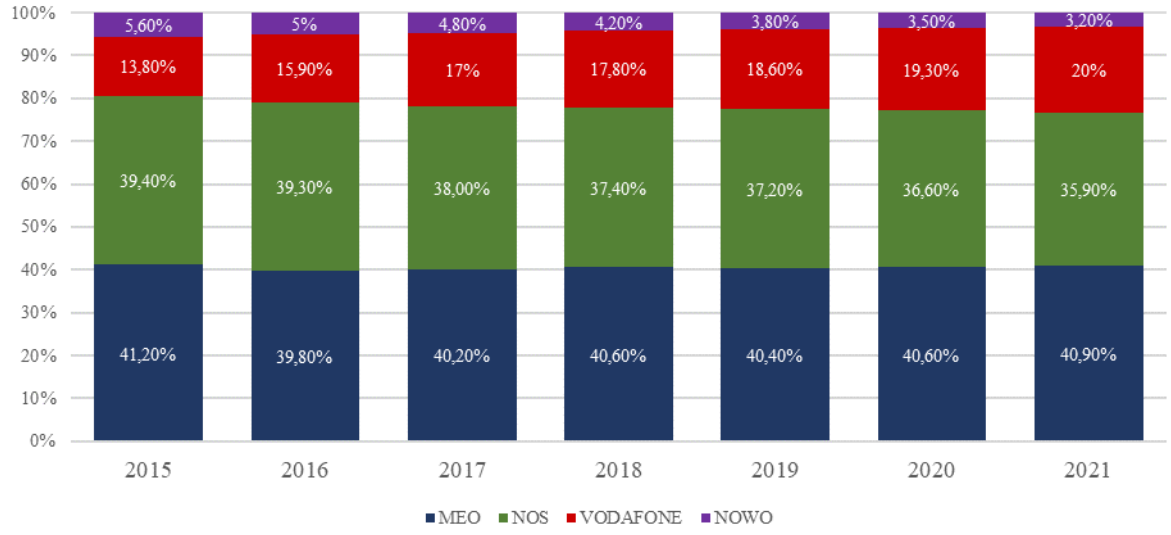

Figure 1. Portugal's Telecommunications Companies Market Share Analysis (2015-2021)

The bundling of multiple communication services has become a prevalent strategy in the telecommunications industry, particularly the four-play bundling, which has gained traction in several countries such as Portugal, France, and Spain. The most well-known and traditional 4P commercial offer available from Portuguese operators consists of television, internet, fixed-line phone, and mobile phone.

Consumers have responded positively to these bundled offerings due to the simplification of service management and the potential reduction in overall costs. For example, in France, approximately 42% of households with fixed broadband opted for bundled mobile services by the end of 2012, a number that skyrocketed to 71% by the end of 2019. Similarly, in Spain, the adoption of bundled services among fixed broadband households increased from 21% in 2012 to 82% by 2019 (S&P Global Market Intelligence, 2021).

According to ANACOM (2023), operators are progressively offering bundled services as many of these providers now use a joint broadband infrastructure to deliver fixed-line telephony, broadband internet, and pay-TV services via fiber-optic networks. Most Portuguese households have internet and broadband connections, often bundled with other services. By the end of the first quarter in 2022, about 89.6% of households were using these services, generating a revenue of approximately 4.46 billion euros.

The most well-known and traditional 4P commercial offer available from Portuguese operators consists of television, internet, fixed-line phone, and mobile phone. As per the same ANACOM 2023's survey, 12.7% of residential electronic communications service subscribers chose quadruple play subscriptions, which is an increase of 0.7 percentage points from the previous period and stands out as the most substantial growth among all the services offered.

This study aims to utilize supervised machine learning techniques in order to develop and evaluate a predictive model that addresses customer churn in NOWO, a telecommunications company based in Portugal. Specifically, the focus is on the four-play customer segment. The research aims to establish a robust framework for understanding predictive and descriptive analytics techniques that are relevant to churn prediction in the telecom industry. It explores the underlying processes of data manipulation, pre-processing, cleaning, and precise churn definition, providing a comparative analysis based on real-world data.

The methodological approach includes a case study using actual data, driven by a recognized gap in knowledge regarding the implementation of descriptive analysis techniques tailored for effective churn prediction in the telecom sector. The primary objective is the construction of an advanced predictive model. This phase entails sophisticated feature

engineering, meticulous feature selection, and a thorough comparison of diverse algorithms tailored for accurate churn prediction, scrutinizing their performance metrics.

Additionally, the study emphasizes the strategic significance of Customer Relationship Management (CRM), its evolutionary trajectory, and the strategic utilization of loyalty programs as indispensable tools within the analytical framework of CRM. Furthermore, the study acknowledges the inherent scope and limitations and proposes potential avenues for further research, extending beyond the current research scope, to enhance the efficacy of churn prediction methodologies.

The second goal is to delve into complex consumer behavioral patterns, aiming to detect subtle indicators and triggers that precede churn events, as indicated by the outputs of the models. By analyzing variables such as the frequency of product/service usage, engagement with promotional offers, and response to customer service interactions, this study aims to uncover nuanced behavioral cues that may indicate potential churn. This exploration into detailed behavioral nuances intends to enhance the predictive model by gaining a comprehensive understanding of customer interactions and preferences, thereby contributing to a more refined and accurate churn prediction framework for NOWO.

As a result, the following research question was formulated:

> *RQ: How to develop a machine learning model to predict residential customer churn in the fixed communication market in highly complex products?*

from which we can derive the following two sub-research questions:

> *SRQ1: Which machine learning algorithm is the most appropriate for solving this particular prediction problem?*
>
> *SRQ2: Which variables hold the utmost significance in forecasting customer churn for quadruple-play customers?*

The dissertation is structured into six chapters. In general, this chapter offers a comprehensive glimpse into the telecommunications industry's landscape, the dynamics of customer behavior, and the significance of grasping churn for enhancing strategies to retain

customers.

Chapter 2 offers an overview of previous approaches to churn prediction and technical information on commonly used methods. It consolidates existing knowledge and techniques in the field. Chapter 3 focus on explaining and tackling class imbalance in data by employing specialized methodologies and strategies. It encompasses a wide range of techniques, including oversampling, undersampling, and harnessing the power of robust algorithms.

Chapter 4 delves into the specific methodologies tailored for the churn model within the CRISP-DM framework. It covers the metrics, utilized models, and details the strategies developed to address this issue. Additionally, it offers a comprehensive analysis of the obtained results, providing insights into the performance and effectiveness of the applied models. This chapter offers effective methods to tackle this issue.

Chapter 5 centers on the evaluation of churn prediction model outcomes. It thoroughly examines the results derived from the applied methodologies, providing valuable insights into the models' performance and effectiveness. The chapter offers a comprehensive analysis and interpretation of the findings, enhancing the understanding of the models' efficacy.

The final chapter addresses any constraints or limitations encountered during the research process. It also explores potential areas for development and improvement, suggesting avenues for further research.

By organizing the dissertation into these chapters, we aim to provide a logical and comprehensive framework for understanding and analyzing churn prediction in the given context, using real customer data.

CHAPTER 2

Literature Review

Our research primarily focuses on predicting customer churn within the telecommunications industry. However, it is crucial to recognize that churn behavior is prevalent across various sectors. To gain a comprehensive understanding of customer churn within the fixed telecom industry, we delve into churn behaviors within other industries as well. This section explores existing predictive models for churn across diverse sectors, while also examining explanatory studies that shed light on specific factors within the telecommunications field, covering different scenarios, methodologies, algorithms, measurements, and limitations.

While this dissertation does not cover all the research in the field, it offers a comprehensive summary of significant works. These studies have made substantial contributions to the advancement of churn prediction, providing valuable insights and methodologies for effectively addressing this challenge in the telecom industry. However, it is worth noting that comparing these studies can be challenging due to variations in the adopted modeling approaches. Churn prediction has been addressed in diverse domains, including telecommunications (Huang et al., 2012; Lu et al., 2014; Huang et al., 2015), banking (Xie et al., 2009; Larivière & Van den Poel, 2005), subscription services (Coussement & Van den Poel, 2008), game businesses (Castro & Tsuzuki, 2015), and retailing (Buckinx & Van den Poel, 2005).

The pursuit of understanding churn has ignited a multitude of efforts. Generally, research in this field centers on two main goals: pinpointing the pivotal factors influencing customer churn and creating predictive models—a pursuit that retains significant relevance (Coussement & Van den Poel, 2009). Churn prediction encompasses two main types of models: binary and survival models (Berry and Linoff, 2004). Binary models aim to predict whether a customer will churn within a predefined timeframe, typically a fixed interval. On the other hand, survival models estimate the duration until a customer churns (Wassouf et al., 2020).

Churn prediction models for binary classification are motivated by the quest for a learning model that strikes an optimal balance between robust predictive performance and comprehensibility. This pursuit has led to the exploration of various methodologies, each with its own set of trade-offs.

In a study comparing logistic regression (LR) and decision tree (DT) algorithms for user churn prediction, Ballings and Van den Poel (2012) showed that the choice of time window

significantly influenced the predictive accuracy of both LR and DT models. Moreover, DT models generally outperformed LR models, especially when considering longer time windows.

Vafeiadis et al. (2015) compared Support Vector Machines (SVM), Logistic Regression (LR), Artificial Neural Networks (ANN), Decision Trees (DT), Naïve Bayes (NB). They found that ANNs with backpropagation achieved the highest accuracy without boosting, while NB and LR consistently demonstrated the lowest accuracy across both studies.

Traditional machine learning methods often encounter difficulties when dealing with complex and multifaceted data sets. Ensemble learning, a well-established technique in the field, effectively tackles these challenges by combining predictions from multiple weaker models. This results in a significant improvement in overall model performance (Sagi & Rokach, 2018). Particularly in customer churn prediction, this approach has proven to be highly effective. It commonly utilizes three prevalent structures: bagging, boosting, and multi-stage methods. These algorithms leverage the strengths of various models, such as Random Forest (RF) and Light Gradient Boosted Machine (LightGBM), with the aim of enhancing predictive accuracy. LightGBM stands out due to its efficient optimization, parallel processing capabilities, integration of regularization techniques, and sophisticated bagging strategies.

Bagging, or bootstrap aggregating, is an ensemble learning technique that enhances the accuracy and stability of machine learning models. By combining predictions from multiple weaker models trained on different subsets of the training data, bagging reduces variance and prevents overfitting. It can be categorized into two types: homogeneous bagging, which employs multiple copies of the same weak classifier, and heterogeneous bagging, which uses diverse weak classifiers. Random Forest is an example of a homogeneous bagging algorithm. Sabbeh (2018) assessed machine learning methods for customer churn prediction using behavioral data. Random Forests (RFs) demonstrated the highest accuracy, while Naïve Bayes (NB) and logistic regression (LR) were the least accurate models.

Boosting algorithms improve predictive models by combining predictions from multiple weak learners. This iterative process focuses on rectifying errors made by previous learners, continuously refining predictions, enhancing predictive capability, improving performance in handling complex datasets and predictive tasks. For example, Rijnen (2018) employed a two-phase approach to identify customers with a higher likelihood of churn. The initial phase utilized unsupervised learning to identify homogeneous user groups, while the subsequent phase focused on classification using supervised learning algorithms. Notably, the study discovered that the churn class achieved the highest Area Under the Curve (AUC) value of 0.95 when employing the XGBoost, a boosting classifier.

Multi-stage ensemble learning methods involve combining multiple models or algorithms in a sequential or hierarchical manner. These methods utilize a series of steps where each stage improves predictive performance by using output or intermediate predictions from the previous stage. In a comparative study, the Logit Leaf Model (LLM) outperformed other methods like Decision Trees, Logistic Regression, Random Forests, and Logistic Model trees (De Caigny et al., 2018). The LLM combines decision trees for segment classification and logistic regression for each segment, resulting in higher prediction accuracy.

The literature on predictive churn has identified several hybrid models. Óskarsdóttir et al. (2020) developed a relational learner, while Ahmed et al. (2017) utilized a metaheuristic approach that incorporated a hybridized firefly algorithm. Faris (2018) introduced a novel hybrid model that combines Particle Swarm Optimization (PSO) with a feedforward neural network. In a comprehensive comparison, Keramati et al. (2014) evaluated a hybrid model using average scores against individual models such as Decision Trees (DT), Artificial Neural Networks (ANN), K-Nearest Neighbors (KNN), and Support Vector Machines (SVM). Their findings demonstrated the superior performance of the hybrid model, with ANN emerging as the most effective individual model.

Researchers are actively exploring various methodologies to predict customer churn, going beyond the techniques previously presented. Ullah et al. (2019) examined the effectiveness of information gain and correlation attribute ranking in selecting relevant features. Verbeke et al. (2012) utilized Ant-Miner+ and ALBA methods to improve accuracy and comprehensibility. Idris et al. (2017) combined Genetic Programming with Adaboost, while Vo et al. (2021) applied text mining techniques to analyze unstructured call center data.

Customer churn prediction has become a critical research area in the telecommunications industry. Machine learning algorithms, such as random forests, support vector machines, and XGBoost, have shown impressive performance with accuracies of up to 98% (Ullah et al., 2016). These algorithms excel at capturing complex relationships in customer data, enabling accurate churn prediction.

Ensemble machine learning methods, like bagging, are known for their high predictive accuracy. However, they lack interpretability. Interpretable machine learning methods and ensemble models face challenges with complex feature engineering, subjectivity, and scalability limitations.

Deep learning techniques offer a promising solution to overcome these limitations (Diro & Chilamkurti, 2018). Deep learning models, such as feedforward neural networks and convolutional neural networks, have shown comparable performance to traditional methods in

churn prediction (Umayaparvathi & Iyakutti, 2018; Ahmed et al., 2019; Alajmani, et al., 2020).

Deep neural networks (DNNs) can automatically learn and refine feature representations from raw data, eliminating the need for manual feature engineering, improving performance (Bar, 2014). However, deep learning models present challenges like computational complexity, the need for careful hyperparameter tuning, and the risk of overfitting specific datasets (Abhinav & Vijay, 2020).

Deep learning models have found their way into various domains, including medical image analysis and Natural Language Processing. However, there has been limited research and application of these models in predictive churn. In their research, Alboukaeey et al. (2020) explored deep learning techniques in the context of customer churn prediction. They examined the effectiveness of Long Short-Term Memory (LSTM) and Convolutional Neural Network (CNN) models, ultimately finding that the LSTM models outperformed the CNN models. However, it is important to note that both LSTM and CNN models are susceptible to overfitting, as evidenced by the findings of this study.

The attention mechanism plays a crucial role in churn prediction by empowering deep neural networks to effectively manage information flow and filter out irrelevant stimuli. This results in improved customer retention and better business outcomes by identifying patterns and relationships that contribute to customer attrition. Several attention-based models, such as the Transformer architecture, have showcased exceptional performance in capturing the intricate dynamics of customer behavior (Correia & Colombini, 2021).

Survival analysis, specifically in the context of customer lifetimes and churn probability, holds significant value yet remains underutilized (Wong, 2011). CLV prediction, extensively studied in contractual and non-contractual scenarios, involves employing models such as Cox regression for contractual churn and probability models for non-contractual situations (Fader et al., 2010). Non-contractual contexts involve the utilization of the Pareto/NBD model and its variations to capture customer purchases and the time until departed (Fader et al., 2010; Jerath et al., 2011).

In the financial services industry, Zopounidis et al. (2008) used survival analysis to measure switching rates and survival periods, identifying appealing products based on churn indicators. Neural networks demonstrate promise in CLV prediction, as evidenced by Chamberlain et al. (2017) in online fashion retail and Moro et al. (2015) in bank deposit subscriptions.

In the gaming industry, Viljanen et al. (2016) utilize survival analysis to calculate churn rates. By combining predictions using techniques such as random forests or neural networks,

accuracy is enhanced compared to individual models.

Despite these advancements, many studies have refrained from disclosing specific features, often due to the utilization of multiple datasets or the protection of business rules.

In this section, we have explored several studies that focus on data analysis and predictive models for churn. The telecommunications sector, which is undergoing significant growth and transformation, emerges as a prominent area of interest in machine learning research related to business quality.

To summarize the key points discussed earlier, the studies commonly used accuracy as the primary metric, followed by specificity, sensitivity, precision, and recall. These metrics are often combined into a score, known as the F1 score, which takes into account both precision and recall. Some studies also incorporated the AUC (Area Under the Curve) as an evaluation metric. Logistic Regression, K-Nearest Neighbors, decision trees, and SVM are among the most commonly used algorithms in this field (De Caigny et al., 2018). However, it is important to note that the performance levels of these algorithms can vary significantly depending on the specific data context of each study.

From the literature analysis, it becomes evident that there is no universally optimal algorithm or method for churn prediction that consistently demonstrates excellent performance across all types of data. It is worth mentioning that the mere presence of advanced machine learning calculation libraries or software does not guarantee favorable results.

Generally, the studies analyzed approach the detection of potential churn as a binary classification task (churn vs. non-churn) based on historical customer data. The data, methods experimented with, as well as the preprocessing techniques and criteria for feature selection, vary across different cases.

It is important to emphasize that the analysis of churn in a subscription-based service differs from that in a service without a subscription or contractual obligation for customers to inform the company about their intention to discontinue the service. In the latter scenario, determining whether there has been abandonment or just a period of inactivity poses a greater challenge, which is commonly encountered in areas such as retail.

Predicting churn in telecommunication companies poses significant challenges due to the handling of massive datasets. These datasets comprise hundreds of millions of daily transactions stored in complex databases. Extracting pertinent information from these databases is a complex task, given the extensive number of tables and rows. Not only is the dataset large, but it is also high-dimensional, demanding the use of specialized and efficient data mining algorithms.

Class imbalance, also referred to as data imbalance, occurs when the distribution of classes in a dataset is significantly skewed. This results in an under-representation of one class (minority class) and an over-representation of the other (majority class). This imbalance is a common issue in binary classification scenarios and can have adverse effects on the interpretability and reliability of results.

Moreover, the data often suffers from class imbalance, with a substantial majority of non-churn instances compared to churn instances. This creates additional obstacles in model training and evaluation.

Our analysis also uncovers a prominent gap in the existing research landscape. While much attention has been given to churn prediction for mobile subscribers, the fixed telecommunications sector remains relatively underrepresented. The unique characteristics and complexities of the fixed sector, particularly among quadruple-play subscribers, warrant dedicated research efforts. Our dissertation seeks to bridge this divide by creating a churn model specifically tailored to fixed-telecommunications customers utilizing quadruple-play services in a Portuguese telecommunication company. By addressing this research gap, we aim to extract valuable insights into the factors contributing to churn in the fixed sector, ultimately assisting telecom companies in enhancing customer retention efforts within this industry segment.

CHAPTER 3

Class Imbalance

Class imbalance has far-reaching implications, affecting critical areas like the accuracy of medical diagnoses, effectiveness of fraud detection in credit card transactions, reliability of software testing, relevance of e-commerce recommendations, and even strategies for stock selection. It is crucial to address this issue to ensure fair and dependable outcomes across these diverse domains.

Traditional machine learning classifiers often make the implicit assumption that the distribution of classes is balanced. Imbalanced data often results in class imbalance, which can have several negative implications for supervised models. These include small sample sizes, class overlapping, and within-class imbalance.

A small sample size can pose challenges in classification tasks, as there may not be enough data for certain classes. This can make it difficult for the classifier to learn the patterns of these classes effectively. Additionally, distinguishing between classes can be challenging due to class overlap or limited separability. This occurs when different classes share similar features, making it hard to establish clear boundaries. Another issue is class imbalance, where subgroups within a class vary significantly in size. Training a classifier on an imbalanced dataset with within-class imbalance can lead to deteriorated performance, as the classifier may exhibit bias towards the more prevalent subgroups (Chawla et al., 2002).

Collectively, these factors can significantly impact the performance of the models. Moreover, this imbalance can result in biased models that favor the majority class and perform poorly on the minority class. As a consequence, suboptimal outcomes may arise, such as inaccurate predictions and biased decision-making. In critical domains, where precise predictions are paramount, addressing data imbalance becomes crucial for optimal performance.

To address imbalanced data, there are two crucial levels where effective strategies can be implemented: algorithmic approaches and data manipulation techniques.

3.1 Algoritmic level

Algorithmic strategies, seamlessly integrated into the model training process, involve specialized algorithms to tackle imbalanced datasets. Two main approaches in this domain are cost-sensitive learning and ensemble machine learning methods. The choice of strategy depends

on the dataset characteristics and desired classification performance, requiring careful evaluation.

Cost-sensitive learning tailors the algorithm by assigning distinct misclassification costs to minority and majority classes. This prioritizes accurate classification of minority class instances, mitigating misclassification due to the overwhelming majority class. Various classification algorithms, including Logistic Regression (Shen et al., 2020), KNN (Zhang, 2020), and Decision Tree (Jabeur et al., 2020) have been employed in cost-sensitive approaches.

Ensemble techniques in classification involve the collaboration of multiple classifiers to arrive at a cohesive final decision (Cao et al., 2014; Park & Ghosh, 2012). For example, Salunkhe and Mali (2016) developed ensemble classifiers by utilizing diverse training datasets and classifier models. Furthermore, researchers have successfully integrated data-level methodologies with classifier ensemble techniques.

Additional techniques include thresholding, data transformation, active learning, and feature selection (Johnson et al., 2019; Aggarwal et al., 2021).

3.2 Data level

Data level have emerged as a prominent strategy for tackling imbalanced data. These approaches utilize sampling techniques during data preprocessing, before the classification stage. The primary objective is to address the data imbalance by augmenting the samples of the minority class through oversampling, reducing the samples of the majority class through undersampling, or employing a hybrid sampling technique that combines both.

A notable advantage of data-level approaches is their versatility, as they can seamlessly integrate with various classification algorithms. To assess their effectiveness, one can consider factors such as computational complexity, distortion of data distribution, generalization across multiple classes, and compatibility with various learning algorithms. Evaluating these aspects can provide valuable insights into the performance of these techniques.

3.2.1 Oversampling methods

Oversampling techniques are employed to address dataset imbalances through duplication of existing observations or generation of artificial data. These methods can be categorized as either uninformed, where samples are randomly chosen for replication, or informed, where emphasis is placed on areas with the greatest impact. Random oversampling and Synthetic Minority

Oversampling Technique (SMOTE) are commonly utilized approaches to rebalance datasets, each offering distinct strategies to tackle this challenge.

Random oversampling is a simplistic data augmentation technique that involves randomly selecting and duplicating instances from the minority class (Ghazukhani et al., 2012). This approach is repeated multiple times to achieve the desired class balance, without relying on any specific logic or heuristic guidance. This raises the chances of overfitting, increasing the risk of the model learning the training data too well and performing poorly on new, unseen data.

The Synthetic Minority Oversampling Technique (SMOTE), introduced by Chawla et al. (2002), is a widely used oversampling method in machine learning. It aims to address the issue of imbalanced dataset by creating new data points through interpolation between existing minority class samples, ensuring they fall within the decision boundary of the minority class.

While random oversampling duplicates existing minority class samples and can lead to overfitting, SMOTE generates new data points that better represent the distribution of the minority class. This helps the classifier focus on the underlying patterns and improves performance.

The SMOTE algorithm generates new samples by interpolating randomly between existing samples and their neighboring ones. It leverages the K-Nearest Neighbors (KNN) algorithm to create these synthetic samples, thus expanding the dataset, as seen in Figure 2.

SMOTE operates through a three-step process to generate synthetic samples. The specific process of SMOTE can be outlined as follows:

1. Calculate the distance between each minority sample i (i $= 1, 2, \ldots, n$) and other samples in the minority set. This is done by following specific rules to identify the k nearest neighbors for each sample.

2. We apply oversampling magnification to select a subset of k nearest neighbors from each sample i. These neighbors, denoted as j (j $= 1, 2, \ldots, m$), are then used to calculate an artificially constructed minority sample p_{ij} using the follow equation:

$$p_{ij} = x_i + rand\ (0,1)\ (x_{ij} - x_i)$$

3. The process pauses until the accumulated or fused data reaches a specific ratio of imbalance. This ratio is determined by generating a random number from a uniform distribution

in the range of [0,1] using the function *rand*(0, 1) (Chawla et al., 2002).

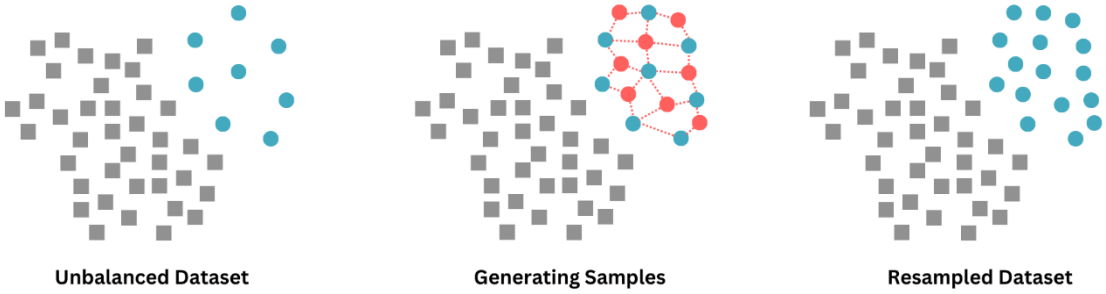

Figure 2. Illustration of SMOTE Oversampling for Imbalanced Classification (own production)

SMOTE has emerged as a highly effective technique for addressing class imbalance across various environments. Extensive research has been conducted on the effectiveness of SMOTE in combination with various classifiers. Bhagat and Patil (2015) showcased the synergy between SMOTE and random forests, especially for large datasets. Elreedy and Atiya (2019) evaluated the performance of SMOTE across decision trees, support vector machines, and K-nearest neighbors. Their findings indicate that the impact of SMOTE on prediction accuracy varies among classifiers and is more pronounced in highly imbalanced datasets.

In their 2016 study, Amin et al. examined the effectiveness of six different oversampling techniques in addressing this issue within the telecommunications industry. By applying these methods to four publicly available telco datasets, they discovered that the Mega-Trend-Diffusion (MTDF) technique and rule generation using genetic algorithms yielded the most favorable results. Another notable study conducted by Hartati et al. in 2018 explored the effectiveness of combining random undersampling and SMOTE with bagging, resulting in a remarkable 56% increase in the F1-score when the data was balanced.

The SMOTE algorithm, while effective in addressing between-class imbalance, has some limitations. It overlooks within-class imbalance and small disjuncts, resulting in the inflation of densely populated minority areas and the persistence of sparsely populated ones. Additionally, SMOTE may amplify noise in the data when interpolating noisy minority samples near majority class instances. The algorithm also fails to enforce the decision boundary, as it oversamples instances regardless of their proximity to the class border. Addressing these weaknesses could improve classifier performance.

The algorithm has gained immense popularity owing to its simplicity and robustness. Several variations inspired by SMOTE have been proposed, solidifying its position as the cornerstone concept in addressing imbalanced learning problems through oversampling. These

include SMOTEBoost (Chawla et al., 2003) Adaptive Synthetic (ADASYN) sampling (He et al., 2008) and SMOTE for discrete attributes (SMOTE-D) (Torres et al., 2016). Despite these advancements, effectively handling discrete variables in oversampling techniques remains a critical area for further research.

3.2.2 Undersampling methods

Random undersampling and Tomek's Links are two commonly used techniques for rebalancing datasets. They provide different strategies to address the issue of imbalanced data. Random undersampling, a non-heuristic technique for data balancing, is a commonly used baseline approach. This method entails randomly selecting and eliminating instances from the majority class until its count matches that of the minority class.

Tomek's Links is a heuristic undersampling methodology that enhances the Condensed Nearest Neighbors (CNN) method by incorporating specific rules for entry elimination (Galar et al. 2018).

Tomek's Links are a form of data noise that may arise in imbalanced datasets. They occur when two samples from different categories serve as each other's nearest neighbors, meaning that there are no other samples from their respective categories that are closer to them. The presence of Tomek's Links can pose challenges for machine learning algorithms in accurately determining the decision boundary between the categories.

Removing Tomek's Links can enhance the performance of machine learning algorithms on imbalanced datasets. In this approach, the samples involved in Tomek's links are eliminated. The process entails the following steps:

1. Determine the closest neighbor based on Euclidean distance for each sample in the dataset.
2. If the nearest neighbor of a sample belongs to a different category, and the nearest neighbor of that sample is the original sample itself, then we have identified a Tomek's Links.
3. Remove the sample from the majority class that is connected to the Tomek's Links.

Tomek's Links effectively identify instances with noisy or borderline characteristics. To improve the robustness of the training set, it is recommended to exclude majority class examples contributing to a Tomek's Links. This practice ensures the preservation of minimally

distanced neighbors belonging to the same class, promoting well-defined clusters within the training set (Moreno-Torres et al., 2009).

As a result, classification performance is significantly improved by enhancing the discernibility and separation of class boundaries as seen in Figure 3.

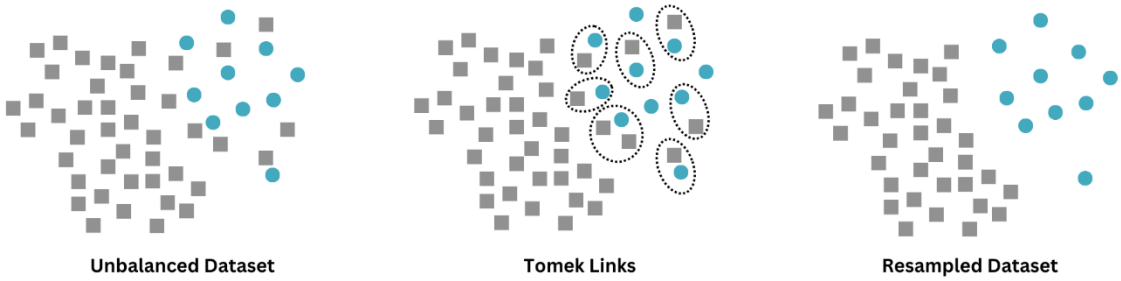

Figure 3. Illustration of Tomek's Links Undersampling for Imbalanced Classification (own production)

Similar to Condensed Nearest Neighbors, this iterative process continues until no eligible pairs remain. However, it is important to note that Tomek's Links, although not exclusively designed for kNN applications, shares the inherent limitation of not guaranteeing class equilibrium (Galar et al. 2018).

Prior to commencing an experiment, it is crucial to determine the approach for addressing the challenge of class imbalance, particularly as the minority class holds significant importance in this research. Within this study, an observed class imbalance ratio of approximately 49:1 indicates that out of 49 non-churned customers, there is a likelihood that 1 customer will churn within three months.

In this study, we primarily focus on data-level methods to tackle class imbalance. These methods, known for their simplicity and adaptability compared to algorithmic approaches, provide a practical solution. The key advantage of using data-level methods is their wide applicability to various learning algorithms, making model design and implementation more streamlined and efficient.

CHAPTER 4

Methodology

A binary churn prediction model is used to identify customers at risk of leaving. Our approach to churn analysis revolves around a structured framework. Initially, we conceptualized a broader methodology to be employed across all specific models. The research strategy is designed to define objectives and establish a clear plan for addressing research questions and objectives.

This serves as the foundation for problem analysis, which is elaborated on in this chapter. Within this framework, we identify key indicators, gather relevant data from multiple sources, and conduct comprehensive algorithmic analyses. These systematic steps synergize to provide a profound comprehension of patterns and intricacies within the dataset under examination.

4.1 Churn-Predictive Variables Analysis

Customer churn prediction has evolved beyond traditional customer characteristics to encompass a broader range of factors. While demographics, socioeconomic status, and commitment data remain important (Risselada et al., 2010; Verhoef et al., 2009), researchers have explored beyond these boundaries by incorporating relationship dynamics and external influences into churn prediction models.

Nguyen et al. (2011) conducted a study in the mobile telecommunications field to develop a predictive model for assessing customer churn probabilities. The study utilized a dataset categorized into six groups, including demographics, billing-related details, recharge patterns, communication patterns, call records, and network data.

Meeke Rijnen (2018) analyzed customer data including demographics, contracts, and online behavior to understand the factors influencing churn in telecommunication services. Sahlberg (2018) focused on churn modeling in the insurance sector, highlighting the impact of address changes and customer attributes on discontinuation of services. Results showed that income, customer age, and duration as insurance clients were important predictors of churn, with long-time and older customers less likely to churn.

Ullah et al. (2019) utilized customer data from a GSM telecom provider in South Asia. The dataset consisted of 64,107 instances and 29 numerical features. They examined various factors, such as total calls, total minutes, charged calls, and charged minutes.

To refine churn prediction, researchers have explored unconventional data sources, such as analyzing visual data from social media behavior to extract churn-predictive features (Ballings et al., 2015). This pictorial data, which captures sentiments and mindsets, complements traditional data sources. Integration of spatio-temporal and high cardinality data has also been undertaken (Kaya, 2019; Moeyesoms et al., 2011).

Valuable tools in addressing the customer churn challenge are the variables generated by the Recency, Frequency, and Monetary (RFM) model (Wang et al., 2015). These RFM-derived variables provide insights into critical behavioral patterns indicating churn risk, encompassing recency, frequency, and monetary value. Furthermore, marketing-related variables, such as promotions, retention strategy calls, and helpdesk interactions have proven to be effective predictors (Verbeke et al., 2012).

4.2 Data Collection

The dataset, collected by the Marketing and Business Development team, is stored in different CRM databases within the company and has been anonymized to comply with GDPR regulations. This dataset is expected to provide valuable information for this study.

The available data, consisting of customer information and churn labels, is suitable for the binary model to learn patterns distinguishing churners from non-churners. However, a survival model is not appropriate due to the absence of censored data for churn events that have not yet occurred.

Churn prediction models are developed using a dataset that typically spans three distinct periods (Verbeke et al. 2012, Wassouf et al. 2020):

1. The first is the Observation Window (OW), which captures customer behavior prior to churn or non-churn events. The length of the Observation Window can be customized to meet specific training requirements and accurately classify customers based on their churn likelihood. It is important for customer downtime forecasting in the telecom industry. The duration of the OW affects prediction accuracy and modeling efficiency prior to churn or non-churn events.

2. To improve customer retention, a latency period is established between the observation window and the prediction period. This phase, known as the Activity Window, involves collecting and analyzing data, training models, and implementing retention strategies. The effectiveness of these efforts determines how valuable the mining model is in using information from active customers during the AW.

3. The Churn Window (CW) is responsible for tracking and recording the final purchase

or payment made by customers who are leaving. This valuable data includes their state as a dependent variable. Each service feature has its own criteria for determining the time window. For instance, Yang et al. (2019) discovered that 95% of mobile game customers did not return after a three-day absence, establishing it as the churn period. On the other hand, Lee et al. (2018) defined the churn section in PC game services as the period when 75% of customers remained consistently disconnected. Their cumulative data revealed that customers were disconnected for more than 75% of the time for a span of 14 weeks. Therefore, the churn period is set at fourteen weeks.

These three windows significantly impact the timing of churn prediction models and capture important events that can affect customer churn.

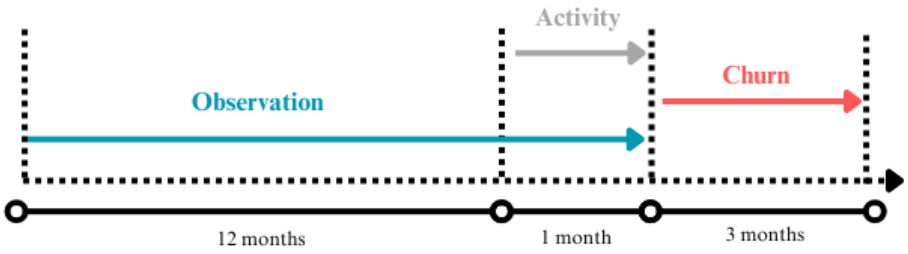

Figure 4. Periods of churn prediction used in this model (own production)

Decisions concerning time periods and client selection made for this study are presented in Figure 4. The model is trained using only active clients over a thirteen-month observation period. Churn, the target variable, is defined within a three-month period after a latency period.

When dealing with sectional data in a database context, the customer entity plays a pivotal role as discussed in the previous section. The supporting data model reflects this significance by showcasing the customer as the core connecting three primary data entities, as depicted in the accompanying diagram. The client acts as the entry point, aggregating data from the account, which, in turn, is associated with at least one service – in this case, four services (Television, Internet, Telephone, and Mobile).

In this scenario, we will create a database where only customers who voluntarily terminate their contract with the company are classified as churners. Involuntary churners have been excluded from this dataset.

Due to the availability of only one year and one month of data, the training and test datasets were limited to the period between December 2021 and January 2023. By using a one-year timeframe, we can improve efficiency and obtain accurate averages of service usage and

interactions with the company's touchpoints, which effectively reflect longitudinal consumer behavior. The churn window will commence from February 1st to April 30th, 2023. This study specifically targets residential customers with active contracts that span at least one year and one month, concluding on January 31st, 2023. These customers must maintain their subscription to a quadruple play bundle throughout this defined period.

Customers who chose to downgrade their services or made any changes to their subscription by adding more services have been deliberately excluded from this dataset.

To maximize the accuracy and effectiveness of a churn prediction model, it is key to ensure that the data used for training and evaluation is of high quality. Any errors or inconsistencies in the dataset, such as incorrect values or missing values, can have a significant impact on the performance of the model. Therefore, maintaining data integrity and reliability is essential in maximizing the predictive power of the model.

During the data preparation process, raw data was carefully analyzed to ensure its quality before proceeding with additional analysis. This involves various steps such as collecting, integrating, transforming, cleansing, reducing and discretizing the data (Kochański, 2003). The main objective of data cleaning is to address missing or incorrect values and remove any inconsistencies that could potentially impact the accuracy of machine learning models.

The database was thoroughly cleaned and aggregated to remove duplicate records and summarize data at the appropriate aggregation level, such as the yearly average.

When dealing with sectional data in a database context, each entry corresponds to an individual customer which ensures a clear and distinct representation of each customer.

The dataset includes information from numerous essential categories that consistently emerged throughout the literature review: customer demographics, personal details, customer support service records, billing and payment information, and customer usage patterns. Due to limitations in the CRM system and/or a substantial lack of data in the entire database, the addition of further data was not feasible. A detailed overview of the resulting database can be found in Table 2, which is located in Chapter 5.

Several various modifications were implemented. Categorical variables experienced one-hot encoding. Converting categorical variables into one-hot encoded vectors offers several advantages for machine learning algorithms. One-hot encoding effectively transforms non-numerical data into a format that algorithms can understand and process, enabling them to handle a wider range of data types. This technique also preserves the unique identity of each category, allowing algorithms to capture the underlying relationships between categorical

features. Additionally, one-hot encoding is relatively straightforward to implement and can be easily integrated into machine learning pipelines.

This encoding is particularly valuable for machine learning algorithms that require numerical input data. In the context of a variable with only two categories, a single binary variable is sufficient to represent both categories, as illustrated in this dissertation.

The final dataset encompasses a total of 72 variables.

Since all fields in each client's dataset were fully populated, there was no need for any treatment or adjustments regarding missing values. This ensured that our analysis used complete and reliable information, eliminating the requirement for imputation techniques commonly applied to incomplete datasets.

In outlier detection, exceptional values may be associated with individuals who demonstrate unique behaviors even in typical circumstances, particularly when investigative individual behaviors or rare occurrences. These anomalies can offer valuable insights for our models. Consequently, no actions were implemented to address or eliminate outliers (Chawla et al., 2003).

4.3 Algorithmic Approach

According to the Parsimony Law, it is established that no single machine learning model can consistently outperform all others in every possible scenario, irrespective of the task or data (Wolpert & Macready, 1997).

This dissertation utilized PyCaret, an open-source Python library for machine learning. PyCaret automates ML tasks, streamlines workflows, and simplifies data preparation by creating ML pipelines. It facilitated the comparison of classification algorithms, data encoding, and automated data splitting for training and testing purposes.

A total of fifteen models were employed, encompassing a diverse array of algorithms. These included AdaBoost, Gradient Boosting Classifier (GBC), CatBoost, LightGBM, Decision Tree (DT), Random Forest, Extra Trees, Logistic Regression (LR), Ridge Regression, Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), K-Nearest Neighbors (KNN), Naive Bayes classifiers, Support Vector Machines (SVM), and Dummy Classifier AdaBoost, Gradient Boosting Classifier (GBC), CatBoost, and LightGBM are all boosting algorithms used in machine learning. Boosting algorithms in machine learning are ensemble methods that combine multiple weak learners to create a strong learner. These weak learners, although individually less accurate, contribute collectively to a more accurate model.

Boosting algorithms sequentially train these weak learners, with each one aiming to improve upon the errors of the previous learner. They each have unique features and techniques that contribute to their effectiveness in creating ensemble learners. AdaBoost and GBC focus on combining weak learners, while CatBoost addresses the challenge of handling categorical features. LightGBM stands out for its histogram-based technique and feature bundling, enhancing efficiency and performance.

A Decision Tree (DT) is a predictive model that recursively divides data into subsets based on features to make decisions. In contrast, Random Forest creates an ensemble of decision trees by using bootstrapped samples and feature randomness. It then averages the predictions to improve robustness and reduce overfitting. Another variation, Extra Trees, further enhances diversity by introducing additional randomness during tree construction through random threshold selection for feature splits.

Logistic regression (LR) is a linear classification algorithm that predicts the probability of an instance belonging to a class. It is well-suited for binary classification tasks. Ridge Regression, on the other hand, extends linear regression by preventing overfitting through regularization. It penalizes large model weights, improving generalization performance.

Linear Discriminant Analysis (LDA) is a dimension-reduction algorithm that efficiently projects high-dimensional data onto a lower-dimensional space while preserving class discrimination. It assumes a normal distribution and equal covariance matrices within each class. In contrast, Quadratic Discriminant Analysis (QDA) is a supervised classification algorithm that allows for non-linear separation between classes. QDA offers more flexibility than LDA, as it does not rely on the assumption of a linear frontier.

K-Nearest Neighbors (KNN) classifies new instances by considering the majority class among its k nearest training instances. This non-parametric approach allows for accurate predictions based on the similarities observed in the data.

Naive Bayes classifiers, which are part of the linear probabilistic classifier’s family, make use of Bayes' theorem by assuming that the features of the data are independent, given the class label. Support Vector Machines (SVM) function by identifying the most effective hyperplane to distinguish between different classes within a dataset. In a two-dimensional setting, this hyperplane takes the form of a line, while in higher dimensions, it manifests as a hyperplane.

A dummy classifier is a basic type of classifier that makes predictions without attempting to learn patterns from the data. It is commonly used as a benchmark to compare

against more sophisticated classifiers.

4.4 Model evaluation

Binary classification problems yield four categories of classification results based on the ground truth and prediction labels. Once the predictions of the test dataset are obtained and the ground-truth classes are known, a confusion matrix can be constructed to evaluate the model's performance. This matrix, presented in Table 1, represents how well the model performs based on the following criteria:

- True positives (TP): the number of positive cases that are correctly identified as positive.
- False positives (FP): the number of negative cases that are incorrectly identified as positive.
- True negatives (TN): the number of negative cases that are correctly identified as negative
- False negatives (FN): the number of positive cases that are incorrectly identified as negative.

Table 1. Confusion matrix for the binary classification tasks

	Positive Prediction	Negative Prediction
Positive Class	True Positives (TP)	False Negatives (FN)
Negative Class	False Positives (FP)	True Negatives (TN)

To evaluate the effectiveness of the algorithms employed in this dissertation, various key metrics are computed based on the output of the confusion matrix. Accuracy, which measures the proportion of correct predictions, serves as the primary gauge for classification tasks. It is given by

$$\text{ACCURACY} = \frac{TP + TN}{TP + TN + FP + FN},$$

whereas the error rate is

$$\text{ERROR RATE} = \frac{FN + FP}{TP + TN + FP + FN} = 1\text{- ACCURACY}$$

Traditionally, accuracy and error rate are commonly employed metrics for evaluating classifier performance. However, they may not be suitable for imbalanced distribution problems. For example, in a dataset consisting of 99% majority class samples and only 1%

minority class samples, accuracy may appear high but fail to reflect the misclassification of minority class samples. Thus, we will consider alternative evaluation metrics in addition to accuracy. Alongside accuracy, other metrics are taken into account.

Precision quantifies the accuracy of correctly classified positive predicted samples. It serves as a valuable metric for assessing imbalanced problems by considering misclassified negative samples (FP). However, relying solely on Precision is inadequate since it overlooks misclassified positive samples (FN). On the other hand, Recall evaluates the accuracy of correctly classified positive samples and remains unaffected by class imbalance. Additionally, Specificity indicates the percentage of correctly classified negative samples.

$$\text{PRECISION} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{RECALL} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{SPECIFICITY} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

The F-measure, a widely used metric known as the F1-score, calculates the weighted harmonic mean of precision and recall. It is defined by the following formula:

$$\text{F} - \text{MEASURE} = \frac{2 \times \text{PRECISION} \times \text{RECALL}}{\text{RECALL} + \text{PRECISION}}$$

Additionally, this study employs the Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) curve as a graphical representation of the model's performance. ROC Curve is a commonly used tool in machine learning to evaluate the performance of classifiers. It plots the True Positive Rate (TPR) against the False Positive Rate (FPR) and provides insights into the trade-off between precision and specificity. A high AUC value indicates a better classifier performance, while a perfect classifier aligns in the top left-hand corner of the curve. The ROC curve is a valuable visual aid for interpreting the performance of prediction models.

In this dissertation, we will primarily evaluate the performance using F1-score and AUC, as they have been shown to be effective for unbalanced datasets (Wang et al., 2021; He et al. 2019). These indicators effectively capture the performance of the minority class. By focusing on these metrics, we can better understand the performance dynamics and nuances of the data.

4.5 Hyperparameterization

Hyperparameterization is the process of setting the parameters of a machine learning model before training. These parameters, called hyperparameters, are not learned from the data but set before training begins. In the field of machine learning, hyperparameters play a crucial role in shaping the behavior and performance of algorithms. These parameters, whether categorical or continuous, have a significant impact on factors like model complexity and computational efficiency (Bischl et al., 2023). Selecting the optimal values for hyperparameters is essential for achieving optimal model performance (Claesen et al, 2015).

It involves exploring different configurations, training models, evaluating the resulting models based on predetermined objectives (such as out-of-sample prediction error and selecting the best-performing setup). Extensive research has been done to develop effective strategies for finding promising configurations.

Hyperparameter optimization, from a mathematical perspective, strives to optimize an objective function that maps hyperparameters to an evaluation metric. The ultimate goal is to achieve maximum accuracy or minimum loss. Through this process, we identify the set of hyperparameters that performs exceptionally well on the objective function:

$$\lambda^* = argmin_{\lambda \in \Lambda}(L(X_{validation}, A_\lambda(X_{train}))$$

In this context, λ represents a collection of hyperparameter values, and Λ signifies the entire range of potential values for these hyperparameters. L denotes the loss function, which takes $X_{validation}$ (the validation dataset) and $A_\lambda(X_{train})$ as inputs. A_λ represents the function learned by the algorithm A, configured with hyperparameters λ, after training on X_{train} (the training dataset). The loss function compares the outputs of the resulting function with the actual ground truth from $X_{validation}$ (Brochu et al., 2010).

One of the main challenges in hyperparameter optimization is the computational cost associated with evaluating the objective function. This typically involves training a model with a training dataset and evaluating it on a validation set. As different hyperparameter settings are explored, this process may need to be repeated multiple times.

There are various methods for hyperparameterization, each with its own trade-offs. Manual tuning is the traditional approach, but it can be quite time-consuming. Random search and grid search are more efficient alternatives, although they may not always yield the optimal set of hyperparameters. Bayesian optimization, on the other hand, offers a more sophisticated

approach that can be more efficient than grid search, albeit with increased implementation complexity (Bergstra et al., 2012).

The optimal approach for hyperparameterization depends on the specific machine learning model and data at hand. It is advisable to experiment with various methods to determine the most effective one for your needs.

Hyperparameterization offers several benefits, including enhanced model performance, reduced modeling time, and improved interpretability. However, it also presents challenges such as computational cost, overfitting, and the difficulty of selecting an appropriate method. The dataset needs to be divided into training and testing. However, this reduces training samples and affects model performance. Cross-validation is a technique in machine learning for comparing algorithms and assessing their real-world performance. It eliminates biases and ensures the entire dataset is used for testing. One common technique is k-fold cross-validation, where data is divided into k folds. The model is trained using k-1 folds and tested on the remaining fold. This is repeated k times. Cross-validation helps detect and prevent overfitting, where a model fails to generalize. It is valuable for improving model performance and accuracy (Burman, 1989; Jung, 2018). This study employs 10-fold cross-validation.

4.6 Interpreting Model Predictions

Understanding the significance of variables in machine learning models is complex, especially in intricate models like neural networks. Linear models simplify this by assigning weights to attributes, determining their importance. However, complex models like neural networks, often referred to as 'black boxes,' pose challenges in explaining predictions due to their lack of transparency, raising trust concerns.

The interpretability of prediction models greatly influences their acceptance and adoption. Despite achieving high accuracy, the lack of understanding in complex models hampers practical application and diminishes trust. Balancing accuracy and interpretability remains a key challenge in model development.

To overcome the inherent complexity of machine learning models, SHapley Additive exPlanations (SHAP), a model-agnostic approach, is employed. This methodology involves treating the trained model as an opaque entity, often referred to as a 'black box,' and extracting decision rules or feature values from its predictions to create an interpretable counterpart. This strategy enables the initial training of a high-performance yet opaque classifier, followed by the development of an interpretable model with decision rules that mimic its behavior.

In 2017, Lundberg and Lee introduced SHAP values, a method for interpreting machine

learning models. SHAP, a model-agnostic and post-hoc Explainable Artificial Intelligence (XAI) tool based on Shapley values, serves as a valuable feature selection technique. Its global approach and reliance on mathematical concepts make it particularly useful for interpreting complex models. The Shapley value, rooted in game theory, assigns contributions to features based on their impact on the final prediction. However, it is important to note that its computation time increases exponentially with the number of features. The Shapley value for feature Xj in a model can be expressed as follows:

$$Shapley(X_j) = \sum_{S \subseteq N \setminus \{j\}} \frac{k-(p-k-1)!}{p!} f(S \cup \{j\}) - f(S)$$

Consider a scenario where p represents the total number of features in a predictive mode. $N\setminus\{j\}$ denotes the set of all possible feature combinations excluding X_j. S represents a feature set in $N\setminus\{j\}$. The function $f(S)$ represents the model's prediction using the features within set S. Adding X_j to this set, denoted as $f(S \cup \{j\})$ represents the model's prediction with the inclusion of feature X_j to the feature set S.

The equation at the core of Shapley values plays a crucial role in quantifying the incremental contribution of a specific feature to the overall model prediction. It accomplishes this by averaging across all possible models that can be built using different combinations of features.

By systematically evaluating how each feature, when added to different features subsets, influences the model's predictions, Shapley values provide a comprehensive understanding of the impact of each feature on the model's output. This comprehensive assessment enhances our comprehension of the relative importance and influence of individual features within the model.

SHAP core contribution lies in its ability to generate locally additive feature attributions, which can be represented by the following equation:

$$\hat{y}_i = shap_0 + shap(X_{1i}) + shap(X_{2i}) + \ldots + shap(X_{ji})$$

Each SHAP value ($shap(X_{ji})$ quantifies the specific contribution of feature p to the predicted value for observation i ($\hat{y}_i$), relative to the mean prediction ($shap_0$), the collective sum of shap values represents the discrepancy between the actual prediction and the mean prediction (Lundberg and Lee, 2017).

CHAPTER 5

Application

In this section, we thoroughly explore and compare various models to identify the optimal solution for the given problem. Our evaluation framework focuses on two key metrics: AUC value and F1 score. To address the inherent data imbalance, we employed two techniques: SMOTE (Synthetic Minority Oversampling Technique) and Tomek's Links. By balancing the dataset, we conducted a rigorous comparison of results obtained using these techniques against a baseline approach that did not consider imbalance mitigation. Throughout this section, we meticulously examine and compare the algorithms based on several metrics discussed in section three.

The research dataset provides valuable insights into client information for the telecom company. It offers a comprehensive overview of the data, including the variables, their corresponding values, and a concise description of their significance. Table 2 serves as a reference point, providing crucial information about the dataset, such as the nature of the variables and their meanings.

We will analyze in detail how churn-related variables influence these models, emphasizing their impact and significance. Additionally, we will conduct a comprehensive comparison to gain a deeper understanding of how these variables collectively contribute to the predictive capability of the classifiers.

ETL stands for extract, transform, load. It is a data integration process that is used to gather data from various sources, clean and transform it, and then load it into a target data warehouse or data mart. To optimize ETL processes and strike a harmonious balance between efficient data storage and rapid data processing, a multifaceted approach has been employed, encompassing string standardization, duplicate record elimination, data summarization, and the utilization of a specialized, unique database. The final dataset includes 52,877 clients, with a low churn rate of only 2.6% (1375 clients within three-month period). The data was analyzed using descriptive statistics, including mean, frequencies, and standard deviation. Additionally, correlation analysis and multiple regression analysis were conducted.

The following are some key findings:

1. On average, clients have contracts that last for approximately 122 months, and only 15.6% of them completed the full lock-in period during this time. Among those who did not churn, it took an average of around 13.28 months to complete the lock-in period which is nearly

twice as long compared to an average of about 6.63 months for churned clients.

2. Churned clients tend to have a higher average number (about six instances) of support requests that do not require on-site technician visits compared to non-churned clients (approximately three instances). Additionally, churned clients have slightly more support interventions with technician visits (an average count of roughly 0.72) compared to non-churned clients who had about 0.38 such interventions on average. A higher repeat count is observed in churned customers compared to non-churners, indicating a pattern of persistent customer service requests (1.45 repeats against 0.94 of non-churners). Repeats refer to customer interactions involving a specific customer service issue that requires multiple attempts to resolve.

3. Clients who churn tend to have higher data usage, both for fixed and mobile services. For fixed services, churn clients use an average of 198GB, while non-churn clients use an average of 142GB. In terms of mobile data usage, churn clients consume an average of 3426MB compared to 2312MB for non-churn clients.

These insights provide valuable information about the behavioral patterns in the dataset and can support further analysis and decision-making.

Table 2. Description of the selected variables

Variable	*D*escription	Classes/Distribution
COD_SUBSCR	Customer Number	Discrete
SUBS_AGE	Duration of the Contract	Discrete (in months)
SUBS_DD	Payment method	1 - Direct Debit; 0 - Other
SUBS_FE	Invoice method	1 - Electronic Bill; 0 -Other
CLIENT_AREA	If the company application (where the account is managed, subscriptions are handled, and payments are made) is active.	1 - User; Non-User
REFERRAL_FRIEND	If the client has already referred the company to friends - the company has a discount policy for each referred friend.	1 -Yes; 0 - No
DESC_CHN	Membership sales channel	Agents; Call Center; Door to Door; Intern; Stores; Retail, Web; Unknown
HUB	Location of the client's technical hub	ACB; ALS; ARL; ASM; ATB; AVR; BEI; BLM; CBR; CDR; CMB; CRR; DST; ELV; EST; EVR; FNF; GUL; LSD; LXO; MOI; NULL; PLM; PTM; SAC; SEI; SMF; STA; TRT
TV_PRICE_DES	Television package	1 – Max; 0 – Basic
NET_SPEED_DES	Internet package	1 - Max; 0 - Basic
TIME_TO_END_LOYALTY	Time to end the lock-in period of the contract (in months)	Discrete (in months)
BOX	Quality of the box subscribed	1 - New; 0 - Legacy
APP_TV	User of the TV APP in Electronic Devices other than TV	1 - User; 0 - Non User
SUBS_FID	Subscriber of the program for clients where customers can have access to promotions from top brands in various categories.	1 - User; 0 - Non User
FEE_MIX_FULL	The average value of the last 12 months paid by the client, which includes fixed and variable amounts (such as, for example, extra subscriptions).	Continuous (in euros)
SUBS_PAYTV	Subscriber of extra paid channels	1 - Yes; 0 - No
BOXS_QTT	Quantity of TV BOX subscribed	Discrete
VOD_TOT_12M	Quantity of videoclub's movies subscribed	Discrete
GIG_NET	Average fixed internet usage (over the past 12 months).	Continuous (in GB)
MIN_MOVEL	Average mobile phone minutes usage (over the past 12 months).	Continuous (in minutes)
SMS_MOVEL	Average mobile phone SMS usage (over the past 12 months).	Continuous
DATA_MOVEL	Average mobile phone internet usage (over the past 12 months).	Continuous (in MB)
MIN_FIXED	Average fixed phone minutes usage (over the past 12 months).	Continuous (in minutes)
WO_TEC	Count of technical support instances with on-site technician visits (over the past 12 months)	Discrete
QTT_TICKET	Count of support instances that do not require on-site technician visits (over the past 12 months)	Discrete
DEBT_INV	Count of unpaid/in overdue invoices (over the past 12 months)	Discrete
COMPLAINTS	Count of complaints filed by the client (over the past 12 months)	Discrete
REPEATS	Count of customer's persistence for issue resolution - that is, for the same reported problem, the number of times the customer insisted for its resolution (over the past 12 months)	Discrete
EXTRA_CREDITS	Count of extra credits assigned to the customer - usually assigned for billing errors made by the company (over the past 12 months)	Discrete
EXTRA_DEBITS	Count of extra debits assigned to the customer - usually assigned for payments made outside the defined period (over the past 12 months)	Discrete
STORE_VISIT	Count of store visits (over the past 12 months)	Discrete
CC_CALL	Count of calls made to call center (over the past 12 months)	Discrete
RATE_DISCOUNT	Average discounts granted compared to the price of the same subscribed product.	Discrete
SUBS_STATUS_N3	Customer Status (churn or non-churn) after the end of the churn window.	Discrete

This subsection provides a description of the machine learning training process. Developing an accurate customer churn prediction model requires careful consideration of model selection and model training. To conduct an experimental evaluation, a range of machine learning techniques were chosen.

Table 3. Metrics for baseline models

	Accuracy	AUC	Recall	Precision	F1	TT (Sec)
catboost	**0.9825**	**0.9628**	0.6050	**0.8843**	**0.7115**	11.6890
lightgbm	0.9801	0.9564	0.4960	0.8546	0.6204	1.8040
gbc	0.9794	0.9571	0.4560	0.8710	0.5931	12.0080
ada	0.9742	0.8782	0.3470	0.7222	0.4658	3.8260
dt	0.9676	0.6931	0.4760	0.4417	0.4485	1.2740
rf	0.9770	0.9416	0.2360	0.8974	0.3712	6.8300
lda	0.9697	0.8615	0.3100	0.4020	0.3462	1.3830
et	0.9757	0.9066	0.1420	0.8760	0.2429	5.5180
nb	0.9087	0.8179	0.5190	0.1590	0.2429	0.4890
lr	0.9737	0.9024	0.1470	0.6260	0.2338	21.8830
svm	0.9660	0.0000	0.0690	0.1980	0.0964	1.4200
qda	0.1406	0.6035	0.9220	0.0279	0.0541	1.2830
knn	0.9693	0.0000	0.0030	0.0730	0.0057	0.7510
ridge	0.9694	0.4761	0.0020	0.1430	0.0038	2.9070
dummy	0.9744	0.5000	0.0000	0.0000	0.0000	0.4270

In the absence of any imbalance mitigation strategies, CatBoost established itself as the top-performing model, attaining the highest AUC and F1-score among all models considered. As evident in Table 3, it demonstrates good separation of classes and a balanced trade-off between precision and recall.

Table 4. Metrics for models with Tomek's Links

	Accuracy	AUC	Recall	Precision	F1	TT (Sec)
catboost	**0.9900**	**0.9795**	0.6473	**0.9433**	**0.7661**	11.6890
lightgbm	0.9867	0.9709	0.5322	0.9124	0.6690	1.8040
gbc	0.9860	0.9695	0.4890	0.9360	0.6392	12.0080
ada	0.9811	0.9455	0.3728	0.7767	0.5009	3.8260
dt	0.9724	0.7448	0.5049	0.4644	0.4827	1.2740
rf	0.9806	0.9547	0.2534	0.9597	0.4002	6.8300
lda	0.9715	0.9229	0.3338	0.4307	0.3740	1.3830
et	0.9781	0.9399	0.1532	0.9345	0.2616	5.5180
nb	0.9195	0.8805	0.5556	0.1706	0.2609	0.4890
lr	0.9764	0.9260	0.1573	0.6669	0.2523	21.8830
svm	0.9682	0.0000	0.0739	0.2103	0.1037	1.4200
qda	0.1786	0.6489	**0.9895**	0.0299	0.0581	1.2830
ridge	0.9740	0.0000	0.0032	0.0783	0.0061	0.7510
knn	0.9741	0.5110	0.0021	0.1500	0.0041	2.9070
dummy	0.9744	0.5000	0.0000	0.0000	0.0000	0.4270

Table 4 displays the performance of models trained on a dataset that incorporates Tomek's Links to mitigate class imbalance. As anticipated, most models exhibit improved AUC values in all algorithms (except for dummy) compared to the baseline. CatBoost remains a standout performer, demonstrating high AUC and F1-score.

LightGBM and Gradient Boosting Classifier also deliver a good performance, with enhanced AUC and F1-score. Conversely, SVM remains exhibiting lower AUC and F1-score. Overall, the models utilizing Tomek's Links generally show superior AUC values and F1-scores, signifying the effectiveness of class balancing techniques, by refining the training dataset and mitigating the negative impacts of class imbalance.

Notably, CatBoost shines as a robust performer in both scenarios, underscoring its proficiency in handling class imbalance.

Table 5. Metrics for models with SMOTE

	Accuracy	AUC	Recall	Precision	F1	TT (Sec)
catboost	**0.9905**	**0.9779**	0.6631	**0.9520**	**0.7814**	**0.6220**
lightgbm	0.9896	0.9781	0.6399	0.9320	0.7582	0.6530
gbc	0.9875	0.9736	0.5723	0.9031	0.6999	0.7980
ada	0.9787	0.9414	0.4709	0.6087	0.5298	0.8440
dt	0.9710	0.7425	0.5017	0.4443	0.4703	0.5680
rf	0.9820	0.9619	0.3146	0.9409	0.4702	1.1610
et	0.9800	0.9423	0.2523	0.8819	0.3915	1.4730
lr	0.8479	0.9277	0.8554	0.1289	0.2240	1.3050
ridge	0.8343	0.0000	**0.8965**	0.1235	0.2170	0.5320
lda	0.8343	0.9364	**0.8965**	0.1235	0.2170	0.5720
svm	0.7671	0.0000	0.7585	0.0998	0.1669	0.5400
nb	0.7842	0.8630	0.7784	0.0868	0.1561	0.8580
qda	0.2006	0.6029	0.9567	0.0298	0.0577	0.5790
knn	0.7910	0.5426	0.2398	0.0313	0.0554	6.0960
dummy	0.9744	0.5000	0.0000	0.0000	0.0000	0.5600

Overall, SMOTE models showed better performance than the baseline and Tomek's Links, with CatBoost consistently performing better as seen on Table 5. LightGBM and Gradient Boosting Classifier also exhibited good performance, particularly when SMOTE was applied. On the other hand, SVM, Ridge, and K-Nearest Neighbors performed relatively poorly even with SMOTE. It is important to note that the estimated models may not be expected to yield consistently favorable results across all metrics, particularly when dealing with an unbalanced dataset characterized by significant class imbalance.

Hyperparameter tuning was performed on three primary algorithms (CatBoost, LightGBM, and Gradient Boosting Classifier) to enhance predictive capability using the SMOTE oversampling strategy. Custom hyperparameterization was employed to further improve results, followed by evaluation using 10-fold cross-validation.

Table 6. Performance Metrics Comparison for Catboost Models with Different Techniques

	Accuracy	AUC	Recall	Precision	F1	TT (Sec)
Catboost with SMOTE	0.9905	**0.9779**	0.6631	0.9520	0.7814	**0.6220**
Catboost with SMOTE and RandomGrid_Search	0.9907	0.9768	0.6673	**0.9522**	0.7841	12.430
Catboost with SMOTE and Customized Hyperparameterization	**0.9908**	0.9760	**0.6789**	0.9447	**0.7894**	325.4030

As seen in Table 6, by leveraging SMOTE and implementing customized hyperparameterization in CatBoost, we observed significant improvements in both metrics and overall algorithm performance. The CatBoost model with customized hyperparameterization

stood out, achieving the highest F1 score (0.7894), accuracy, and recall compared to other models.

On the other hand, the CatBoost model without any hyperparameterization showed computational efficiency while maintaining the highest AUC (0.9779). Regarding customized hyperparameterization, we reduced the learning rate to 0.1 and increased the number of estimators to 1250, resulting in improved critical metrics such as F1 compared to Random Grid Search.

Lowering the learning rate allows the model to take smaller steps during training, enabling more precise adjustments. This fine-tuning often leads to better convergence and more accurate updates to the model weights, potentially enhancing its ability to generalize to unseen data. Increasing the number of estimators, especially in ensemble methods like CatBoost, tends to boost predictive performance. With a larger number of estimators, the model can capture more intricate patterns in the data, potentially reducing bias and improving overall accuracy.

Table 7. Performance Metrics Comparison for Light GBM Models with Different Techniques

	Accuracy	AUC	Recall	Precision	F1	TT (Sec)
Light GBM with SMOTE	0.9896	**0.9781**	0.6399	0.9320	0.7582	**0.6530**
Light GBM with SMOTE and RandomGrid_Search	0.9884	0.9775	0.5871	0.9331	0.7201	12.4540
Light GBM with SMOTE and Customized Hyperparameterization	**0.9901**	0.9753	**0.6462**	**0.9504**	**0.7687**	35.4980

The hyperparameter-tuned LightGBM with SMOTE outperformed other configurations in all key evaluation criteria, including accuracy, recall, precision, and F1-score, as observed in Table 7. However, it did have a longer training time of 35.4980 seconds.

By customizing the hyperparameters, setting max_depth to 6 and increasing the number of estimators to 1250, significant improvements were achieved across all metrics except AUC. This configuration is particularly suitable when precision and F1-score are crucial, even though it requires a longer training time.

By setting the max_depth to 6 during the boosting process, we gain effective control over the depth of individual trees. This not only prevents overfitting and improves generalization on unseen data, but also enhances the training speed and reduces the likelihood of memorizing the training data. Moreover, increasing the number of estimators to 1250 empowers the model to learn from a larger pool of weak learners, specifically decision trees. This enables the model to capture intricate patterns in the data, potentially resulting in a more sophisticated overall model.

Table 8. Performance Metrics Comparison for Gradient Boosting Classifier (GBC) with Different Techniques

	Accuracy	AUC	Recall	Precision	F1	TT (Sec)
GBC with SMOTE	0.9875	0.9736	0.5723	0.9031	0.6999	**0.7980**
GBC with SMOTE and RandomGrid_Search	**0.9888**	0.9742	**0.6419**	0.8914	**0.7401**	11.9400
GBC with SMOTE and Customized Hyperparameterization	0.9886	**0.9766**	0.6029	**0.9270**	0.7243	29.4460

In summary in Table 8, the Gradient Boosting Classifier with SMOTE and Random Grid Search outperformed other models in terms of F1 score, recall, and accuracy, making it the preferred choice for this classification task. It successfully addressed class imbalances, exhibited higher accuracy in predicting positive instances, and outperformed other considered models in terms of overall precision.

However, the Gradient Boosting Classifier with SMOTE and hyperparameterization model showed the best performance in AUC and precision, which is valuable in scenarios where minimizing false positives is important. Markedly, the customized hyperparameterization in this model involved reducing the learning rate, increasing the number of estimators, and adjusting the max_depth, resulting in distinct trade-offs in performance metrics.

For example, reducing the learning rate can improve generalization but require more boosting rounds, increasing the number of estimators enables learning complex patterns but demands more computational resources, and adjusting max_depth prevents overfitting but may limit capturing complex relationships in the data.

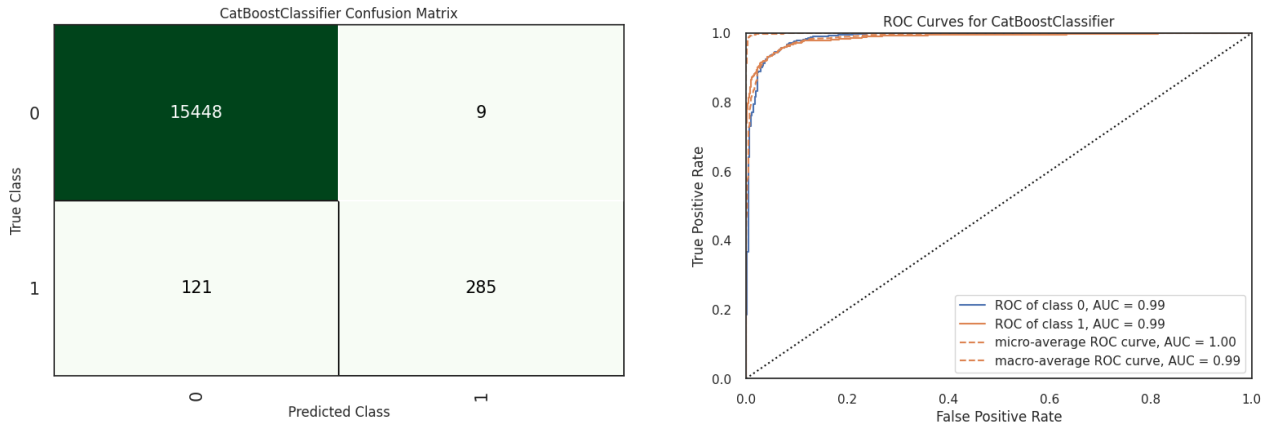

Figure 5. Confusion Matrix and ROC Curve for CatBoost with SMOTE and custom hyperparametrization

The results showcased the effectiveness of a comprehensive strategy that combines oversampling and hyperparameterization to boost the performance of predictive models on

imbalanced datasets. The findings demonstrate that by combining CatBoost with SMOTE and customized hyperparameterization, we achieve exceptional performance metrics. This is evidenced by the ROC curve and the confusion matrix depicted in Figure 5. As a result, this robust model will be used in the next phase of the research to gain deeper insights into churn factors and patterns, facilitating informed decision-making and strategy development.

Figure 6 provides valuable insights into the significance of the top 20 features for CatBoost, without any unbalancing technique and the top 20 CatBoost features when SMOTE is applied with customer hyperparameterization. To further enhance our analysis, we incorporate Shapley Additive exPlanations (SHAP) values. These values reveal the individual impact of each feature on the model's predictions.

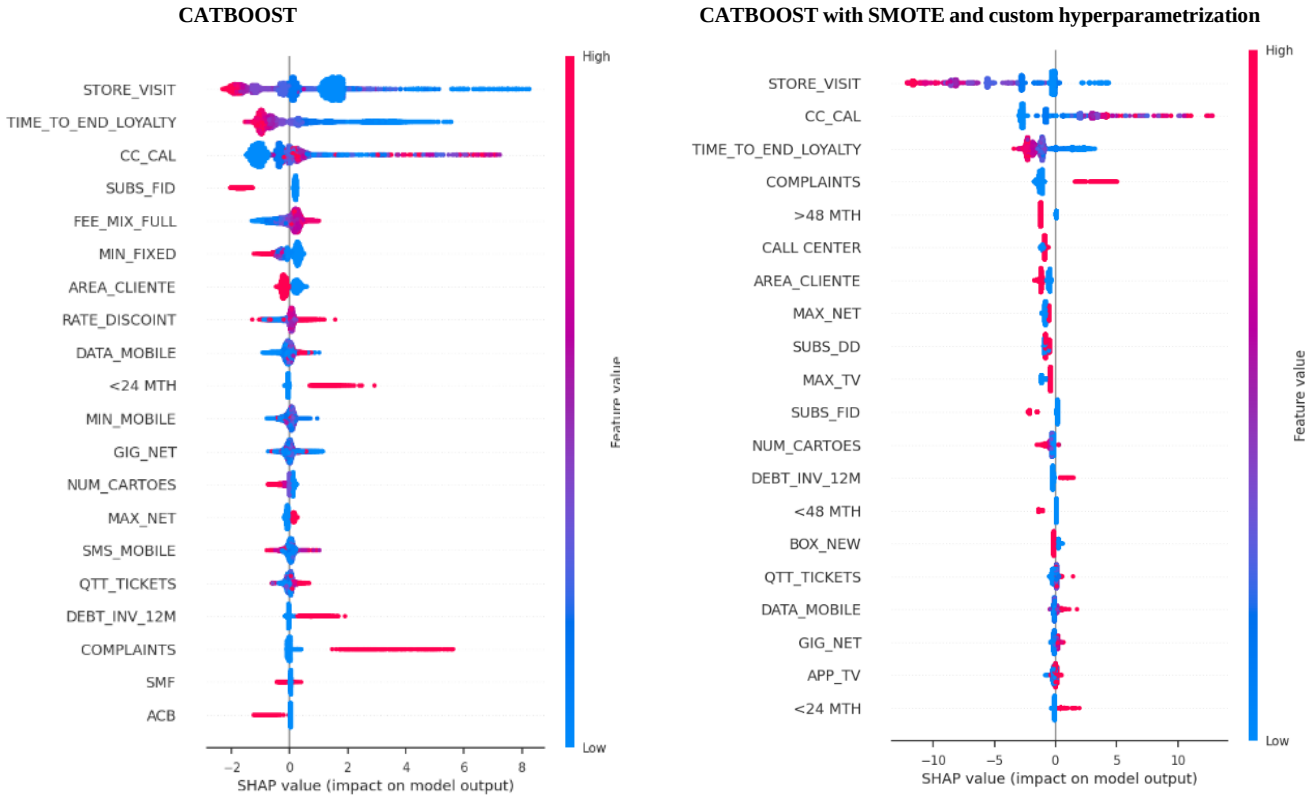

Figure 6. SHAP plot for CATBOOST and CATBOOST with SMOTE and custom hyperparametrization

Based on the findings of the SHAP analysis, it appears that customers who visit the physical stores have a lower likelihood of churning. This finding reinforces the belief that face-to-face interactions can greatly enhance customer loyalty and satisfaction. In-person visits offer personalized assistance, immediate issue resolution, and build trust through physical presence. These factors collectively contribute to heightened customer loyalty and satisfaction (Qualtrics,

2023).

Engaging with the call center can increase the likelihood of customer churn, implying that interactions through the call center may not be as effective in retaining customers compared to face-to-face interactions. The effectiveness of call center engagement plays role in influencing customer churn, highlighting the importance of continuously improving call center operations. Factors such as service quality, communication, and issue resolution contribute to customer dissatisfaction and the potential for churn (Dean, 2007).

Engaging with the call center can increase customer churn for businesses. To reduce churn, businesses should focus on improving call center interactions, including service quality, communication, and issue resolution.

Furthermore, the analysis indicates that customers with a shorter remaining lock-in period are prone to churn. The SHAP values, in both cases, indicate that this feature strongly influences customer retention in a negative way. This suggests that customers nearing the end of their contractual commitments may be exploring alternative options or reevaluating their relationship with the company.

Implementing a subscriber program featuring exclusive promotions from top brands can effectively minimize customer churn, as per the SHAP analysis of the initial model. By offering valued perks, subscribers feel appreciated, resulting in heightened satisfaction and stronger customer relationships. Additionally, exclusive promotions help mitigate price sensitivity and have the potential to transform subscribers into enthusiastic brand advocates.

Additionally, the presence of customer complaints is a strong predictor of churn, as indicated by the substantial positive SHAP value. Customers who have filed complaints are more likely to churn, indicating that unresolved issues or dissatisfaction significantly impact their decision to leave. To mitigate customer attrition, it is essential to respond to complaints and improve the process of solving issues.

Lastly, the SHAP figure demonstrates that longer customer tenure has a strong negative impact on churn prediction. This suggests that long-term customers have likely developed a sense of loyalty, trust, and a deeper connection with the brand, which increases their inclination to stay (Krautz & Hoffmann, 2017).

CHAPTER 6

Discussion

In the telecommunications industry, understanding customer behavior is key to stimulate usage, mitigate customer turnover, and boost overall customer contentment. Robust evaluations employ diverse methodologies, data analysis techniques, and predictive algorithms to identify patterns and triggers of churn. Customer segmentation and profiling enable tailored approaches for different customer groups. By leveraging this approach, telecom companies can reduce attrition and foster growth in a competitive industry.

This dissertation effectively tackles the research questions and accomplishes the stated goals, offering a comprehensive understanding of churn prediction in imbalanced datasets. Moreover, it makes a valuable contribution to the existing body of knowledge.

The company's database contained genuine data, but it was noticeably imbalanced. Less than 3% of the entries represented churning clients, posing challenges for our models. Training on such skewed data could lead to overfitting, undermining prediction accuracy. This study investigates handling imbalanced data in classification tasks. It proposes using effective sampling techniques to rebalance the dataset, regardless of the chosen classifier. The introduction of both Tomek's Links and the Synthetic Minority Oversampling Technique (SMOTE) significantly improved model performance.

In both scenarios, CatBoost consistently demonstrated superior performance, achieving exceptional AUC and F1 scores. Notably, CatBoost with SMOTE surpassed other models, reaching an AUC of 0.9779 and an F1 score of 0.7814 on this dataset of residential telecom customers with four-play services. This performance stems from SMOTE's ability to effectively synthesize minority class data, mitigating the impact of class imbalance and enhancing the model's overall learning capabilities. While Tomek's Links also proved beneficial, SMOTE's superior data augmentation technique yielded more robust results. Further refinement through manual hyperparameter tuning elevated the F1 score to 0.7894, cementing CatBoost with SMOTE as the optimal choice for knowledge extraction.

The SHAP analysis unveiled that physical store visits play a pivotal role in reducing churn, whereas call center engagement may actually contribute to churn. Additionally, factors such as shorter remaining lock-in periods, customer complaints, and longer customer tenure were identified as significant influencers of churn predictions.

Through this dissertation, we have successfully demonstrated the suitability of

classification models applied to the dataset. These models have effectively generated predictions for anticipating customer churn. The dataset represents authentic and comprehensive data from a telecom company, making this model practically valuable for the company's benefit. By utilizing such a model, the administration can accurately forecast and focus attention on clients who are approaching contract termination.

To enhance the developed model, it is suggested to include additional variables such as sociodemographic indicators of customers. Current model focuses solely on customer consumption and account information, overlooking key attributes like age, gender, family size, and marital status. Studies in literature review have shown the significant impact of these attributes on model quality. Additionally, customer churn can be influenced by word of mouth, rumors, commentaries, and the churn decisions of other customers, highlighting the importance of social interactions in churn decisions (Pinheiro, 2011) not included in this model.

Since our dataset spans a 14-month timeframe, the inclusion of older data may have a negative impact on our models. It would be beneficial to explore a data frame that encompasses the client's entire history with the company, from day one. This analysis would help us determine if such an approach yields superior results and how it affects the overall efficiency of our models.

Voice of the Customer (VoC) plays a crucial role in churn models, as it captures customer feedback and complaints related to service quality. However, the company currently lacks the capability to effectively incorporate unsolicited and direct VoC interactions into its data structures. This limitation hinders the seamless integration of VoC into the churn model, resulting in overlooked valuable insights and customer feedback that could enhance the model's effectiveness (Witell et al., 2011). Addressing this gap in data collection and analysis is essential for a more comprehensive and robust approach to churn prediction and mitigation.

During the data collection process, we faced challenges with inconsistent data sources, prompting the development of a unified CRM system. Despite addressing these issues, the overall quality of the dataset was impacted. Limitations and obstacles affected the quantity and quality of available data due to strategic constraints within the company. As a result, the size and scope of the final dataset used for analysis were influenced.

To improve the modeling process, exploring alternative techniques for data pre-processing, such as imputing missing values and detecting outliers might help improving results. Efforts in dimensionality reduction can refine the dataset and enhance modeling effectiveness.

The algorithms underwent parameter tuning to optimize the models. However,

additional methodologies for fine-tuning were not explored during testing. These unexplored approaches have the potential to further enhance the final models beyond the tested parameters.

While traditional machine learning models have yielded satisfactory results in this project, given the abundance of available data, it is worthwhile to delve into a deep learning approach. This strategy entails harnessing the power of Artificial Neural Networks (ANNs) in future research endeavors.

References

Abhinav, K., & Vijay, K. (2020). A Review on Deep Learning Techniques for Churn Prediction in Telecommunications Industry. International Journal of Recent Technology and Research, 9(2), 108-115.

Aditya Kapoor. (2017, June 28). Customer Retention and Churn Rate in the US Telecommunications Sector. Retrieved from https://www.statista.com/statistics/816735/customer-churn-rate-by-industry-us/

Aggarwal, U., Popescu, A., & Hudelot, C. (2021). Minority Class Oriented Active Learning For Imbalanced Datasets. In 2020 25th International Conference on Pattern Recognition (ICPR) (pp. 4355-4362). IEEE. doi:10.1109/ICPR48806.2021.9412182

Ahmed, A. M., Hassan, M. A., & Abdel-Salam, H. (2017). Churn Prediction Model in Telecommunications Using Firefly Algorithm. Wireless Networks, 23(5), 1477-1487.

Ahmed, M. S., Umayaparvathi, K., & Iyakutti, K. (2019). Customer Churn Prediction in Mobile Telecommunication Using Deep Learning. Wireless Networks, 25(6), 3529-3545.

Alajmani, S., & Jambi, K. (2020). Assessing Advanced Machine Learning Techniques for Predicting Hospital Readmission. International Journal of Advanced Computer Science and Applications, 11(2), 377-391.

Alboukaey, N., Joukhadar, A., & Ghneim, N. (2020). Dynamic Behavior Based Churn Prediction in Mobile Telecom. Expert Systems with Applications, 162, 113779. doi:10.1016/j.eswa.2020.113779

Amin, S., Guo, L., Zhang, S., & Kumar, V. (2016). Effective Oversampling Techniques for Imbalanced Data Classification in Telecommunications Industry. In 2016 13th International Conference on Service Systems and Service Management (ICSSSM) (pp. 1-6). IEEE.

ANACOM. (2017). Relatório de Mercado das Comunicações Eletrónicas 2017. [Relatório eletrónico]. Recuperado de https://www.anacom.pt/render.jsp?contentId=1489821

ANACOM. (2023). O Sector das Comunicações 2022. [Relatório eletrónico]. Recuperado de https://www.anacom.pt/render.jsp?contentId=1745016

Ballings, M., & Van den Poel, D. (2012). Customer Churn Prediction in Service Industries: Comparing Logistic Regression and Decision Trees. International Journal of Electronic Commerce Studies, 3(2), 19-30.

Ballings, M., Verhoef, P. C., & den Poot, H. (2015). Unveiling Churn Dynamics Through Visual Social Media Analytics: A Case Study of the Dutch Telecom Industry. Journal of Business Research, 68(11), 2344-2355.

Bar, H. (2014). Deep learning for real-time customer churn prediction. arXiv preprint arXiv:1401.4795.

Bergstra, J., & Bengio, Y. (2012). Random Search for Hyperparameter Optimization. Journal of Machine Learning Research, 13(2), 281-305.

Berry, M. J., & Linoff, G. S. (2004). Data Mining Techniques for Marketing, Sales, and Customer Relationship Management. Wiley.

Bhagat, R. C., & Patil, S. S. (2015). Enhanced SMOTE Algorithm For Classification Of Imbalanced Big-Data Using Random Forest. In Proceedings of IEEE International Advance Computing Conference (IACC) (pp. 1-6). IEEE. doi:10.1109/IACC.2015.7198238

Bhikha, R. (2019). Billing blunders: Errors on mobile bills cost Brits £64 million. Retrieved October 23, 2023, from https://www.uswitch.com/media-centre/2018/06/billingblunders-errors-mobile-bills-cost-brits-64-million/

Bischl, B., Binder, M., Lang, M., Pielok, T., Richter, J., Coors, S., Thomas, J., Ullmann, T., Becker, M., Boulesteix, A.-L., Deng, D., & Lindauer, M. (2023). Hyperparameter Optimization: Foundations, Algorithms, Best Practices, and Open Challenges. WIREs Data Mining and Knowledge Discovery, 13(2), e1484.

Blattberg, R. C., Kim, J., & Neslin, S. A. (2008). Customer Churn in Multi-Service Businesses: What's Driving it and How To Stop it. Harvard Business Review, 86(11), 138-145.

Brochu, E., Cora, V. M., & Freitas, N. D. (2010). A Tutorial on Bayesian Optimization of Expensive Cost Functions, with Application to Active User Modeling And Hierarchical Reinforcement Learning. Retrieved from https://arxiv.org/pdf/1012.2599.pdf

Buckinx, W., & Van den Poel, D. (2005). Customer Base Analysis: Partial Defection of Behaviorally Loyal Clients in a Non-Contractual FMCG Retail Setting. European Journal of Operational Research, 164(1), 252-268.

Burez, J., & Van den Poel, D. (2009). Customer Relationship Management. Wiley.

Burman, J. M. (1989). Cross-Validation in Statistical Models. Journal of the Royal Statistical Society. Series B (Methodological), 51(2), 239-245.

Cao, X. Y., Wei, Y. G., & Li, X. F. (2014). A Novel Hybrid Ensemble Learning Method for Imbalanced Data Classification. Applied Soft Computing, 23, 655-662.

Castro, E. G., & Tsuzuki, M. S. G. (2015). Churn Prediction in Online Games Using Players' Login Records: A Frequency Analysis Approach. IEEE Transactions on Computational Intelligence and AI in Games, 7(3), 255-265.

Correia, P. M., & Colombini, M. (2021). Transformer-Based Customer Churn Prediction for Telecommunication Services. IEEE Transactions on Computational Social Systems, 8(2), 769-780.

Chamberlain, B., Cardoso, Â., Liu, C. H., Pagliari, R., & Deisenroth, M. (2017, August). Customer Lifetime Value Prediction Using Embeddings. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1753-1762). ACM. DOI: 10.1145/3097983.3098123

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-Sampling Technique. Journal Of Artificial Intelligence Research, 16, 321-357.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2003). Outlier Detection on Categorical Data: A Data Mining Perspective. In Proceedings of the 2003 SIAM International Conference on Data Mining (pp. 151-162).

Claesen, M., & De Moor, B. (2015). Hyperparameter search in machine learning. arXiv preprint arXiv:1502.02127. [Link to arXiv preprint: https://arxiv.org/abs/1502.02127]

Coussement, K., & Van den Poel, D. (2008). Churn Prediction in Subscription Services: An Application of Support Vector Machines. Expert Systems with Applications, 35(3), 1333-1341.

De Caigny, A., Coussement, K., & Van den Poel, D. (2018). A New Logit Leaf Model for Customer Churn Prediction. Expert Systems with Applications, 97, 265-277.

Dean, D. (2007). Customer churn: How To Predict And Retain Your Customers. John Wiley & Sons.

Diro, G., & Chilamkurti, N. (2018). Customer Churn Prediction Using Deep Learning. In 2018 IEEE International Conference on Big Data (Big Data 2018) (pp. 5929-5934). IEEE.

Elreedy, D., & Atiya, A. F. (2019). A Comprehensive Analysis of Synthetic Minority Oversampling Technique (SMOTE) for Handling Class Imbalance. Information Sciences, 505, 32-64.

Fader, P. L., Hardie, B., & Shang, J. (2010). Customer Lifetime Value for Non-Contractual Relationships: the Case of the Swedish Telecom Industry. Journal of Marketing Research, 47(2), 266-282.

Faris, H. M. (2018). Hybrid PSO-ANN Model for Customer Churn Prediction in Telecommunication Industry. Journal of Network and Computer Applications, 116, 34-42.

Fang, Y. (2021). Customer Loyalty in the Telecommunications Industry: A Systematic Review ãnd Meta-Analysis. Journal of Business Research, 125, 562-574.

Feick, L., & Price, J. L. (1987). The Effect of Workplace Competition on Job Satisfaction and Turnover: a Field Study. Academy of Management Journal, 30(4), 706-723.

Galar, M., Fernández, A., García, S., & Herrera, F. (2018). Evolutionary Undersampling for Classification Tasks: Analysis and Insights on Imbalanced Data. Applied Soft Computing, 71, 1312-1329.

Ghazikhani, A., Yazdi, H. S., & Monsefi, R. (2012). Class Imbalance Handling Using Wrapper-Based Random Oversampling. In 20th Iranian Conference on Electrical Engineering (ICEE2012) (pp. 611-616). Tehran, Iran: Institute of Electrical and Electronics Engineers (IEEE). doi:10.1109/IranianCEE.2012.6292428

Hartati, S. D., Wardoyo, R. I., & Kusumawardhani, D. P. (2018). Handling Imbalanced Dataset Using SMOTE and Random Under-Sampling Techniques with Bagging Method. In 2018 6th International Conference on Electrical Engineering, Computer Science and Informatics (ICECSI) (pp. 1-4). IEEE.

He, H., & Garcia, E. A. (2011). Learning from Imbalanced Data. IEEE Transactions on knowledge and data engineering, 21(9), 1263-1294.

He, H., Bai, Y., Garcia, E. A., & Li, S. (2008). ADASYN: Adaptive Synthetic Sampling Method For Imbalanced Datasets. In IEEE International Conference on Data Mining (ICDM) (pp. 1321-1326). IEEE.

Holmlund, M. (2020). The Impact of Marketing Intelligence on Customer Satisfaction and Engagement. Journal of Marketing Management, 36(1-2), 133-160.

Huang, Y., Kamel, M. S., & Bergmann, D. R. (2012). Can We Beat the Churn? A New Approach for Customer Churn Prediction Based on Ensemble Learning. In 2012 IEEE International Conference on Data Mining (pp. 594-602). IEEE.

Huang, Y., Shen, S., & Kamel, M. S. (2015). Churn Prediction in Telecommunications: A Deep Learning Approach. Expert Systems with Applications, 42(16), 6599-6608.

IBM. (2019). Cross Industry Standard Process for Data Mining (CRISP-DM). Retrieved from https://www.ibm.com/docs/en/spss-modeler/saas?topic=dm-crisp-help-overview

Idris, A., Selamat, N. S., & Mansor, M. (2017). Customer Churn Prediction Using Genetic Programming with Adaboost Ensemble. Journal of Network and Computer Applications, 80, 103-114.

Jabeur, I., Aouamri, S., & Hammami, M. (2020). An Improved Decision Tree-Based Approach for Imbalanced Data Classification. Cognitive Systems Research, 89, 100163.

Jerath, K., Schweidel, G., & Sarma, A. (2011). Retaining Customers in The Telecom Industry: An Analysis of Churn Drivers And Retention Strategies. Journal of Decision Systems, 20(4), 472-487.

Johnson, J. M., & Khoshgoftaar, T. M. (2019). Deep Learning and Thresholding with Class-Imbalanced Big Data. In 2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA) (pp. 755-762). Boca Raton, FL, USA: IEEE. doi:10.1109/ICMLA.2019.00134

Jung, H.-W. (2018). Cross-Validation Methodology and its Applications in Machine Learning. Frontiers in Machine Learning, 3, 86.

Kaya, M. R. (2019). A Spatio-Temporal Data Mining Approach for Customer Churn Prediction in Telecommunications Industry. International Journal of Information Management, 46, 285-297.

Keramati, M., Rezvani, MR., & Rafiei, D. (2014). A Hybrid Model for Customer Churn Prediction Based on Ensemble Learning And Decision Trees. Journal of Business Research, 67(2), 251-258.

Kochański, J. (2003). Data Preparation for Data Mining. In Data mining and knowledge discovery handbook (2nd ed., pp. 545-577). Springer.

Kotler, P., & Keller, K. L. (2012). Marketing Management. Pearson Education Limited.

Krautz, C., & Hoffmann, S. (2017). Customer Tenure: A Review of the Literature and Implications for Research. International Journal of Service Industry Management, 28(2), 302-326.

Larivière, B., & Van den Poel, D. (2005). Customer Churn in Retail: A Comparison of Modeling Approaches. Expert Systems with Applications, 28(4), 595-609.

Lee, Y., Kim, K., & Park, J. (2018). Identifying Churn Factors in the Mobile Game Market: A Case Study of Korean Mobile Game Users. Internet Research, 28(5), 1070-1093. doi:10.1108/intres-02-2017-0018

Lies, P. (2023). Micro-targeting and geo-marketing in the telecommunications industry: A review of the literature. Journal of Business Research, 136, 396-407.

Lu, J., Chen, H., & Kou, G. (2014). Customer Churn Prediction for Telecommunication Services: A Deep Learning Approach Using Customer Knowledge. In 2014 IEEE International Conference on Big Data (Big Data 2014) (pp. 28-39). IEEE.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. arXiv preprint arXiv:1705.07874.

Moeyesoms, J., Martens, L., & Van den Poot, H. (2011). A Data-Driven Journey into Customer Churn: A Case Study of a Belgian Mobile Telecommunication Operator. European Journal of Operational Research, 214(3), 644-655.

Moussaoui, S., Zineb, A., & El Amrani, M. (2022). Analysis of the Impact Of Technological Advancements on the Telecommunications Industry. International Journal of Informatics and Communication Technology, 13(3), 1-15.

Moreno-Torres, J. L., Sáez-López, J. M., & Pita-López, M. D. (2009). Improving Classification Accuracy Through Tomek's Links snd GA Optimization. Pattern Recognition, 42(9), 1864-1876.

Moro, S., Cortez, P., & Rita, P. (2015). Predicting Bank Customer Churn with Churn-Diffusion Ensembles. Expert Systems with Applications, 42(13), 5718-5731.

Muneeb, M. A., Butt, A. M., & Han, J. (2019). The Impact of Overall Service Experience on Customer Loyalty in the Telecommunications Industry: A Mediating Role of Customer Satisfaction. Journal of Business Research, 102, 242-251.

Nguyen, T. M. H., & Duong, T. T. (2011). Cost-Sensitive Learning for Customer Churn Prediction in the Telecommunications Industry. In 2021 International Conference on System Science and Engineering (ICSSE) (pp. 1-5). IEEE.

Óskarsdóttir, M., Gíslason, G. D., & Axelsson, S. (2020). Relational Learning for Churn Prediction in Mobile Telecommunications. Expert Systems with Applications, 150, 113360.

Park, S. H., & Ghosh, J. (2012). Ensemble Learning for Imbalanced Data Classification: A Review. Information Systems Frontiers, 14(2), 357-381.

Pinheiro, C. (2011). Social Network Analysis in Telecommunication. John Wiley & Sons.

Qualtrics. (2023). The Impact of Face-To-Face Interaction on Customer Loyalty. Retrieved from https://www.qualtrics.com/au/experience-management/customer/customer-loyalty/

Rijnen, W. J. M. (2018). A Two-Phase Approach for Customer Churn Prediction in the Telecommunications Industry. Expert Systems with Applications, 96, 321-328.

Risselada, S. D., van der Vorst, H. C. M., Blokpoel, H., & Verhoef, P. C. (2010). Factors Influencing Customer Churn in a Digital Service Environment: A Classification Study of Mobile Telephony. Journal of Service Management, 21(5), 651-668.

S&P Global Market Intelligence. (2021). Quadruple-Play Bundling In Europe: A Strategic Analysis.

Sabbeh, T. G. M. (2018). Analysis Of Machine Learning Techniques For Customer Churn Prediction Using Behavioral Data. Journal of Big Data, 5(1), 1-23.

Sahlberg, E. (2018). Churn Modeling in the Insurance Sector—A Study of the Impact Of Address Changes and Customer Attributes on Discontinuation of Services. Expert Systems with Applications, 94, 255-265.

Sagi, S., & Rokach, L. (2018). Ensemble Learning Methods for Improving Classification Accuracy: A Review. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8(5), e1314.

Salunkhe, S. S., & Mali, M. (2016). Ensemble Learning for Imbalanced Data Classification: A Study. In International Conference on Computer Communication and Informatics (ICCCI) (pp. 1-6). IEEE.

Sharma, S., & Rajan, R. (2017). Factors Influencing Customer Churn in the Telecommunications Industry: A Review of Literature. International Journal of Applied Business and Economic Research, 15(17), 114-132.

Shen, Y-J., Huang, Y., & Li, J. (2020). A Novel Customer Churn Prediction Model Based on Feature Importance Analysis and Ensemble Learning. Expert Systems with Applications, 147, 113192. doi:10.1016/j.eswa.2020.113192

Torres, A., García, S., & Sánchez, D. (2016). SMOTE for Discrete Attributes: A New Oversampling Method for Imbalanced Data Sets in Classification Tasks. In International Conference on Intelligent Data Analysis (pp. 405-413). Springer, Cham.

Ullah, A., Ullah, Z., & Iqbal, M. T. (2016). Customer Churn Prediction Using Hybrid Machine Learning Models. In 2016 IEEE International Conference on Cybernetics and Computational Intelligence (ICCCI 2016) (pp. 1-6). IEEE.

Ullah, F., Irshad, M. I., & Khan, Z. A. (2019). The Impact of Customer Retention Strategies on Profitability in The Telecommunications Industry: A Case Study of Pakistan. International Journal of Emerging Markets, 14(4), 711-731.

Umayaparvathi, K., & Iyakutti, K. (2018). Customer Churn Prediction Using Deep Learning Classification Algorithms. In 2018 International Conference on Communication, Computing and Information Technology (ICCCIT) (pp. 1-6). IEEE.

Vafeiadis, T., Giaglis, G. M., & Koulocheras, I. (2015). Predicting Customer Churn in Telecommunications Using Machine Learning Techniques: A Comparative Study. Journal of Business Research, 104, 291-304.

Verbeke, W., Dejaeger, K., Martens, D., Mues, C., Baesens, B., & Van den Poel, D. (2012). New Insights into Churn Prediction in the Telecommunications Sector: A Profit-Driven Data Mining Approach. European Journal of Operational Research, 218(2), 211-229.

Verhoef, P. C., Kooijman, M. P., & Franses, P. H. (2009). Customer Churn and Network Externalities: A Dynamic Analysis of the Dutch Mobile Telecommunications Market. Journal of Marketing Research, 46(3), 378-390.

Viljanen, A., Etzenhofer, L., & Lääränen, M. (2016). Churn Prediction in Mobile Social Games: A Comparison of Machine Learning and Game Analytics Approaches. Entertainment Computing, 23, 34-43.

Voorhees, C. M., Brady, M. K., & Smith, A. K. (2017). Customer Lifetime Value: Theory and Practice. Routledge.

Vo, D. M., Tjoa, A. M., & Le, B. S. (2021). Customer Churn Prediction in Telecommunications Using Text Mining Techniques. International Journal of Information Management, 57, 102423.

Wang, L., Han, M., Li, X., Zhang, N., & Cheng, H. (2021). Review of Classification Methods on Unbalanced Datasets. IEEE Access, 9, 64606-64628.

Wang, G., Chen, Y., & Zhou, J. (2015). Customer Churn Prediction based on RFM And SVM. In Proceedings of the 2015 7th International Conference on Intelligent Systems and Informatics (ISIC) (pp. 361-365). IEEE.

Wassouf, T., Al-Debei, M., & Al-Shammari, F. (2020). A Comparison of Binary And Models for Customer Churn Prediction. International Journal of Information Management, 51, 102061.

Wansink, B. (2020). BuddeComm Intelligence Report: Mobile Broadband Subscriptions (Global). BuddeComm.

Witell, L., Kristensson, P., Gustafsson, A., & Löfgren, M. (2011). Customer Co-Creation in New Service Development: A Matter of Communication? Journal of Product Innovation Management, 28(2), 291-307.

Wolpert, D. H., & Macready, W. G. (1997). No Free Lunch Theorems for Optimization. IEEE Transactions on Evolutionary Computation, 1(1), 67-82.

Wong, H. Y. (2011). A New Paradigm for Customer Churn Prediction: Combining Survival Analysis and Machine Learning. Expert Systems with Applications, 38(4), 3891-3902.

Xie, Y., Xu, L., & Chen, X. (2009). Customer Churn Prediction using Improved Bayes Networks. In 2009 International Conference on Networking, Sensing and Control (ICNSC) (pp. 1-4). IEEE.

Yang, E.-J., Lee, J., & Kim, S. (2019). Predicting Customer Churn for Mobile Games Using Machine Learning Techniques. Expert Systems with Applications, 129, 114-123. doi:10.1016/j.eswa.2019.02.023

Zopounidis, C., Doumpos, M., & Iliopoulou, E. (2008). Customer Segmentation and Churn Prediction in the Greek Telecommunications Market. European Journal of Marketing, 42(5/6), 571-591.

Zhang, Y. (2020). An Improved Imbalanced KNN Algorithm for Imbalanced Data Classification. In Proceedings of the 2020 5th International Conference on Intelligent Computing and Information Science (ICICIS) (pp. 110-114). IEEE.