



INSTITUTO
UNIVERSITÁRIO
DE LISBOA

Anticipating Financial Risk: Machine Learning for Debt Management in Telecommunications

Filipe de Oliveira Faustino Arsénio

Master's in Data Science

Supervisors:

PhD António Raimundo, Assistant Professor
Iscte - University Institute of Lisbon

PhD João Pedro Pavia, Assistant Professor
Iscte - University Institute of Lisbon

september, 2025



BUSINESS
SCHOOL



TECNOLOGIAS
E ARQUITETURA

Departamento de Métodos Quantitativos para Gestão e Economia
Departamento de Ciências e Tecnologia da Informação

Anticipating Financial Risk: Machine Learning for Debt Management in
Telecommunications

Filipe de Oliveira Faustino Arsénio

Master's in Data Science

Supervisors:

PhD António Raimundo, Assistant Professor
Iscte - University Institute of Lisbon

PhD João Pedro Pavia, Assistant Professor
Iscte - University Institute of Lisbon

september, 2025

I dedicate this work to all those who supported me, but most importantly, I dedicate it to
myself.

Acknowledgments

I would like to express my sincere gratitude to my internship supervisor for the patience in helping me understand the data and for always being available to answer my questions.

I am also deeply thankful to my Masters supervisors, for guiding me throughout the development of this thesis and for being consistently supportive and present during the entire process.

To Francisca, for standing by my side in the most difficult moments, whether facing challenges in writing, problems with the code, or the struggle of balancing work and university. Your support made all the difference.

Finally, I would like to thank my family for their constant support in every aspect of my life, especially for the financial assistance, without which this journey would not have been possible.

Abstract

The telecommunications industry is characterized by intense competition and rapid technological evolution, making financial stability a critical factor for sustained growth. This dissertation focuses on leveraging machine learning techniques to analyze and predict customer payment behavior within a Portuguese telecommunications company, aiming to reduce financial losses associated with unpaid debts.

Using the CRISP-DM methodology, the project first develops supervised learning models to predict whether customers will remain good payers, based solely on internal data. Among the algorithms tested, Random Forest achieved the highest accuracy of 99%, enabling early identification of potential defaulters. Complementing this, unsupervised learning methods, specifically Principal Component Analysis for dimensionality reduction and K-Means clustering, uncover hidden behavioral segments within the customer base. The optimal clustering identified five distinct groups, some of which show near-homogeneous target values (close to 0 or 1), allowing for strong characterization of compliant and non-compliant profiles.

The findings demonstrate the effectiveness of combining supervised and unsupervised learning for risk analysis. Supervised models allow scenario testing by altering feature values to simulate changes in payment behavior. In unsupervised learning, analyzing ambiguous clusters through comparison with more definitive ones helps estimate likely client outcomes and supports proactive management.

Future work may explore focused clustering of non-compliant clients, alternative data preprocessing, and time series forecasting to further improve predictive accuracy and operational utility.

Keywords: *Supervised Learning, Unsupervised Learning, Financial data, Telecommunications, Risk Management*

Resumo

A indústria das telecomunicações é caracterizada por uma intensa concorrência e rápida evolução tecnológica, tornando a estabilidade financeira um fator crítico para um crescimento sustentado. Esta dissertação foca-se na utilização de técnicas de machine learning para analisar e prever o comportamento de pagamento dos clientes numa empresa de telecomunicações portuguesa, com o objetivo de reduzir as perdas financeiras associadas a dívidas não pagas.

Utilizando a metodologia CRISP-DM, o projeto desenvolve primeiro modelos supervisionados para prever se os clientes continuarão a ser bons pagadores, com base apenas em dados internos. Entre os algoritmos testados, a Random Forest alcançou a maior precisão, com 99%, permitindo a identificação precoce de potenciais incumpridores. Complementarmente, métodos de aprendizagem não supervisionada, nomeadamente análise de componentes principais para redução de dimensionalidade e clustering com K-Means, revelaram segmentos comportamentais ocultos na base de clientes. O clustering ótimo identificou cinco grupos distintos, alguns dos quais apresentam valores da variável alvo quase homogêneos (próximos de 0 ou 1), permitindo uma forte caracterização dos perfis de clientes cumpridores e incumpridores.

Os resultados confirmam a eficácia da combinação de supervised e unsupervised learning na análise de risco. Os modelos supervised permitem simular cenários, alterando variáveis para refletir mudanças no comportamento de pagamento. Já a aprendizagem unsupervised possibilita estudar clusters ambíguos, comparando-os com grupos mais definidos, o que ajuda a estimar resultados prováveis e a apoiar a gestão proativa de clientes.

Trabalhos futuros poderão focar-se em clustering dirigido a incumpridores, novas estratégias de pré-processamento e previsões com séries temporais, visando aumentar a precisão preditiva e a utilidade operacional.

Palavras-Chave: *Aprendizagem Supervisionada, Aprendizagem Não Supervisionada, Dados Financeiros, Telecomunicações, Gestão de Risco*

Table of Contents

Acknowledgments	iii
Abstract	v
Resumo	vii
List of Figures	xiii
List of Tables	xv
List of Acronyms	xix
Chapter 1. Introduction	1
1.1. Context	1
1.2. Motivation	2
1.3. Research questions	3
1.4. Objectives	3
1.5. Methodology	4
Chapter 2. Literature Review	7
2.1. Related Work	8
2.1.1. Supervised and unsupervised learning techniques	8
2.1.2. Supervised and unsupervised learning applied to financial data	12
2.2. Methodology	14
2.2.1. PRISMA Methodology	14
2.2.2. Search and selection process	15
2.3. Results and findings	18
2.3.1. Supervised Learning	18
2.3.2. Unsupervised Learning	22
2.3.3. Supervised and unsupervised learning combined	23
2.4. Chapter Overview	25

2.5. Different Approach	26
Chapter 3. Data Analysis	29
3.1. Business Understanding	29
3.2. Data Understanding	29
3.3. Data Preparation	31
3.3.1. Missing values and New columns	31
3.3.2. Outliers	33
3.3.3. Normalization	34
3.3.4. Transformation of categorical variables to numerical	35
3.3.5. Target variable	36
3.3.6. Class Balancing	36
3.4. Final dataset overview	37
Chapter 4. Modeling and Results	39
4.1. Supervised learning models	39
4.1.1. Dataset preparation	39
4.1.2. Logistic regression	41
4.1.3. Random forest	43
4.1.4. k-NN	43
4.1.5. Neural Network	44
4.1.6. Best supervised learning model	46
4.1.7. Feature importance analysis	47
4.2. Unsupervised learning models	47
4.2.1. PCA Suitability Analysis	48
4.2.2. Choice of number of PC	49
4.2.3. Interpreting of the Principal Components	50
4.2.4. t-SNE method	52
4.3. Clustering the data	53
4.3.1. K-Means	53
4.3.2. Hierarchical clustering	53
4.3.3. DBSCAN	54
4.3.4. Optimal number of clusters	55

4.3.5. Interpreting and naming the clusters	55
4.4. Second experiment	61
4.4.1. Supervised Learning models	62
4.4.2. Unsupervised Learning models	62
4.5. Comparison with Existing Literature	62
Chapter 5. Discussions and Conclusions	65
5.1. Overview of the Dissertation	65
5.2. Discussion of Results	65
5.3. Conclusion	67
5.4. Recommendations for Future Work	67
Bibliography	69
Appendix A. Literature Review	79
Appendix B. Data Analysis	83
B.1. Variables description	84
B.2. Variable '60 a 90' relevance	86
Appendix C. Modeling and Results	91
C.1. Correlations	92
C.2. Supervised Learning	96
C.3. Unsupervised learning	102
C.4. Second experiment	119

List of Figures

Figure 1.1	CRSIP-DM [12]	4
Figure 2.1	Demonstration of supervised learning logic, illustrating the process of model training and prediction. Source: [17]	8
Figure 2.2	Demonstration of unsupervised learning logic, illustrating the process of data clustering and pattern discovery. Source: [20] . . .	10
Figure 2.3	Article selection process following Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) methodology. Adapted from [34]	17
Figure 3.1	Percentage of missing values for each variable in the dataset. . . .	31
Figure 3.2	Class distribution of the target variable	37
Figure 4.1	Heatmap of correlations between variables	40
Figure 4.2	Neural network accuracy (right) and loss (left) graphics	45
Figure 4.3	Scree plot showing the eigenvalues of each Principal Component (PC)	49
Figure 4.4	3D visualization of K-Means clusters top-down perspective	61
Figure C.1	Correlation between the variables and the target variable	95
Figure C.2	Top15 feature importance of logistic regression model	99
Figure C.3	Top15 feature importance of random forest model	100
Figure C.4	Top15 feature importance of k-Nearest Neighbors (k-NN) model .	100
Figure C.5	Top15 feature importance of neural network model	101
Figure C.6	t-SNE - Scatter plot with 100,000 observations	104
Figure C.7	t-SNE - Scatter plot with 5,000 observations	105
Figure C.8	Elbow method	107
Figure C.9	Silhouette method	107
Figure C.10	Silhouette Plot for k-means with k=3 Clusters	108
Figure C.11	Silhouette Plot for k-means with k=4 Clusters	108

Figure C.12	Silhouette Plot for k-means with k=5 Clusters	109
Figure C.13	Silhouette Plot for k-means with k=6 Clusters	109
Figure C.14	Dendogram of the Ward linkage method	110
Figure C.15	Silhouette plot for Ward method with k=3 Clusters	110
Figure C.16	Silhouette plot for Ward method with k=4 Clusters	111
Figure C.17	Silhouette plot for Ward method with k=5 Clusters	111
Figure C.18	Silhouette plot for Ward method with k=6 Clusters	112
Figure C.19	Silhouette plot for Ward method with k=7 Clusters	112
Figure C.20	Dendogram of the Complete linkage method	113
Figure C.21	Silhouette plot for Complete linkage method with k=3 Clusters .	113
Figure C.22	Silhouette plot for Complete linkage method with k=4 Clusters .	114
Figure C.23	Silhouette plot for Complete linkage method with k=5 Clusters .	114
Figure C.24	Silhouette plot for Complete linkage method with k=6 Clusters .	115
Figure C.25	Silhouette plot for Complete linkage method with k=7 Clusters .	115
Figure C.26	Cluster formation with the DB-SCAN method	116
Figure C.27	3D visualization of K-Means clusters diagonal perspective	117
Figure C.28	3D visualization of K-Means clusters side perspective	118
Figure C.29	3D visualization of K-Means clusters opposite side perspective . .	119
Figure C.30	Neural Network - accuracy (right) and loss (left) curves using common features only	123

List of Tables

Table 4.1	Summary of classification metrics for the test set across models (class 1 as positive)	46
Table 4.2	Variables that appear in 4 and 3 models	47
Table 4.3	Silhouette scores obtained by algorithm and number of clusters . .	55
Table 4.4	Average PC scores, debt fulfillment percentage, and observation distribution per cluster	56
Table 4.5	Average original variable values per cluster (non-normalized) . . .	58
Table A.1	Overview of articles on churn prediction using supervised learning in financial data in telecommunications	80
Table A.2	Overview of articles on fraud detection using supervised learning in financial data in telecommunications	81
Table A.3	Overview of articles on profitability prediction using supervised learning in financial data in telecommunications	81
Table A.4	Overview of unsupervised learning applications in financial data for telecommunications	81
Table A.5	Overview of supervised and unsupervised learning techniques com- bined in financial data for telecommunications	82
Table B.1	Datasets variables description	84
Table B.2	Final dataset variables description	85
Table B.3	Logistic regression - Confusion Matrix for Train Set ('with 60 a 90')	86
Table B.4	Logistic regression - Classification Report for Train Set ('with 60 a 90')	86
Table B.5	Logistic regression - Confusion Matrix for Test Set ('with 60 a 90')	86
Table B.6	Logistic regression - Classification Report for Test Set ('with 60 a 90')	86

Table B.7	Random Forest - Confusion Matrix for Train Set ('with 60 a 90')	87
Table B.8	Random Forest - Classification Report for Train Set ('with 60 a 90')	87
Table B.9	Random Forest - Confusion Matrix for Test Set ('with 60 a 90')	87
Table B.10	Random Forest - Classification Report for Test Set ('with 60 a 90')	87
Table B.11	k-NN - Confusion Matrix for Train Set ('with 60 a 90')	88
Table B.12	k-NN - Classification Report for Train Set ('with 60 a 90')	88
Table B.13	k-NN - Confusion Matrix for Test Set ('with 60 a 90')	88
Table B.14	k-NN - Classification Report for Test Set ('with 60 a 90')	88
Table B.15	Neural Network - Confusion Matrix for Test Set ('with 60 a 90')	88
Table B.16	Neural Network - Classification Report for Test Set ('with 60 a 90')	89
Table C.1	Pairs of variables with correlations greater than 0.7 or less than -0.7	92
Table C.2	Correlation of variables with the target variable	93
Table C.3	Logistic regression - Confusion Matrix for Train Set	96
Table C.4	Logistic regression - Classification Report for Train Set	96
Table C.5	Logistic regression - Confusion Matrix for Test Set	96
Table C.6	Logistic regression - Classification Report for Test Set	96
Table C.7	Random Forest - Confusion Matrix for Train Set	97
Table C.8	Random Forest - Classification Report for Train Set	97
Table C.9	Random Forest - Confusion Matrix for Test Set	97
Table C.10	Random Forest - Classification Report for Test Set	97
Table C.11	k-NN - Confusion Matrix for Train Set	97
Table C.12	k-NN - Classification Report for Train Set	98
Table C.13	k-NN - Confusion Matrix for Test Set	98
Table C.14	k-NN - Classification Report for Test Set	98
Table C.15	Neural Network - Confusion Matrix for Test Set	98
Table C.16	Neural Network - Classification Report for Test Set	99
Table C.17	Eigenvalues of Principal Components	102
Table C.18	Cumulative explained variance by component	103
Table C.19	PC loadings	106
Table C.20	Logistic Regression - Confusion Matrix for Train Set using common features only	119

Table C.21	Logistic Regression - Classification Report for Train Set using common features only	120
Table C.22	Logistic Regression - Confusion Matrix for Test Set using common features only	120
Table C.23	Logistic Regression - Classification Report for Test Set using com- mon features only	120
Table C.24	Random Forest - Confusion Matrix for Train Set using common features only	120
Table C.25	Random Forest - Classification Report for Train Set using common features only	121
Table C.26	Random Forest - Confusion Matrix for Test Set using common features only	121
Table C.27	Random Forest - Classification Report for Test Set using common features only	121
Table C.28	k-NN - Confusion Matrix for Train Set using common features only	121
Table C.29	k-NN - Classification Report for Train Set using common features only	122
Table C.30	k-NN - Confusion Matrix for Test Set using common features only	122
Table C.31	k-NN - Classification Report for Test Set using common features only	122
Table C.32	Neural Network - Confusion Matrix for Test Set using common features only	122
Table C.33	Neural Network - Classification Report for Test Set using common features only	123

List of Acronyms

ANN	Artificial Neural Network
AUC	Area Under Curve
BiGRU	Bidirectional Gated Recurrent Unit
BTG	Bridge to Graph
CNN	Convolutional Neural Network
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
DCNN	Deep Convolutional Neural Network
DT	Decision Trees
ECLAT	Equivalence Class Clustering and bottom-up Lattice Traversal
FCLCNN-LSTM	Fully Connected Layer Convolutional Neural Network - Long Short Term Memory
FCM	Fuzzy C-Means
GBM	Gradient Boosting Machine
GK	Gustafson-Kessel
KMO	Kaiser-Meyer-Olkin
k-NN	k-Nearest Neighbors
LSTM	Long Short-Term Memory
MOEEC	Multi-Objective Evolutionary Ensemble Clustering
PCA	Principal Component Analysis
PC	Principal Component
PCM	Possibilistic C-Means
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RF	Random Forest
RNN	Recurrent Neural Network
RPART	Recursive PARTitioning
RRF	Restricted Random Forest

RST	Rough Set Theory
SVM	Support Vector Machine
t-SNE	t-Distributed Stochastic Neighbor Embedding
UL	Unsupervised Learning
XGBoost	Extreme Gradient Boosting
NPV	Negative Predictive Value

CHAPTER 1

Introduction

1.1. Context

The telecommunications industry plays a critical role in the global economy, serving as the backbone of connectivity and enabling numerous other industries to thrive. As stated by the authors of [1, p. 1], “The telecommunications sector has become one of the main industries in developed countries. The technical progress and the increasing number of operators raised the level of competition. Companies are working hard to survive in this competitive market depending on multiple strategies.” This extremely competitive industry is intensified by “rapid changes, market liberalization, technological advancements, and diverse attractive plans” [2, p. 102], making financial stability crucial for sustaining growth and innovation. It is because of this competitiveness that the ability to analyze financial data effectively becomes a crucial factor for telecommunications companies to make informed decisions, optimize operations, and maintain competitiveness in order to be at the top of the industry [3].

Supervised learning, a subset of machine learning, has emerged as a valuable tool for analyzing financial data, enabling companies to make data-driven decisions [4]. These techniques have been widely applied in various domains, showcasing their versatility and effectiveness in addressing complex financial problems. Supervised learning provides a structured framework to uncover patterns and relationships in financial data, allowing telecommunications companies to predict customer churn [5], optimize credit risk management [6], and enhance profitability forecasts [7], among others.

While supervised learning has been highly effective in financial data analysis, it is not without limitations. One of its primary challenges is the need for labeled data. This is where unsupervised learning comes into play, offering a complementary approach to uncover hidden structures in data without requiring explicit labels [8], and proving essential for analyzing complex datasets by enabling pattern discovery, dimensionality reduction, and clustering, with applications not only in telecommunications and finance, but also

in medicine, education, biology, computer science, and social sciences [9]. Moreover, as highlighted by the authors of [8], such techniques are particularly valuable in domains like image recognition, speech processing, and anomaly detection, where extracting meaningful insights from unlabeled data is crucial.

In telecommunications, unsupervised learning is widely applied in churn prediction, as it helps companies identify patterns in customer behavior. By using dimensionality reduction and clustering techniques, businesses can group customers based on their usage patterns and service preferences, which in turn enables the development of more targeted retention strategies [10]. Typically, customer data encompasses various attributes, such as account duration, international usage, and customer service interactions, all of which contribute to identifying distinct behavioral segments [9]. By uncovering hidden patterns within these datasets, unsupervised learning allows companies to anticipate churn risk more effectively and, as noted by the authors of [9, p. 832], “...improve their marketing strategies which indulge in profit maximization.”

Given that financial data analysis is crucial for the sustainability and competitiveness of telecommunications companies [3], and that both supervised [5] and unsupervised learning techniques [8] provide valuable insights into data patterns, the combination of these elements becomes highly relevant.

1.2. Motivation

This project is driven by both practical needs and academic innovation. It addresses a significant issue faced by the telecommunications company: financial losses and operational costs resulting from unpaid debts. Currently, the company assigns debt scores to customers based on data provided by an external entity with access to broader financial information. While this system is effective, it only comes into play once a history of delinquency has been established, and relies on third-party data.

The innovation of this project lies in the prediction of the likelihood of customers being good payers using only internal data, even for customers who currently have no overdue amounts, such as customers in the dataset who have active services or installment plans and pay regularly, meaning they have no debt in arrears. By analyzing customer characteristics and early financial signals, such as current payment behaviors and contract details, the model seeks to identify patterns associated with future delinquency before

external scoring is available. This enables the company to assess risk earlier, using its own data, and make more informed decisions regarding customer engagement, service provision, and debt management strategies.

Another motivation behind this work is to explore the potential value that can be extracted from the company’s extensive internal dataset, which contains millions of records and a wide variety of variables. These data are often underutilized, despite representing a rich source of information about customer behavior and profiles.

1.3. Research questions

This project is guided by the following key research questions:

- (1) How accurately can a supervised learning model predict whether a customer will remain a good payer?
- (2) What behavioral clusters can be identified within the customer base, and how do these relate to observed payment performance?
- (3) Can supervised and unsupervised learning techniques be combined in a complementary way to improve the reliability of customer risk assessments?

1.4. Objectives

The main objectives of this project are threefold. First, to develop a supervised learning model capable of predicting whether a customer is likely to remain a good payer or not.

Second, to reduce the dimensionality of the dataset and apply clustering techniques in order to uncover hidden patterns in the data using unsupervised learning methods.

Third, to explore how both supervised and unsupervised learning approaches can be used together in a complementary manner to generate added value. This integrated approach consists of predicting whether a customer is a good payer and then verifying which cluster they belong to, in order to assess whether they are grouped with typical good payers or bad payers, allowing for a dual confirmation of the prediction and opens the door to more informed actions. For instance, if a customer is predicted to be a good payer but falls within a cluster of bad payers, the company can take preventive measures such as closer monitoring. Conversely, if a customer is predicted to be a bad payer but

belongs to a cluster of good payers, it may be worth offering incentives or support to help reverse the negative trend.

1.5. Methodology

The methodology adopted for this dissertation is the Cross Industry Standard Process for Data Mining (CRISP-DM), a structured and iterative framework developed in 1996 by a consortium of companies, and widely recognized for its flexibility and applicability across different business domains [11]. As illustrated in Figure 1.1, CRISP-DM consists of six main phases and is represented as an iterative cycle, allowing continuous revisions between phases.



FIGURE 1.1. CRSIP-DM [12]

The six phases of CRISP-DM are [11]:

- (1) **Business Understanding:** This phase involves understanding the project objectives and business needs. The problem to be solved is defined, and success criteria for the model are established.
- (2) **Data Understanding:** In this phase, data is collected and explored to understand its structure, quality, and potential limitations.

- (3) Data Preparation: This step includes data cleaning, transformation and selection of the most relevant variables for modeling in order to ensure that the data is properly prepared for modeling.
- (4) Modeling: Various machine learning algorithms are applied to the data, with tests to different configurations and approaches to optimize performance.
- (5) Evaluation: Once the models are built, their performance is evaluated through various metrics to ensure they meet business objectives, with possible adjustments being made before proceeding to deployment.
- (6) Deployment: The final phase involves implementing the model in a production environment, making insights available for decision-making.

CHAPTER 2

Literature Review

The prediction of customer payment behavior has become increasingly important in credit management, especially as financial services become more digitalized, making it essential to understand how customers financial characteristics influence risk and decision-making processes.

In this context, the literature review aims to provide a comprehensive analysis of both theoretical and practical approaches to predicting debt payment, with a focus on supervised, unsupervised, and hybrid techniques applied to customer financial behavior. The objective is to explore existing methodologies, including the use of historical financial data and dimensionality reduction techniques, in order to identify patterns and enhance predictive models.

This literature review is guided by the following research questions:

- What supervised learning approaches have been applied to financial data in the telecommunications sector?
- What unsupervised learning approaches have been used to analyze financial behavior in telecommunications?
- How have supervised and unsupervised learning techniques been combined to model financial data in the telecommunications industry?

With these research questions in mind, the structure of this literature review will first provide background information on supervised, unsupervised and hybrid techniques, including their application to financial data. Following this, a review of relevant articles will be presented, highlighting what has been done in this field. Finally, the approach proposed in this dissertation will be discussed.

2.1. Related Work

2.1.1. Supervised and unsupervised learning techniques

Supervised learning is a foundational area of machine learning that uses labeled datasets to train algorithms in mapping inputs to corresponding outputs. This process enables models to make predictions on new, unseen data. Tasks within supervised learning are broadly categorized as classification (for example, predicting customer churn or fraud detection) or regression (for example, forecasting sales or financial metrics), with these methods being widely applied across various fields such as finance, business, and telecommunications [13], [14], [15], [16]. Figure 2.1 illustrates this process, demonstrating how a model learns from labeled data to generate predictions on new inputs.

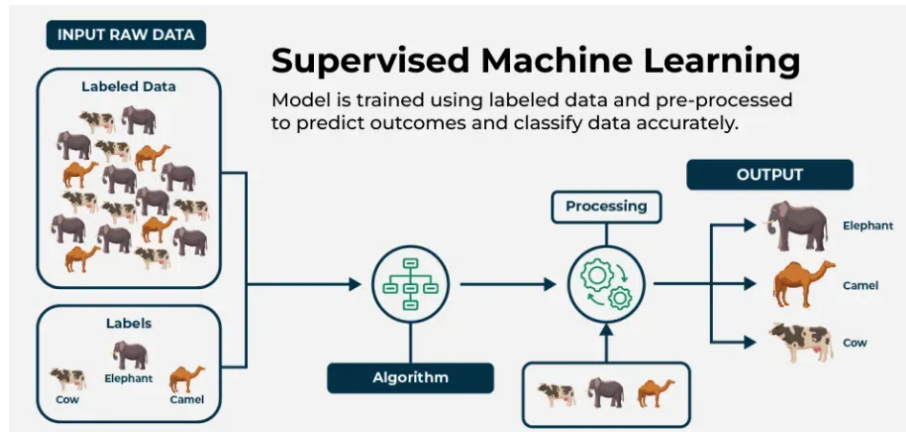


FIGURE 2.1. Demonstration of supervised learning logic, illustrating the process of model training and prediction. Source: [17]

Supervised learning includes several key techniques that are widely used across various fields of research [4], such as:

- **Linear Regression:** This algorithm predicts continuous values by modeling the relationship between variables as a straight line. It is commonly applied in financial forecasting and employee turnover prediction [4], [13], [14].
- **Logistic Regression:** Designed for binary classification tasks, logistic regression estimates the probability of an outcome (for example, churn or credit approval) using a logistic function [15], [18].
- **k-NN:** k-NN classifies data points based on the majority class of their nearest neighbors. While effective for small datasets, it faces scalability issues with larger data volumes, often seen in telecom or financial applications [15], [18].

- Decision Trees (DT): These models use a tree-like structure to split data based on features, providing interpretable predictions. However, they can overfit complex datasets if not pruned adequately [14], [18].
- Random Forest (RF): By averaging predictions from multiple decision trees, Random Forest reduces overfitting and improves robustness, making it ideal for fraud detection and credit scoring [19].
- Gradient Boosting Machine (GBM) and Extreme Gradient Boosting (XGBoost): These methods sequentially build models to minimize errors, excelling in high-accuracy domains like financial risk assessment and customer behavior prediction [18].
- Support Vector Machine (SVM): SVM classifies data by finding the optimal hyperplane that separates classes, making them effective for structured prediction problems such as churn detection and credit scoring [4], [13].
- Naïve Bayes: A probabilistic model using Bayes' theorem, Naïve Bayes is often applied in spam detection and text classification. Its simplicity makes it efficient for high-dimensional data [15].
- Artificial Neural Network (ANN): Inspired by the human brain, ANNs handle non-linear and high-dimensional data, frequently used in financial risk modeling and customer segmentation [19].
- Convolutional Neural Network (CNN): Primarily designed for image data, CNNs also find applications in financial risk analysis involving visual or structured data [15].
- Recurrent Neural Network (RNN): RNNs, especially Long Short-Term Memory (LSTM) networks, process sequential data effectively, such as financial time series or customer transactions [19].

Therefore, supervised learning encompasses a variety of robust techniques widely applied across domains such as predictive modeling, classification, and regression, demonstrating their adaptability to diverse research and industry challenges [4], [14], [16].

Unsupervised learning is a machine learning approach that enables a system to identify patterns and relationships in datasets without the need for predefined labels. Unlike supervised learning, where models are trained with labeled data to learn specific

associations, unsupervised learning explores data autonomously, detecting hidden structures without direct human intervention, as illustrated in Figure 2.2. As highlighted by the authors of article [8, p. 911], “unsupervised machine learning, unlike supervised learning, does not need the presence of a person to oversee the model.”

Furthermore, the authors of article [8, p. 911] state that “Unsupervised Learning (UL) algorithms are better at solving complicated tasks than supervised learning algorithms. However, supervised learning models provide more accurate results because a programmer explicitly teaches the system what to search for in the data presented.” So, when choosing which technique to use, the trade-off between flexibility and accuracy needs to be considered.

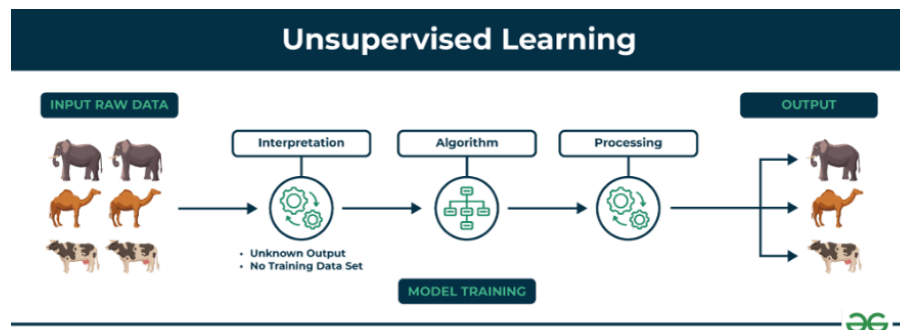


FIGURE 2.2. Demonstration of unsupervised learning logic, illustrating the process of data clustering and pattern discovery. Source: [20]

Unsupervised learning includes several key techniques that are widely used across various research fields, such as:

- **Hierarchical Clustering:** This method organizes data into a tree-like structure using a dendrogram to illustrate relationships between clusters [10], [21]. Common approaches include:
 - **Single Linkage:** Defines clusters based on the shortest distance between points [10].
 - **Average Linkage:** Computes the average distance between clusters to determine grouping [10].
 - **Complete Linkage:** Merges clusters based on the farthest distance between points [10].
- **k-Means Clustering:** A widely used partition-based clustering algorithm that assigns data points to k predefined clusters by minimizing intra-cluster variance.

It is effective for customer segmentation and document classification but is sensitive to initial centroid selection [10], [21].

- **k-Medoids Clustering:** Similar to k-means but selects actual data points (medoids) as cluster centers, making it more robust to outliers and noise. It is often applied in medical diagnosis and market analysis [10].
- **Fuzzy C-Means (FCM):** Unlike k-means, where each data point belongs strictly to one cluster, FCM allows partial membership, assigning probabilities to clusters. This approach is frequently used in image segmentation and medical data analysis [10], [21].
- **Gustafson-Kessel (GK) Clustering:** A variation of fuzzy clustering that adapts cluster shapes based on data distribution, making it suitable for complex datasets with elliptical structures [10].
- **Possibilistic C-Means (PCM):** An extension of FCM that reduces sensitivity to noise by considering cluster memberships independently, often applied in medical imaging and anomaly detection [10].
- **Density-Based Spatial Clustering of Applications with Noise (DBSCAN):** A density-based clustering algorithm that groups points based on the density of neighboring points. It does not require the number of clusters to be predefined and is effective for datasets with irregular shapes and outliers, being that it is particularly useful in geospatial data analysis and identifying noise in datasets [10], [21].
- **Apriori Algorithm:** A classic algorithm for mining frequent itemsets and learning association rules in transactional databases [8].
- **Equivalence Class Clustering and bottom-up Lattice Traversal (ECLAT) Algorithm:** An efficient algorithm for mining frequent itemsets, known for its use of vertical data formats and intersection operations [8].
- **Frequent Pattern Growth Algorithm:** An algorithm that identifies frequent itemsets without the need for candidate generation, therefore improving efficiency compared to Apriori [8].

The combination of supervised and unsupervised learning techniques is highly beneficial for enhancing the performance of machine learning models [22]. By leveraging the strengths

of both approaches, it is possible to achieve better accuracy and adaptability in complex tasks. Supervised learning, with its ability to predict outcomes based on labeled data, is effective for tasks like classification and regression. On the other hand, unsupervised learning allows the model to discover hidden patterns and structures in data without predefined labels, making it ideal for exploring large datasets and uncovering relationships that might not be immediately apparent [23], [24].

There are two techniques that combine supervised and unsupervised learning techniques:

- **Ensemble approach:** This technique combines multiple individual models (unsupervised or supervised) to improve overall performance. By aggregating predictions from several models, ensemble methods increase accuracy, robustness, and generalization, making them highly effective for complex tasks like churn prediction [23], [24].
- **Hybrid approach:** According to the authors of article [22, p. 1], “Ensemble machine learning techniques involve combining multiple models to enhance predictive performance.” This approach is often used to improve model accuracy by combining the strengths of both unsupervised and supervised learning techniques.

2.1.2. Supervised and unsupervised learning applied to financial data

After introducing the fundamental concepts of supervised and unsupervised learning, it is important to explore their practical applications in financial data analysis. Supervised learning techniques are widely used in forecasting, risk assessment, and fraud detection, enabling institutions to extract actionable insights from historical datasets to predict future outcomes. Financial data are often high-dimensional, heterogeneous, and imbalanced, making the use of advanced supervised learning methods both challenging and valuable [25], [26], [27].

Credit scoring is a key area for supervised learning. While logistic regression remains popular for its interpretability, modern machine learning models such as Random Forest, GBM, and SVM offer higher accuracy by capturing non-linear relationships in the data [25], [26], [28].

Detecting fraudulent transactions is another critical application. Machine learning models, including Naïve Bayes, SVM, and deep learning architectures like LSTM and

CNN, are widely used for identifying anomalies in transaction patterns [27]. A notable example is the integration of XGBoost with Bidirectional Gated Recurrent Unit (BiGRU) and self-attention networks for internet loan fraud detection, which achieved exceptional accuracy and precision on benchmark datasets [29], [30]. These methods address challenges such as data imbalance by leveraging techniques like outlier scoring and feature engineering.

Supervised learning has also revolutionized financial market forecasting. Time-series models such as Support Vector Regression are used to predict stock prices, debt levels, and other financial indicators. In the Montenegrin electricity industry, this model outperformed traditional models by efficiently capturing non-linear dependencies in data, demonstrating its efficacy in debt risk prediction [25], [28].

Deep learning models have shown significant promise in financial applications. Architectures like LSTMs and CNNs are increasingly employed for processing sequential and high-dimensional data, as they are capable of capturing temporal dependencies and extracting hierarchical features, which is particularly valuable in financial contexts where customer behavior evolves over time and data often contains diverse patterns. For instance, LSTM combined with attention mechanisms has proven effective in credit risk modeling by accurately identifying complex patterns in customer behavior [28], [29]. Similarly, hybrid approaches that combine feature selection with deep learning models have enhanced both interpretability and performance in financial forecasting [30].

The application of supervised learning extends to the insurance sector, particularly in predicting lapses in mortgage life insurance policies. Models such as Random Forest, XGBoost, and ANN have been utilized to identify clients at higher risk of policy cancellation [31].

Unsupervised learning techniques are particularly useful in scenarios where labeled data is scarce or when identifying hidden structures can enhance decision-making. In churn prediction, these methods play a crucial role by revealing customer behavior patterns that are not explicitly labeled, allowing for more strategic interventions. A fundamental application of unsupervised learning in churn prediction is customer segmentation through clustering. By analyzing customer usage patterns and service preferences, clustering techniques such as K-Means, DBSCAN, and hierarchical clustering help classify users into

distinct behavioral groups, enabling telecommunications companies to tailor retention strategies and improve customer satisfaction [9].

Beyond customer segmentation, anomaly detection is another critical application of unsupervised learning in churn prediction. By identifying irregular customer behaviors that may indicate an increased likelihood of churn, techniques like Rough Set Theory and density-based clustering help detect deviations in customer activity, providing early warnings for proactive intervention [32].

While supervised and unsupervised learning techniques each offer unique advantages in churn prediction, combining both approaches has proven to be highly effective in enhancing predictive accuracy and decision-making. Hybrid models leverage the strengths of supervised learning in classification and prediction, while incorporating unsupervised methods for feature engineering, data preprocessing and better interpretability. A hybrid approach that integrates supervised and unsupervised learning has shown great effectiveness in churn prediction, as shown by the authors of article [33], where clustering techniques were applied in order to segment customers based on their behavior, and then used supervised learning models, like SVM and Recursive PARTitioning (RPART), to predict churn within each segment, improving prediction accuracy.

Additionally, ensemble models that combine clustering and classification have been proposed to enhance churn prediction performance, as demonstrated by the authors of articles [23] and [24], where a combination of clustering and classification algorithms was used to improve the predictive power and robustness of churn prediction models in the telecom industry. These models first cluster customers based on shared behavioral traits and then use supervised classifiers to predict churn likelihood within each cluster, improving interpretability and precision.

By combining supervised and unsupervised approaches, telecom companies can improve their decision-making regarding customer debt payment behavior.

2.2. Methodology

2.2.1. PRISMA Methodology

The PRISMA methodology is a standardized guideline designed to enhance the transparency and rigor of systematic reviews and meta-analyses. Initially developed in 2009

and updated in 2020, this methodology assists researchers in reporting systematic reviews comprehensively by addressing why a review was conducted, the methods used, and the findings [34]. The protocol includes a 27-item checklist and a flow diagram to document each stage of the study selection process, from initial identification to final inclusion, which ensures consistency and reduces bias [35].

PRISMA 2020 was updated to address concerns about the quality of systematic reviews and meta-analyses, which had declined due to the rapid increase in these publications. With the growing availability of literature review tools, many studies were conducted with less rigor, impacting their reliability and reproducibility. PRISMA 2020 provides stricter guidelines to ensure that the rising volume of publications maintains high standards [35]. Moreover, this methodology is widely endorsed for its role in evidence-based research, providing a standardized yet adaptable structure proven effective in fields such as medicine, management, and tourism [36].

This methodology is particularly advantageous for ensuring that systematic reviews are transparent and reproducible, as it minimizes bias by applying structured protocols across all review stages. By mandating clear documentation at each step, such as study identification, screening, eligibility, and inclusion, researchers are able to present a comprehensive overview of existing literature with high methodological rigor [34], [36].

In this literature review, PRISMA will be applied to screen and select articles for the literature review on supervised and unsupervised learning in financial data, particularly within the telecommunications sector. This approach will ensure a systematic and replicable process, providing a robust and unbiased foundation for further research in this domain.

2.2.2. Search and selection process

The search and selection of articles for this systematic literature review followed the PRISMA methodology to, as the authors of article [35] claim, ensure transparency and rigor in the process of study selection. The search began by identifying groups of keywords related to the research question:

- Telecommunications: ("Telecom data", "Telco data", "Internet Service Provider").
- Financial Data: ("Debts", "Client", "Churn", and "Financial data").

- Machine Learning: ("Supervised Learning", "Unsupervised Learning", "Classification", "Artificial Intelligence", "Predictive Modeling", "Clustering", "Dimensionality reduction").

After testing various combinations of these keywords, it was determined that the following queries returned the most focused and relevant articles:

- (Telecommunications data OR "Telco data" OR "Telecom data") AND ("Supervised Learning") AND ("Financial data" OR "debt data" OR "Churn")
- (Telecommunications data OR "Telco data" OR "Telecom data") AND ("Unsupervised Learning" OR "Clustering" OR "Cluster" OR "Dimensionality reduction") AND ("Financial data" OR "debt data" OR "Churn")

Figure 2.3 summarizes the entire process, from the collection of articles via Scopus and Web of Science to the final selection of articles included in the review.

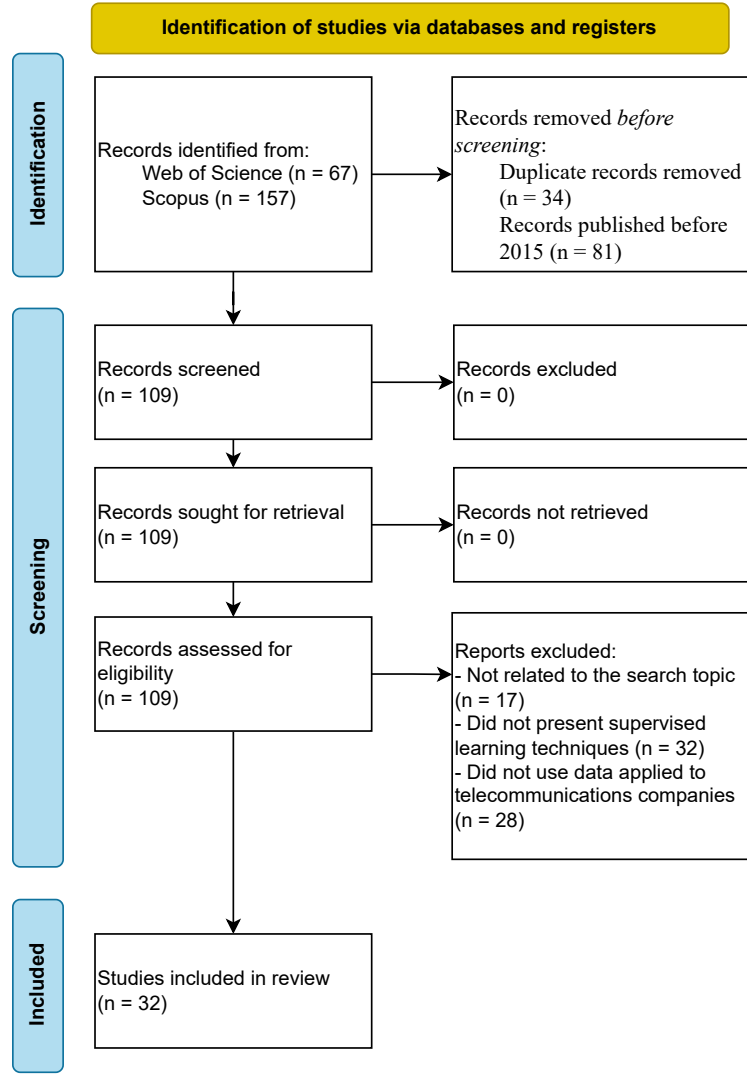


FIGURE 2.3. Article selection process following PRISMA methodology. Adapted from [34]

These queries retrieved 88 and 136 articles (224 in total), respectively, being that 67 were from Web of Science and 157 from Scopus. The Scopus and Web of Science databases were accessed on 04/02/2025, ensuring that the most recent articles up to that date were included. Following the removal of articles published before 2015 (81 articles) and duplicates (34), 109 articles remained for further evaluation.

The exclusion of studies before 2015 was justified by the significant advancements in machine learning techniques, computational power, and big data infrastructure over the past decade. Articles published prior to 2015 may not reflect the most recent advances in

supervised learning techniques, data availability, and computational frameworks. Moreover, the rapid growth in big data technologies and increased computational power has enabled the development and implementation of more sophisticated models that were previously infeasible. Furthermore, limiting the scope to recent studies ensures the inclusion of research that reflects the state-of-the-art techniques and addresses current challenges faced by these industries [37].

Of these 109, 18 articles were inaccessible due to institutional subscription limitations. These articles were retrieved via ResearchGate, an alternative platform to Scopus and Web of Science, bringing the total number of articles available for analysis to 109. Further refinement led to the exclusion of 77 articles based on the following criteria:

- Not related to the research topic: 17 articles.
- Did not present supervised or unsupervised learning techniques: 32 articles.
- Did not use data applied to telecommunications companies: 28 articles.

This left 32 final articles for inclusion in the systematic review, forming the foundation for the analysis presented in the subsequent sections.

2.3. Results and findings

After conducting a thorough search for relevant articles and narrowing the selection down to 32 final articles, the interpretation of these studies was divided into three topics: supervised learning techniques (15 articles), unsupervised learning methods (6) and the combination of supervised and unsupervised techniques (11 articles).

2.3.1. Supervised Learning

Of the 15 articles that explore supervised learning techniques, an initial analysis revealed that these articles could be divided into three distinct topics: Fraud Detection [38], Profitability Prediction [7], and Churn Prediction [39]. This categorization, based on the authors of articles [7], [38], [39], provides a structured overview of the current state of the art in supervised learning techniques applied to this domain. By organizing the studies into these three categories, it becomes possible to systematically present the methodologies, performance metrics, and key findings reported in the literature. The following sections will provide a detailed discussion of each category, highlighting the most relevant techniques and insights derived from the selected studies.

Churn Prediction

According to the authors of article [39, p. 2007], “the churn customers are the customers, who decides to escape from the service provider and prepare to shift other competitors in the market”, therefore, churn prediction involves predicting the likelihood that a customer will leave a service, whether it be telecommunications, banking, or other subscription-based services. As stated by the authors of article [40, p. 1–2], “Effective churn modeling has two important components: (i) predicting whether a specific customer will churn, (ii) discovering the reasons behind their churn, either at a local or global level. While much of the existing research predominantly focuses on the first aspect”. This highlights a significant gap in the literature, as understanding the reasons for churn is equally important for implementing targeted retention strategies.

However, all studies reviewed in this systematic literature review focus exclusively on the prediction of whether customers will churn, leaving the second component, which is understanding churn reasons, largely unexplored. By using customer data such as usage patterns, demographic information, and customer interactions, companies can predict which customers are at risk of churning, allowing businesses to proactively engage with at-risk customers to retain them. Supervised learning methods, for example, logistic regression, random forest, and XGBoost, are often employed in churn prediction, as they can model complex patterns in customer behavior [41], [42], [43].

Table A.1 in Appendix A, presents the articles classified under churn prediction, detailing the techniques employed and the metrics used to assess their performance.

It is important to address the use of article [44] due to the fact that the technique used (Rough Set Theory (RST)) is not directly supervised learning but is closely linked to it through its ability to preprocess data and generate decision rules that enhance supervised models. By identifying relevant features and eliminating redundancies, RST simplifies datasets while preserving critical information, making it an effective tool for tasks like classification and prediction [44].

The analysis of studies presented in the literature on churn prediction in the telecommunications sector highlights the widespread use of supervised learning techniques, with varying results depending on the methods employed and the metrics used for evaluation. Decision tree-based algorithms, such as RF and GBM, have shown robust performance

in various implementations. In particular, RF often achieves metrics for precision, recall, and F1-Score close to 99%, as demonstrated by the authors of studies [5], [42]. While GBM is a consistent technique, its performance was low in study [40] where it achieved an F1-Score of 60% and a precision of 67%, despite a good ROC score of 86%.

On the other hand, models based on logistic regression tend to produce moderate but still competitive results. For example, the authors of article [2] noticed that logistic regression achieved an F1-Score of 81%, while its boosted variant performed slightly worse. This demonstrates that the simplicity of logistic regression can be useful, although it often does not achieve the results of more complex methods.

Among more sophisticated methods, deep learning-based techniques, such as optimized convolutional neural networks, also stand out. The authors of article [39] reported a precision of 94.76%, indicating that these methods can be highly effective for large volumes of data. However, in another case, the Fully Connected Layer Convolutional Neural Network - Long Short Term Memory (FCLCNN-LSTM) showed mixed results, with a relatively low F1-Score of 50.93%, despite an overall precision of 94.15%, suggesting that the complexity of models does not always translate into consistent performance across all metrics.

Algorithms like XGBoost have proven particularly effective, as seen by the authors of articles [1], [41] achieving high performance metrics. These results highlight its ability to handle complex tabular data and large volumes, often outperforming simpler methods.

The studies also indicate that ensemble techniques, which combine multiple algorithms, can offer significant benefits. In article [45] the authors note that the ensemble approach achieved the highest F1-Score (85%) among the tested methods, suggesting that combining techniques may capture patterns more comprehensively.

Fraud Detection

Fraud, as defined by the authors of [38, p. 91], is “the use of one’s occupation for personal enrichment through the deliberate misuse or misapplication of the employing organization’s resources or assets.”. Detecting such activities is crucial, given their significant impact across sectors like online banking, e-commerce, and telecommunications, resulting in multi-billion dollar losses globally [46]. In telecom, fraud detection typically involves

spotting anomalies, such as unusual patterns of behavior in customer accounts, usage, or financial transactions. Machine learning techniques like Random Forest, SVM, and XGBoost are commonly used to predict and detect fraud in real-time by learning from historical fraud data [38], [46], [47].

Table A.2 in Appendix A presents the articles categorized under fraud detection, along with the techniques used and the metrics applied to evaluate their performance.

Fraud detection in telecommunications benefits significantly from machine learning techniques, with advanced methods such as graph-based approaches, deep learning, and ensemble models standing out in recent research. The Bridge to Graph (BTG) method, as highlighted by the authors in article [47], demonstrated excellent performance with an Area Under Curve (AUC) of 92.45% and an F1-Score of 87.01%, showcasing the power of graph-based models in capturing relational data for fraud detection.

Deep learning techniques, particularly Deep Convolutional Neural Network (DCNN), also showed promise, achieving the highest accuracy (82%) by the authors of the comparative study [46]. However, the results obtained in this article also show that traditional machine learning methods, such as RF and Gradient Boosting Classifiers, provided competitive performance. ANNs emerged as another effective approach, as seen by the authors of article [38], showing a superior performance compared to classical algorithms like SVM and Decision Trees.

Profitability Prediction

Profitability Prediction aims to forecast a company's future profitability or financial performance based on various factors, including market conditions, customer behavior, and historical financial data [7], [48]. This process helps businesses make informed decisions about investments, pricing strategies, and product offerings. Supervised learning models like SVM[7], Random Forest[48], and Neural Networks[7] are frequently applied in profitability prediction tasks to assess the factors that contribute most significantly to a company's financial outcomes.

Table A.3 in Appendix A provides an overview of the articles focused on profitability prediction, highlighting the techniques employed and the evaluation metrics used to assess their performance.

As referred by the authors of article [7], k-NN demonstrated the most promising results, achieving a precision of 95.5%, recall of 94.7%, and an F1-Score of 95.1%, even though other models, such as Naïve Bayes, ANN and SVM, also performed well. These results highlight that machine learning methods, even with varying complexities, can deliver accurate predictions of profitability. However, the field of profitability prediction in telecommunications remains relatively under explored compared to other applications like churn or fraud detection. The limited number of studies available makes it challenging to draw broader insights or generalize findings. Therefore, further research and the collection of more articles in this domain would be necessary to gain a more comprehensive understanding of the effectiveness and applicability of different machine learning techniques in profitability prediction.

2.3.2. Unsupervised Learning

Unlike the articles analyzed that focused on supervised learning techniques, which were based on churn prediction, fraud detection and profitability prediction, the 6 articles that focus on unsupervised learning techniques are only applied to churn prediction.

The reviewed studies demonstrate the effectiveness of various clustering algorithms in identifying meaningful customer groups, extracting key patterns, and enhancing decision-making processes. Table A.4 in Appendix A shows the key techniques and metrics used in these studies.

A predominant approach is spectral clustering, as applied by the authors of articles [49], [50]. The authors of article [49] introduced an enhanced spectral clustering algorithm with S-Distance, achieving a Jaccard Index of 92.45% and an F-score of 87.01%, demonstrating that modified distance measures can improve clustering quality, while the authors of [50] proposed a scalable spectral clustering method using Nyström and Lanczos approximations, focusing on reducing computational complexity rather than predictive accuracy, highlighting a trade-off between efficiency and precision.

Another key trend is the integration of graph theory with clustering methods, particularly in social network analysis, according to the authors of article [51]. In this study, Wards Minimum-Variance Clustering was combined with graph-based metrics, achieving a top decile lift of 2.5-10, depending on the model used. Their findings suggest that churn

behavior is influenced by social connections, where influential customers trigger additional churn within their network.

Traditional clustering techniques, such as K-means and hierarchical clustering, have also been widely used, as the authors of article [10] compared K-means, K-medoids, FCM, GK, and Hierarchical Clustering, reporting an accuracy of 55.15% for FCM, while K-means obtained the highest Rand Index (50.50%) and Fowlkes Mallows Index (61.09%). These results emphasize the importance of selecting the appropriate clustering method based on the characteristics of the dataset.

Additionally, the authors of article [32] combined hierarchical clustering with rough set theory, stating that their approach proved effective in identifying customer behavior patterns. Similarly, the authors of article [9] applied K-means clustering for customer segmentation, confirming its effectiveness in detecting high-risk churn groups.

2.3.3. Supervised and unsupervised learning combined

The integration of supervised and unsupervised learning techniques in the telecommunications sector has been widely explored for customer churn prediction. Table A.5 in Appendix A provides a detailed overview of 11 studies that integrate both learning paradigms, reinforcing the significance of churn prediction as a critical challenge in the industry.

A key observation from these studies is that hybrid approaches consistently outperform individual techniques, demonstrating their effectiveness in enhancing predictive accuracy and model robustness. This improvement is evident across multiple studies where the authors compared the results obtained by using supervised and unsupervised learning in isolation versus in combination.

A distinct approach was identified by the authors of article [52], where unsupervised learning was applied after supervised learning. In this study, XGBoost was first employed to predict customer churn, achieving an accuracy of 85.1% and detecting approximately 11,000 churn customers per month. Subsequently, K-Means clustering was used to segment these customers, leading to a 0.93% reduction in churn by successfully retaining 9,000 customers. This highlights the utility of clustering as a post-processing tool to refine churn management strategies.

Moreover, one particular study [53] has demonstrated that higher performance metrics might not always indicate methodological superiority but rather result from data preprocessing choices. For instance, the authors of article [53] employed Principal Component Analysis (PCA) to reduce dimensionality from 186 to 60 principal components before applying classification models. While this approach led to an extremely high reported accuracy (99.97%), the retention of a significant portion of the original data complexity suggests that the model still operated with substantial information, potentially inflating predictive performance. Additionally, the impact of class imbalance is often underexplored in these studies. Many real-world churn datasets exhibit highly imbalanced distributions, which can lead to misleadingly high accuracy scores if not properly addressed through resampling techniques or appropriate metric selection.

The role of clustering in these studies varies significantly. In article [54], the authors used K-Means as an exploratory tool for customer segmentation without direct integration into the classification models. In contrast, the authors of article [55] exemplify a more intertwined approach, where unsupervised learning is a critical preprocessing step that enhances the classification process. Their Multi-Objective Evolutionary Ensemble Clustering (MOEEC)-1 and MOEEC-2 models integrate clustering within ensemble learning frameworks, leading to increased classification accuracy and robustness.

Several trends emerge from these hybrid models. Regarding the choice of clustering methods, most studies employ centroid-based techniques such as K-Means, Adaptive K-Means, and K-Medoids, likely due to their simplicity and interpretability in customer segmentation. In terms of supervised learning models, the most frequently used classifiers include SVM, RF, Gradient Boosting Trees, and XGBoost. Ensemble techniques, such as AdaBoost and Stacking, also appear frequently, highlighting the advantages of leveraging multiple classifiers to enhance predictive accuracy. Finally, the effectiveness of hybrid models is evident across all studies, as the combination of clustering and classification techniques leads to improved churn prediction. For instance, the authors of [24] reported that RF outperformed SVM, achieving an accuracy of 95.44% and an F1-Score of 97% for non-churners and 84% for churners, emphasizing the advantage of hybrid approaches.

Additionally, some studies introduced novel methodologies beyond traditional clustering and classification. The authors of article [56] proposed a probabilistic possibilistic fuzzy

clustering model combined with an artificial neural network, which allows for more flexible membership assignments compared to traditional K-Means clustering. This approach improves the model's ability to capture uncertainty in customer behavior. Meanwhile, the authors of article [57] employed a Bayesian hierarchical joint model with SVM, which incorporates both time-to-event and count data to refine churn prediction. These methodologies highlight the potential for hybrid models to go beyond standard techniques, integrating probabilistic and hierarchical modeling approaches for enhanced churn prediction.

2.4. Chapter Overview

Firstly, the application of supervised learning techniques in telecommunications has proven to be diverse, covering areas such as customer churn prediction [42], fraud detection [46], and profitability forecasting [7]. Each of these domains use different models, ranging from traditional methods such as logistic regression and Naïve Bayes to more complex approaches such as deep learning and ensemble methods. These techniques have been adapted to meet specific challenges in the telecommunications industry, demonstrating high predictive accuracy and strong performance metrics in their respective applications.

Secondly, the application of unsupervised learning techniques in telecommunications has been widely explored, particularly in customer segmentation and churn prediction. Various clustering methods, such as K-means [10], hierarchical clustering [10], and spectral clustering [49], [50], have been used to identify meaningful customer groups and behavioral patterns. More advanced approaches, including graph-based clustering and rough set theory, have further enhanced the analysis of customer relationships and churn drivers. These techniques have proven effective in extracting insights from large datasets, enabling telecom operators to better understand their customers and optimize retention strategies.

Lastly, recent studies have highlighted the growing relevance of hybrid approaches that integrate both supervised and unsupervised learning. This combination has demonstrated significant improvements in predictive accuracy and robustness, particularly in churn prediction. By leveraging clustering techniques as a preprocessing step, researchers have enhanced the feature engineering process, leading to better-performing classification models [55]. Additionally, clustering has also been used as a post-processing tool, refining the segmentation of high-risk customers after churn prediction [52]. These hybrid methodologies consistently outperform standalone supervised or unsupervised techniques, reinforcing the

idea that a complementary approach can extract more meaningful insights from telecom data.

Nevertheless, it is important to emphasize that, as the authors of article [18, p. 42] state, “...every algorithm differs according to area of application and no algorithm is more powerful than the other in different scenarios”, meaning that the effectiveness of a given method depends largely on the complexity, dimensionality, and volume of the data used in each study. Therefore, it is not possible to definitively conclude that one technique is universally superior to others, being that the choice of the most suitable method should consider factors such as the data characteristics, the specific context of the study, and the objectives at hand. Although some models may demonstrate better performance than others, this performance is heavily influenced by these factors, even though all the studies analyzed are based on telecommunications company data. So, it cannot be concluded that these models are inherently superior, simply because their performance may not be as strong when applied to different datasets.

2.5. Different Approach

In studies using supervised learning techniques, it is common to propose the use of customer segmentation through machine learning and unsupervised learning techniques, as stated by the authors of studies [2], [41]. Furthermore, the authors of articles [40], [42] recommend using more balanced data to significantly improve results. In addition, the authors of article [40] suggest handling outliers carefully, as they can distort machine learning models, reducing accuracy and generalization. Removing or mitigating their impact can improve model performance by preventing extreme values from biasing predictions.

A common limitation in articles is the small amount of available customer data, as noted by the authors of studies [42], [54], [57], with more data allowing for better validation of machine learning methods. The authors of articles [7], [53], [55] recommend the use of more advanced techniques, such as deep learning, and the refinement of feature engineering, exploring additional techniques and features to enhance model performance. Additionally, the authors of [33] suggests determining the optimal number of clusters to achieve the best results.

Although the authors of [18] affirm that no algorithm is inherently more powerful than another, in different scenarios, it can be said, based on the analysis of studies that combine

supervised and unsupervised learning, that this approach consistently outperforms the use of either technique individually.

Thus, the model proposed in this dissertation addresses all the limitations and recommendations identified in the analyzed studies. In addition to combining supervised and unsupervised learning techniques, this study benefits from a large dataset (millions of observations), employs advanced feature engineering techniques, such as balancing the dataset, and applies methods to determine the optimal number of clusters in unsupervised learning. By building upon these strengths, the proposed model is expected to achieve higher predictive accuracy and better customer segmentation compared to traditional approaches. Leveraging the advantages of both supervised and unsupervised learning, it not only improves classification performance but also uncovers hidden patterns within the data.

CHAPTER 3

Data Analysis

3.1. Business Understanding

Managing customer debt is critical for financial and telecommunications companies, as unpaid debts can significantly impact cash flow and profitability. Traditional credit scoring models rely on historical payment behavior, which may not be effective for new customers or those without a detailed credit history. By implementing a predictive model that identifies potential defaulters early, the company can make proactive decisions, such as adjusting credit limits, offering alternative payment plans, or requiring additional guarantees. This approach enhances financial stability and optimizes debt recovery strategies.

Hence, in this study, the primary goal is to predict whether a customer will pay their outstanding debt by leveraging machine learning techniques. The model will be developed using financial data that includes relevant customer attributes and debt-related metrics. Additionally, unsupervised learning techniques will be applied to reduce dimensionality and explore customer segmentation. Finally, insights from both supervised and unsupervised learning approaches will be compared to provide a deeper understanding of customer debt behaviors and improve financial decision making.

The implementation of this model can help reduce financial risk by identifying high-risk customers early, allowing for preventive actions and automating the debt risk assessment process, which increases operational efficiency by reducing the need for manual evaluations. This model can also reduce financial losses associated with unpaid debts by minimizing litigation, as well as operational, administrative, and financial costs related to debt collection and cash flow disruptions.

3.2. Data Understanding

The dataset used in this study comes from a telecommunications company in Portugal and contains information on customer debt as of December 2023. It consists of 90 variables and 2,437,633 observations, with each observation representing a customers contract

account. The dataset provides a snapshot of the customer’s financial situation at that time, covering different customer types (consumer, business, public sector, wholesale). It includes details on total debt, overdue and non-overdue amounts, debt segmented by different maturities, as well as indicators of indemnities, provisions, dissolution or insolvency status, installment payments for equipment purchases, payment methods (such as bank transfer), among others.

A significant portion of the variables in the dataset are purely informational (such as tax identification number, contract account number, company name, account type, etc.). These were excluded as they do not contribute to the models. This selection process was carried out with the assistance of experienced professionals who handle this data regularly, ensuring that no relevant information was discarded. After this filtering step, the dataset was reduced to 27 variables. Table B.1 in Appendix B, section B.1, provides a brief description of each of these variables.

Considering that this data set is about financial/debt variables, some concepts may not be adequately explained in table B.1 in Appendix B, section B.1, and understanding these concepts is imperative. So, here is the explanation:

- Wholesale: In the dataset, the 'Sub-Segmento' variable classifies customers into different categories: consumer, business, public, and wholesale. The Wholesale category specifically refers to relationships between the telecom company and other telecommunications operators, including those from different countries [58]. In this context, wholesale customers are other telecom providers that have service agreements with the company.
- A billing cycle is the period of time between two billing statements issued to a customer. It defines how often a company generates invoices or bills for services provided [59]. In this case, the billing cycle occurs every three days.
- 'VEP' (Installment Sale for Equipment): Financing option that allows customers to purchase equipment and pay for it in installments rather than making a full upfront payment. This structured payment plan is agreed upon at the time of purchase, enabling customers to spread the cost over a defined period [60].
- 'PPP' (Payment Plan): Debt repayment plans can be designed as structured installment-based arrangements that smooth repayment over time, reducing

financial stress and improving repayment outcomes [61]. In the telecommunications context considered in this study, such arrangements typically arise when a customer fails to meet their original payment obligations, leading to the restructuring of the outstanding balance into an installment-based repayment plan.

3.3. Data Preparation

This section describes and explains all the preprocessing steps applied to the dataset before proceeding to the modeling phase.

Before starting the data preprocessing, the columns 'Vencido' and '0 a 30' were excluded. These columns were later used to create the target variable, and their inclusion in the dataset could lead to potential correlation issues. Since the target variable creation was already planned from the beginning, preprocessing these columns seemed unnecessary, as they would eventually be removed(it would only be worthwhile to handle missing values, but these variables had none). As a result, the dataset was reduced to 25 variables.

3.3.1. Missing values and New columns

Firstly, it is essential to analyze the missing values in the dataset, as most models do not handle null values well. Figure 3.1 illustrates the percentage of missing values for each variable.

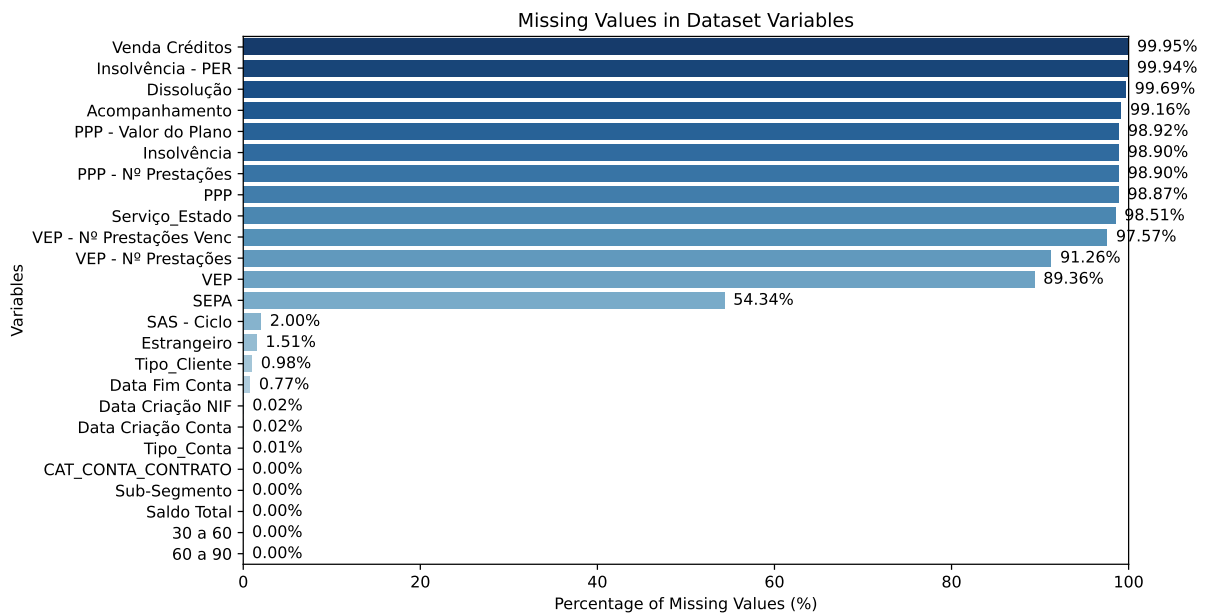


FIGURE 3.1. Percentage of missing values for each variable in the dataset.

As observed, many variables have a high percentage of missing values. After a thorough analysis, it was found that, in most cases, missing values represent the opposite situation of the available values. For instance, in the 'Dissolução' column, all missing values correspond to cases where dissolution did not occur, therefore, these missing values were replaced with 'N' (No), which means no occurrence. The same approach was applied to several other columns. For the remaining columns with missing values, a different approach was required to handle missing data appropriately.

Taking into consideration these notes, the following steps were taken:

- The variable 'Estrangeiro' was filled based on the 'Tipo_Cliente' column, assuming that national clients ('Nac') are not foreign ('N') and international clients ('Int') are foreign ('S' - Yes). After this step, the variable 'Tipo_Cliente' was excluded because it provides the same information as the 'Estrangeiro' variable. A similar approach was applied to the 'Serviço_Estado' variable, which was filled based on the 'Tipo_Conta' column: 'CN' mapped to 'Ativo', 'CT' to 'Desativos', and 'CA' to 'Interrupção', and after this step the variable 'Tipo_Conta' was also excluded for the same reason as 'Tipo_Cliente'
- The variables 'Dissolução', 'Insolvência', and 'Insolvência - PER' originally contained dates indicating when those events occurred. These were transformed into binary variables, where the presence of a date (i.e., event occurred) was encoded as 'S', and missing values (i.e., no such event) were encoded as 'N'. For consistency, the binary variables 'Acompanhamento', 'Venda Créditos', and 'SEPA' were treated similarly: missing values were replaced with 'N', and non-missing values were encoded as 'S'.
- For other date columns such as 'Data Criação NIF' and 'Data Criação Conta', missing values were imputed with the median of the respective variable. Additionally, the 'SAS - Ciclo' variable had missing values filled with the mode of the column to ensure consistency in the data.
- For the 'VEP' and 'PPP' variables, missing values were replaced with 'N', while the columns related to the number of installments ('VEP - N° Prestações', 'VEP - N° Prestações Venc', 'PPP - N° Prestações', and 'PPP - Valor do Plano') had missing values replaced with 0.

- For accounts with a null value in 'Data Fim Conta', a value of 0 was assigned, indicating that the account is still active.

Regarding the creation of new variables, 'Data Criação NIF (Diferença Dias)' was created, representing the difference in days between the NIF creation date and the most recent date in the dataset. After this calculation, the original 'Data Criação NIF' column was removed. Additionally, the variable 'Longevidade' was created based on 'Data Criação Conta' and 'Data Fim Conta', representing the number of days between account creation and closure. The 'Data Criação Conta' and 'Data Fim Conta' columns were removed as they were no longer needed after the 'Longevidade' variable was created.

In summary, no variables were deleted (except those removed as part of the creation of a new variable based on the original), and no observations were removed, as all missing values were successfully treated.

3.3.2. Outliers

To treat outliers, it was decided that winsorization was the most suitable method. The winsorization technique is a statistical method used to reduce the impact of outliers in a dataset by replacing extreme values with less extreme values within a defined range. Unlike outlier removal, which simply excludes these observations, winsorization adjusts values so they remain within a specified limit. This approach is useful for preventing outliers from distorting statistical analyses while preserving all observations in the dataset [62]. This method was chosen because it reduces the influence of outliers without removing any data, ensuring that valuable information is not lost. The following columns underwent winsorization:

Regarding 'Saldo Total', '30 a 60' and '60 a 90', these columns were winsorized by capping the lower 0.5% and upper 5% of values. The lower limit was set to only 0.5% to maintain the representativity of negative values. Since negative values are already less frequent compared to positive values, applying a 5% lower limit could have disproportionately affected them. On the other hand, a 5% upper limit was chosen to handle the concentration of extreme positive values more effectively. However, during the outlier treatment, it was observed that the '60 to 90' variable ended up with no values, as the majority (96.75%) of its values were 0. Nevertheless, to assess its impact, model performance was compared under two scenarios: with outlier treatment (effectively

removing the variable from the dataset) and without it. Tables B.3 to B.16 in Appendix B, section B.2 present the confusion matrices and classification reports of different models when the '60 a 90' variable was included, whereas tables C.3 to C.16 in Appendix B, section B.2 display the results when the variable was excluded. The analysis showed that including '60 a 90' did not lead to significant improvements in any of the models. Since its inclusion would only increase the datasets dimensionality without providing meaningful predictive value, it was decided to exclude this variable from the analysis, or, in other words, proceed with the outlier treatment of this variable.

As for 'VEP - N° Prestações', 'VEP - N° Prestações Venc', 'PPP - N° Prestações', 'PPP - Valor do Plano', and 'Data Criação NIF (Diferença Dias)', winsorization was applied by capping only the top 0.5% of values, ensuring that extreme positive values were adjusted while leaving the lower values unchanged.

In the case of 'Longevidade', winsorization was applied symmetrically at the top and bottom 5% to handle extreme values, including negative ones. Additional analysis was conducted to identify the values falling within these percentiles since some negative values needed to be replaced with zero. This adjustment was necessary because negative values in 'Longevidade' do not make sense, meaning they are likely errors and should be corrected to zero.

The remaining columns in the dataset were not winsorized because the outliers in these variables were considered representative of real patterns and should not be removed, which means that altering these values could distort meaningful trends in the data rather than improve the analysis.

3.3.3. Normalization

Normalization is a crucial step in data preprocessing for machine learning, as it helps improve model performance and ensures that features with different scales or units are treated equitably [63]. In this case, the Min-Max normalization technique was applied. Min-Max normalization is particularly useful when the data does not follow a normal distribution and when it is important to preserve the original shape of the data distribution while maintaining relationships between values. Additionally, this technique is intuitive and easy to implement, making it a popular choice among data science and statistics professionals [63].

Min-Max normalization transforms the data into a specific range, typically between 0 and 1, using the following formula:

$$X' = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \quad (1)$$

where X is the original value, $X_{\min}$ is the minimum value of the feature, and $X_{\max}$ is the maximum value. The transformed value, X' , will be in the range $[0, 1]$. This transformation is applied to each feature individually, ensuring comparability regardless of their original scales [63].

Min-Max normalization was applied to the following columns in the dataset: 'Saldo Total', '30 a 60', 'VEP - N° Prestações', 'VEP - N° Prestações Venc', 'PPP - N° Prestações', 'PPP - Valor do Plano', 'Data Criação NIF (Diferença Dias)', and 'Longevidade'.

3.3.4. Transformation of categorical variables to numerical

In this step, categorical variables were transformed into numerical. This step was crucial to ensure that the models could process the data correctly, as most machine learning algorithms do not work directly with categorical data [64]. The encoding process consisted of two main approaches: binary encoding and one-hot encoding. Regarding binary encoding, some categorical variables contained only two unique values, specifically 'S' (Yes) and 'N' (No). To make them usable in the models, these variables were converted into a binary format, where 'S' was mapped to 1 and 'N' was mapped to 0. The transformed variables were: 'Estrangeiro', 'Dissolução', 'Insolvência', 'Insolvência - PER', 'VEP', 'PPP', 'SEPA', 'Acompanhamento', and 'Venda Créditos'.

For categorical variables with multiple unique values, the one-hot encoding technique was applied. This approach creates a new binary column for each unique category within a variable, allowing models to distinguish between categories without imposing an arbitrary order [64]. The variables subjected to this transformation were: 'CAT_CONTA_CONTRATO', 'Sub-Segmento', 'Serviço_Estado', and 'SAS - Ciclo'. To avoid multicollinearity and unnecessary data expansion, the first category from each variable was removed and used as a reference.

3.3.5. Target variable

The target variable, 'Cumprimento da divida', is a binary variable that can take two values: 'Pagou' (1), which means the client is a good payer, or 'Não Pagou' (0) which means the client is a bad payer. The definition of this variable is based on the 'Vencido' and '0 a 30' column.

To understand the logic behind the target variable, it is necessary to analyze the 'Vencido' variable. This variable contains numerical values that can be positive (meaning that the individual has allowed the payment to exceed the due date, meaning they have an outstanding debt), zero (indicating that the individual makes all payments on time, as there is no overdue amount), or negative (referring to credit notes that have been issued)

However, 'Cumprimento da divida' does not depend solely on 'Vencido', as it also takes into account the '0 a 30' column. This is because some customers pay their bills only after the monthly accounting period closes. For example, suppose the financial statements for March closes on the 30th, but a customer always pays their due amount for March on the 2nd of the following month (April). In this case, the customer would have a positive overdue balance, which would also appear in the '0 a 30' column. If the customers were classified based only on the 'Vencido' variable, this customer would be incorrectly classified as a bad payer.

To address this, the target variable 'Cumprimento da divida' is defined as follows:

$$Cumprimento da divida = \begin{cases} 0, & \text{if } ('Vencido' - '0 a 30') > 0 \\ 1, & \text{otherwise} \end{cases}$$

3.3.6. Class Balancing

According to the authors of [65], dealing with imbalanced datasets is crucial to ensure that predictive models do not become biased toward the majority class. When one class significantly outweighs the other, the model may struggle to learn meaningful patterns from the minority class, leading to poor generalization and misleading performance metrics. To address this issue, resampling techniques such as undersampling and oversampling are commonly used. Undersampling reduces the number of observations in the majority class to match the minority class, preventing the model from being overly influenced by the dominant class. On the other hand, oversampling increases the number of observations

in the minority class, either by duplicating existing samples or generating synthetic ones. Both approaches aim to create a more balanced dataset, improving the models ability to correctly classify both classes and enhancing overall predictive performance.

In order to understand if these techniques need to be applied, the class distribution of the target variable 'Cumprimento da divida' was analyzed. Figure 3.2 shows the distributions of the two classes of the target variable.

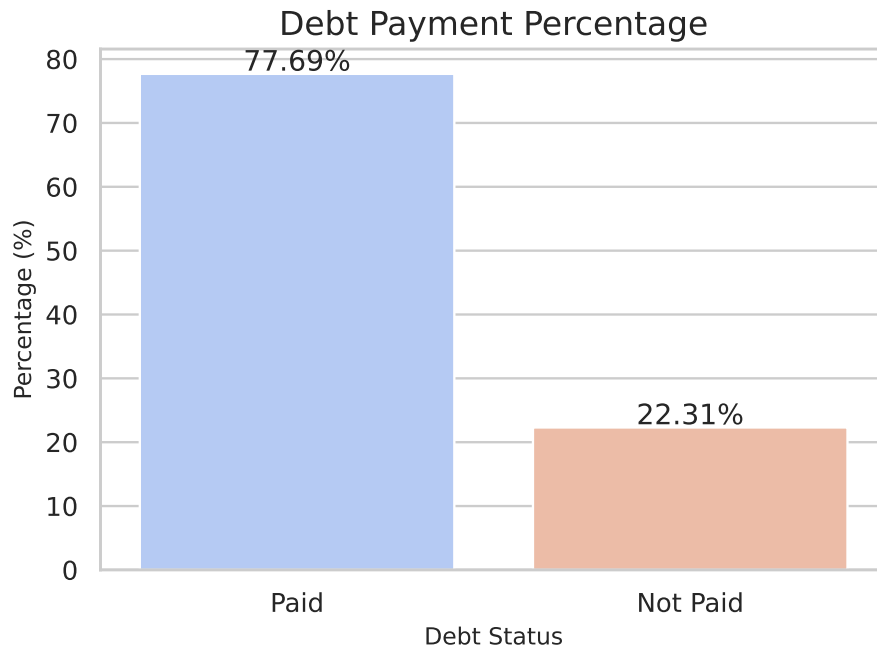


FIGURE 3.2. Class distribution of the target variable

Following this analysis, a class imbalance was observed, with the majority of clients classified as 'Paid'. To address this imbalance, undersampling was applied to the majority class 'Paid', by reducing the number of observations of 'Paid' to match the number of observations of 'Not paid', going from 2,437,633 to 1,087,666 observations, with the classes now balanced perfectly (543,833 observations in each class).

3.4. Final dataset overview

After all the data transformations and cleaning steps, the final dataset was ready for modeling. The resulting dataset contained 1,087,666 observations and 61 variables, including financial variables, payment history, contractual information, and categorical features converted through one-hot encoding, which explains the high dimensionality of the final dataset.

To facilitate understanding of the datasets structure, table B.2 in Appendix B, section B.1 presents an overview of the final variables, along with their data types and brief descriptions. The variables capture multiple facets of customer behavior and financial status. For instance, features like 'Saldo Total', '30 a 60', and 'PPP - Valor do Plano' quantify different forms of debt and installment plans. Others, such as 'Insolvência', 'Dissolução', and 'Venda Créditos', reflect credit risk and legal status. Behavioral elements are also represented, as 'SEPA' identifies customers who pay by direct debit, 'Acompanhamento' indicates monitoring status, and 'Longevidade' denotes the duration of the customer relationship. In addition, categorical information related to contract type, company segment, and service state was preserved through one-hot encoding, resulting in multiple binary variables.

CHAPTER 4

Modeling and Results

In this chapter, an initial experiment was conducted using all available features, where several models were compared across both supervised and unsupervised learning techniques. The best-performing models from each technique were then selected. Subsequently, a second experiment was carried out using the most relevant features identified from both approaches - through feature importance analysis in supervised learning and Kaiser-Meyer-Olkin (KMO) scores in unsupervised learning - with the aim of understanding whether the model could be reduced in dimensionality.

4.1. Supervised learning models

To develop predictive models for identifying customers likely to comply with debt repayment, a series of supervised learning techniques was explored through a structured evaluation process. The analysis began with a correlation study aimed at reducing multicollinearity and enhancing model performance, followed by a balanced division of the dataset into training and testing sets. Subsequently, four classification algorithms including Logistic Regression, Random Forest, k Nearest Neighbors and Neural Network were trained and assessed using key performance metrics. Feature importance was also analyzed to identify the most influential variables driving the model predictions.

4.1.1. Dataset preparation

Correlation analysis plays a crucial role in supervised learning, as it helps identify relationships between predictor variables and the target variable. When two or more variables are highly correlated, multicollinearity can occur, meaning that these variables convey similar information. This can cause issues in model estimation, as it makes it difficult to determine the individual effect of each variable. Addressing multicollinearity is crucial for obtaining stable and reliable model coefficients, ensuring more accurate predictions and reducing the likelihood of overfitting [66].

In this study, a correlation analysis was conducted on the variables used to predict debt compliance. First, the correlation matrix was computed for all variables, as can be seen in figure 4.1, where strong correlations are represented in warmer colors (orange and red), while weaker correlations are shown in cooler colors (light blue and strong blue).

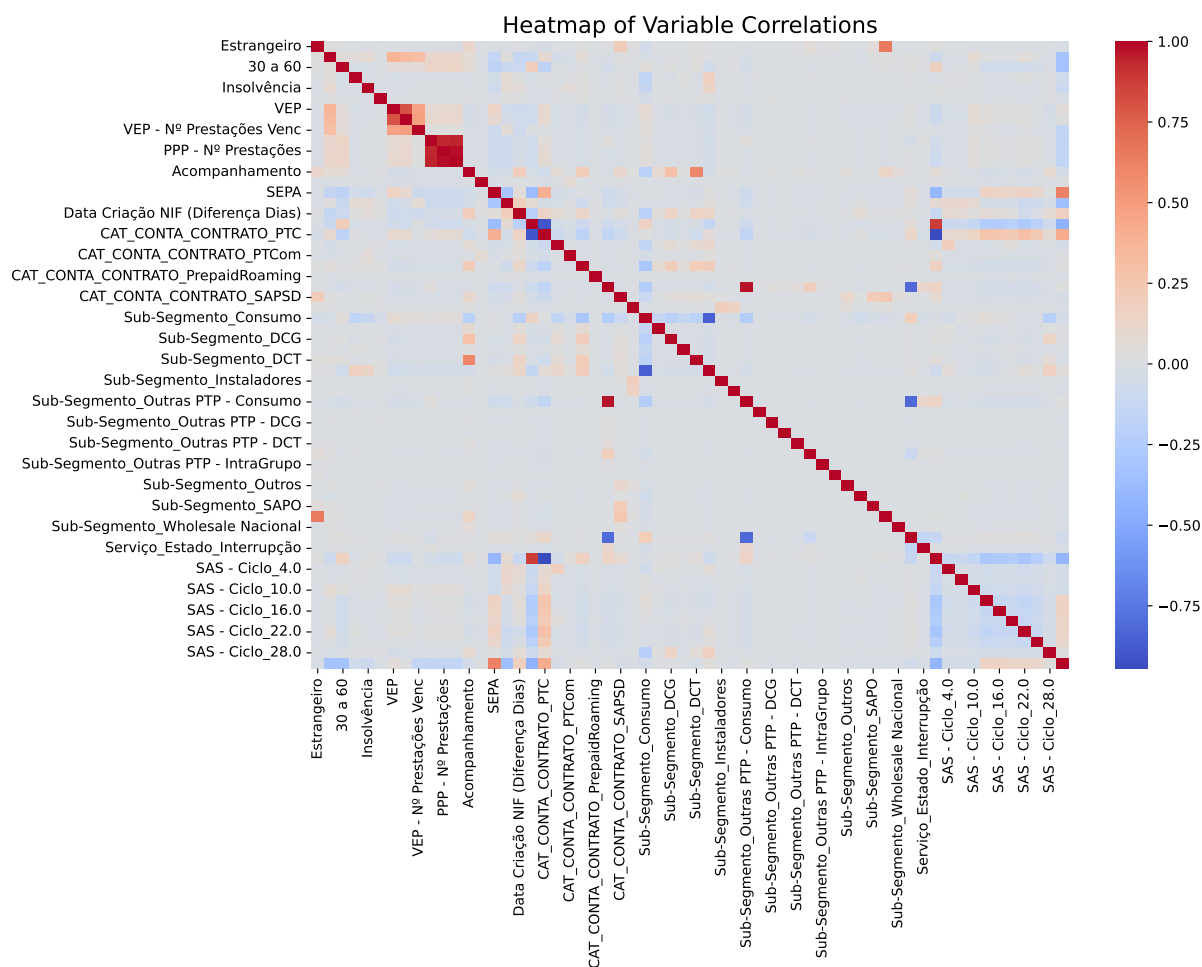


FIGURE 4.1. Heatmap of correlations between variables

Figure 4.1 may not be fully comprehensible due to the high dimensionality of the dataset. However, the goal is not to analyze the correlations individually but to identify variables with very high (positive or negative) correlations. As shown, several pairs of strongly correlated variables can be identified that need to be addressed, as highly correlated variables can lead to multicollinearity issues.

Thus, it was decided to address variables with correlations above 0.7 or below -0.7. The pairs of variables falling into this category, along with their correlation values, are presented in Table C.1 in Appendix C, section C.1. For each correlated pair, the variable

with the lower correlation to the target variable was removed, preserving the one with greater relevance for prediction.

Besides the correlation between variables, analyzing the correlation between the features and the target variable is also essential. This is because strong correlations between the features and the target suggest that the variables contain valuable predictive information, while weak correlations might indicate that certain features do not significantly contribute to the model's ability to predict the target, potentially leading to reduced model performance. Figure C.1 in Appendix C, section C.1, shows the correlations of the target variable with all the variables, and table C.2 in Appendix C, section C.1, provides a detailed breakdown with the exact correlation values. Several strong correlations between features and the target variable can be seen, such as 'SEPA' (62.67%) and 'Serviço_Estado_Desativos' (-69.26%), which is a positive sign because it indicates that these variables contain valuable predictive information, giving the model a solid foundation to produce reliable and accurate predictions.

To ensure a fair evaluation of the model, the dataset was split into training (80%) and test (20%) sets. This division allows the model to learn patterns from the training data while keeping an unseen portion for testing its generalization ability. A stratified sampling approach was used to maintain the original class distribution in both sets, which is important in order to ensure that the model learns from both classes equally.

After analyzing the correlations both between variables and between the variables and the target, and dividing the dataset into training and test sets, the dataset is now fully prepared for modeling.

4.1.2. Logistic regression

Firstly, a logistic regression model was trained to predict debt compliance. The model was trained using the training data, and its performance was evaluated on both the train and test sets. To analyze the results, confusion matrices were generated for each dataset, providing insights into the models classification performance. Additionally, classification reports were produced, detailing precision, recall, and F1-score for each class, helping to assess the models predictive capability. The corresponding confusion matrices and classification reports are presented in Appendix C, section C.2.

The confusion matrix for the train set is presented in Table C.3, and the classification report is shown in Table C.4. The performance on the test set is summarized in Table C.5 and Table C.6, which present the confusion matrix and classification report, respectively.

The logistic regression demonstrated a strong performance on both the training and test sets. The overall accuracy of the model is 94%, with both classes achieving high F1-scores, despite some trade-offs between precision and recall. Examining the confusion matrix provides insights into the classification errors made by the model, with the main focus being the false positives and false negatives, as they directly impact business decisions:

- False Positives (9,536 in the test set)

These are customers whom the model incorrectly predicted as likely to repay their debt (1), but in reality, they did not pay (0). The impact of false positives is significant, as the company may approve contracts for customers who ultimately default, leading to potential financial losses. Therefore, minimizing false positives is crucial to reduce financial risk and improve the allocation of credit resources.

- False Negatives (3,148 in the test set)

These are customers who were incorrectly classified as non-payers, despite actually repaying their debt. While this results in the company rejecting some customers who would have paid, the financial impact may be less severe than granting contracts to non-paying customers. However, reducing false negatives is also important to avoid unnecessary loss of potential revenue and maintain good customer relationships.

The recall for class 1 (payers) is 97%, meaning that only 3% of actual payers are incorrectly classified as non-payers. On the other hand, the precision for class 1 is 92%, indicating that 8% of customers predicted as payers actually default. These trade-offs must be considered depending on the company's risk tolerance and financial objectives.

A key aspect of model evaluation is determining whether the model has overfitted the training data or if it generalizes well to unseen observations. The results indicate that there is no significant overfit, as the performance metrics are nearly identical between the training and test datasets, suggesting that the model learns the underlying patterns in the data rather than memorizing the training set.

4.1.3. Random forest

Secondly, a random forest model was trained, and several hyperparameters were defined to optimize performance while maintaining computational efficiency:

- `n_estimators = 100`: The model was built with 100 decision trees, balancing predictive power and training time.
- `max_depth = 20`: The maximum depth of the trees was limited to prevent excessive complexity and overfitting.
- `min_samples_split = 10`: A node must have at least 10 samples to be split, ensuring that splits are meaningful and reducing overfitting.
- `min_samples_leaf = 5`: Each leaf node must contain at least 5 samples, helping generalization.

After training, the models performance was evaluated in both the training and test sets, using the same methods as in logistic regression. The corresponding confusion matrices and classification reports are presented in Appendix C, section C.2. The confusion matrix for the train set is presented in Table C.7, while the classification report is shown in Table C.8. Similarly, the performance on the test set is summarized in Table C.9, which presents the confusion matrix, and in Table C.10, which provides the classification report.

The random forest model achieved excellent performance, with an overall accuracy of 99% on both the training and test sets, and precision, recall, and F1-Score consistently high across both classes, indicating a well-balanced model. Regarding classification errors, there were 2,165 false positives and 1,079 false negatives in the test set, values that are significantly low compared to the number of true positives and true negatives.

Compared to the logistic regression model, the Random Forest significantly reduces both false positives and false negatives, improving predictive reliability. Additionally, there is no indication of overfitting, as the train and test performance remain nearly identical.

4.1.4. k-NN

Thirdly, a k-NN model was trained to predict debt compliance. This algorithm classifies observations based on the majority class of the k nearest neighbors in the feature space. In this case, the model was configured with $k = 5$, meaning that each observation was classified based on the five closest neighbors.

After training, the models performance was evaluated on both the training and test sets, using the same metrics as in logistic regression and random forest. The corresponding confusion matrices and classification reports are presented in Appendix C, section C.2. The confusion matrix for the train set is presented in Table C.11 presents the confusion matrix and Table C.12 shows the classification report for the k-NN model. Equivalently, the confusion matrix on the test set is presented in Table C.13 and the classification report is in Table C.14.

The k-NN model achieved strong performance, with an overall accuracy of 97% on the training set and 96% on the test set. Precision, recall, and F1-Score remain well-balanced across both classes, indicating a reliable classification model.

Regarding classification errors, the test set contains 3,880 false positives and 3,891 false negatives, which are slightly higher than those observed in the Random Forest model but still relatively low compared to true positives and true negatives.

While the k-NN model performs slightly worse than Random Forest, it still maintains a solid predictive capability with no signs of overfitting, as the train and test performance remain closely aligned.

4.1.5. Neural Network

Lastly, a neural network model was trained. The architecture and training process were defined as follows.

- Architecture:
 - Input layer with 64 neurons (ReLU activation).
 - One hidden layer with 32 neurons (ReLU activation).
 - Dropout regularization (20% rate) applied after each layer.
 - Output layer with 1 neuron and a sigmoid activation function for binary classification.
- Training:
 - Optimizer: Adam.
 - Loss function: Binary Cross-Entropy.
 - Early stopping: training stopped if validation loss did not improve for five consecutive epochs.

The models performance was assessed on the test set through a confusion matrix and a classification report. Unlike previous models, there was no need to analyze training metrics, as they were only used earlier to check for overfitting. Instead, accuracy and loss graphs were generated to visualize the learning process. Tables C.15 and C.16 in Appendix C, section C.2 present the confusion matrix and the classification report, respectively, while Figure 4.2 shows the accuracy and loss curves and training history of the neural network model.

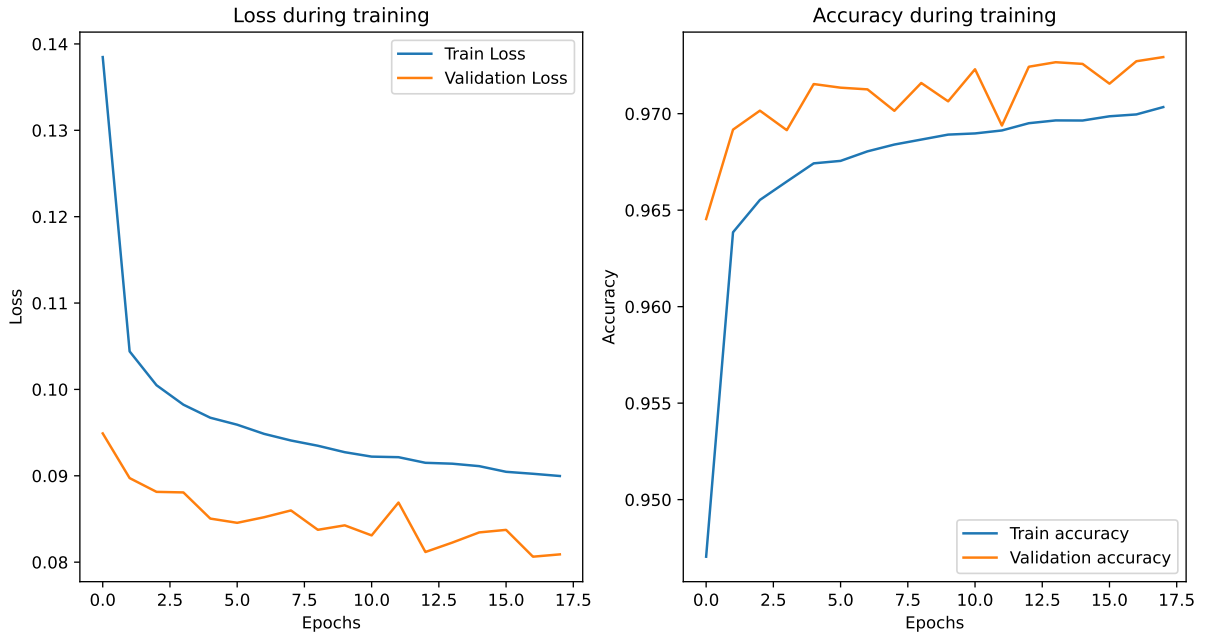


FIGURE 4.2. Neural network accuracy (right) and loss (left) graphics

The neural network model achieved an overall accuracy of 97% on the test set, with precision, recall, and F1-Score maintaining a balanced performance across both classes. Regarding classification errors, the test set contains 3,063 false positives and 2,766 false negatives, representing a slight improvement over the k-NN model. These values remain relatively low compared to the total number of true positives and true negatives, reinforcing the model's reliability. Compared to the other models, the neural network demonstrates a strong predictive ability with minimal error rates.

As it can be seen in figure 4.2, the loss curve (left) shows a steady decrease for both training and validation sets, confirming that the model successfully optimizes its objective function without signs of divergence. The accuracy curve (right) indicates rapid improvement in the early epochs, with both training and validation accuracy stabilizing around 97%, further demonstrating the models ability to generalize well to new data.

The near-identical performance on training and validation sets shows that there is no overfitting which would typically be indicated by a significant divergence between the loss and accuracy curves of the two sets, for example, if the training accuracy continued to improve while the validation accuracy plateaued or declined.

4.1.6. Best supervised learning model

To compare the performance of the four supervised learning models, Table 4.1 summarizes the main classification metrics on the test set. Class 1 is considered the positive class, therefore, recall corresponds to sensitivity for class 1 and to specificity for class 0, while precision corresponds to the Negative Predictive Value (NPV) for class 0.

TABLE 4.1. Summary of classification metrics for the test set across models (class 1 as positive)

Model	Class	Precision / NPV	Sensitivity / Specificity	F1-Score	Accuracy
Logistic Regression	0	0.97 (NPV)	0.91 (Specificity)	0.94	0.94
	1	0.92 (Precision)	0.97 (Sensitivity)	0.94	
Random Forest	0	0.99 (NPV)	0.98 (Specificity)	0.99	0.99
	1	0.98 (Precision)	0.99 (Sensitivity)	0.99	
k-NN	0	0.96 (NPV)	0.96 (Specificity)	0.96	0.96
	1	0.96 (Precision)	0.96 (Sensitivity)	0.96	
Neural Network	0	0.97 (NPV)	0.97 (Specificity)	0.97	0.97
	1	0.97 (Precision)	0.97 (Sensitivity)	0.97	

As shown in Table 4.1, the random forest model clearly stands out as the best performer, achieving the highest scores in all evaluation metrics (99%). Although k-NN and the neural network showed similar results in performance metrics, random forest provides equally high performance with significantly faster training times when compared to k-NN. Additionally, the neural network, while performing well, it is not only slightly slower when it comes to training time, but also computationally expensive and can be challenging to optimize, especially in environments with limited resources. This makes Random Forest the most efficient choice in terms of cost-effectiveness, as it offers robust results without the high computational costs associated with neural networks.

4.1.7. Feature importance analysis

In this section, the top 15 most important features for each of the supervised learning models used in the study were analyzed. Feature importance is crucial for understanding which variables most influence the model's predictions. By comparing them, key drivers of performance can be identified and determine whether a subset consistently ranks highly.

By analyzing the feature importance of the Logistic Regression (Figure C.2), Random Forest (Figure C.3), k-NN (Figure C.4) and Neural Network (Figure C.5) models, all presented in Appendix C, section C.2, it can be observed that some variables are common across all four models. The features that appear simultaneously in all models, as well as those that appear in 3 out of the 4 models, are displayed in table 4.2.

TABLE 4.2. Variables that appear in 4 and 3 models

Variables that appear in 4 models	Variables that appear in 3 models
'Serviço_Estado_Desativos'	'Saldo Total'
'30 a 60'	'SAS - Ciclo_19.0'
'PPP'	'SAS - Ciclo_22.0'
'SEPA'	'VEP - N° Prestações Venc'
'Longevidade'	'Data Criação NIF (Diferença Dias)'
'CAT_CONTA_CONTRATO_MEO'	

4.2. Unsupervised learning models

Seeing that unsupervised models are based on pattern recognition without access to the target variable, before beginning the correlation analysis and modeling, the target variable was removed from the dataset to ensure that the patterns identified are derived solely from the features and not influenced by any prior knowledge of the outcome.

In this section, PCA and t-Distributed Stochastic Neighbor Embedding (t-SNE) were analyzed to determine which technique is most suitable for dimensionality reduction. Following this, partitioning (k-Means), hierarchical (Ward and complete linkage methods) and DBSCAN clustering approaches were applied to group the data and compared to identify the most effective method for clustering the data.

4.2.1. PCA Suitability Analysis

It is crucial to examine whether there are relationships between the variables that can be represented by the PCs. This analysis follows a similar approach to the correlation analysis performed for supervised learning models, where highly correlated variables were identified. However, in this case, all variables were retained, as the purpose of the correlation analysis within the PCA analysis is to determine whether meaningful correlations exist. If the correlations were too weak, PCA would not add significant value. The correlation matrix used for this analysis is presented in Figure 4.1 in Appendix C, section C.2. As concluded in that chapter, several pairs of strongly correlated variables can be identified, which means that the application of PCA is appropriate in this context.

To assess whether PCA was suitable for the dataset, the Bartlett's test was applied to verify if the correlation matrix significantly differs from an identity matrix, confirming the suitability of the data for factor analysis when the result is significant ($p < 0.05$) [67]. The result showed a test statistic of 154,107,230.40 and a p-value close to zero, which led to the conclusion that the variables were sufficiently correlated to proceed with PCA.

Furthermore, the KMO index was calculated. According to [68], a KMO score above 0.6 is recommended, but this dataset's score was only 0.1092, which is well below the threshold. This result indicated weak correlations within the dataset, suggesting that PCA might not be effective. Therefore, the individual KMO values for each variable were calculated, revealing that several variables had very low scores, further indicating little to no correlation with other variables.

Based on these results, the decision was to remove variables with an individual KMO value of less than 0.5. After removing these variables, the KMO was recalculated, and the new overall KMO score increased to 0.7063, which is now above the acceptable threshold for PCA, which indicated that the dataset was now suitable for dimensionality reduction. As a result, the final set of variables retained for PCA are: 'Estrangeiro', 'Saldo Total', '30 a 60', 'Dissolução', 'Insolvência', 'Insolvência - PER', 'VEP', 'VEP - N° Prestações', 'VEP - N° Prestações Venc', 'PPP', 'PPP - N° Prestações', 'PPP - Valor do Plano', 'Acompanhamento', 'Venda Créditos', 'SEPA', 'Longevidade', 'Data Criação NIF (Diferença Dias)', 'Serviço_Estado_Desativos', and 'Serviço_Estado_Interrupção'.

To confirm these improvements, Bartlett's test was re-run on the cleaned dataset. The result showed a test statistic of 23,393,020.94 and a p-value of zero, which further supported the conclusion that the dataset was now appropriate for PCA.

4.2.2. Choice of number of PC

In order to choose the optimal number of principal components, the eigenvalues, scree plot, and cumulative variance were analyzed.

Eigenvalues

The eigenvalues for the 19 PC are shown in Table C.17 in Appendix C, section C.3. As the authors of [68] state, the rule of thumb consists of retaining components with eigenvalues greater than 1. However, none of the eigenvalues in this case exceed 1, indicating that none of the components explain a significant proportion of the variance in the dataset.

Scree plot

Analyzing the scree plot is also important to determine the optimal number of PC, as stated by the authors of [68]. The scree plot is represented in Figure 4.3, in Appendix C, section C.3, which visually represents the eigenvalues associated with each PC.

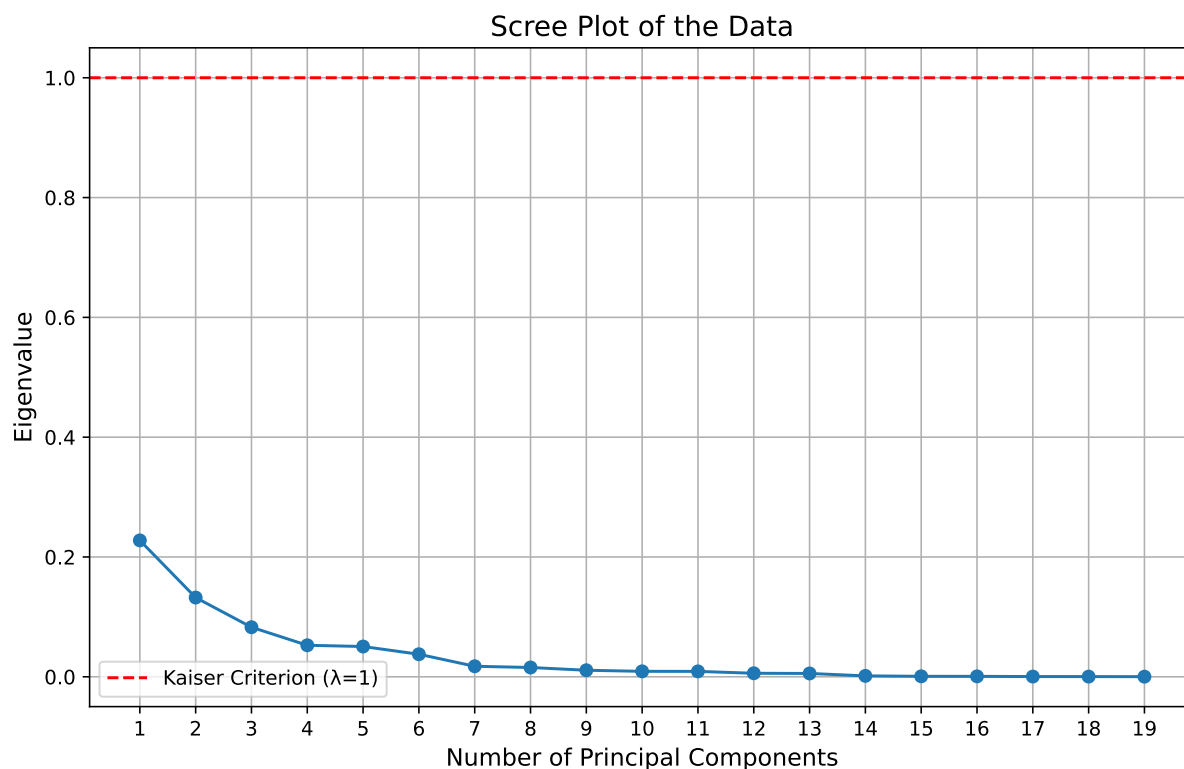


FIGURE 4.3. Scree plot showing the eigenvalues of each PC

In this case, the plot shows an inflection point, often referred to as the elbow, at the fourth principal component, suggesting that three principal components should be chosen for further analysis.

Cumulative variance

According to the authors of [68], the cumulative explained variance is important to analyze in order to understand how each principal component contributes to explaining the data variability. Table C.18 in Appendix C, section C.3 shows the cumulative variance explained by each principal component, illustrating a progressive increase in the total variance accounted for as more components are included. Notably, by the third principal component, 66.97% of the total variance is explained, which surpasses the commonly accepted threshold of 60% used in PCA to retain the most informative features while reducing dimensionality. This indicates that selecting the first three components is sufficient to capture a substantial amount of the data variability, balancing model simplicity and information preservation.

Optimal number of PC

Considering the results from the eigenvalues, scree plot, and cumulative explained variance, the optimal number of principal components to retain is three. The eigenvalues analysis does not provide a strong indication for retaining any components based on the usual threshold of 1. However, the scree plot shows a clear elbow at the fourth component, suggesting that three components capture most of the variance before the curve flattens. Additionally, the cumulative explained variance reaches 68.88% by the third component, surpassing the common threshold of 60%. Therefore, for further analysis, the first three principal components will be used.

4.2.3. Interpreting of the Principal Components

The three PCs selected capture 39.17%, 16.80%, and 12.91% of the total variance, respectively. Each component was analyzed based on its loadings, which are presented in Table C.19 in Appendix C, section C.3, in order to interpret its underlying meaning and assign an appropriate name. So, based on these loadings, the following names were chosen to each PC:

- **PC1: Engagement vs. Churn Risk**

PC1 was named 'Engagement vs. Churn Risk' due to its strong association with service status and payment behavior. The variable 'Serviço_Estado_Desativos' has the highest positive loading (0.71), indicating that customers with high PC1 scores are more likely to have had their service deactivated. Conversely, 'SEPA' has a strong negative loading (-0.63), suggesting that customers who actively use direct debit tend to have lower PC1 scores. Additionally, 'Longevidade' and 'Saldo Total' load positively, indicating that deactivated customers often have longer tenures and outstanding balances.

This component distinguishes engaged, reliable customers (low PC1 scores - active service, SEPA users) from at-risk or churned customers (high PC1 scores - inactive, no SEPA usage). This interpretation aligns with operational expectations, where engaged customers maintain active services and use automated payment mechanisms, while deactivated customers may have accumulated debt over time.

- **PC2: Active Indebtedness**

PC2 was named 'Active Indebtedness' as it is primarily influenced by variables that reflect current financial exposure. The variable 'VEP' shows a high positive loading (0.55), along with 'Saldo Total' (0.48) and 'SEPA' (0.50). These indicate that customers with high PC2 scores have considerable outstanding debt and are likely in the process of repayment, even if they use SEPA to manage those obligations. The loading for 'Serviço_Estado_Desativos' (0.37) suggests that some of these customers may already be deactivated, indicating unresolved debts.

This component captures ongoing financial pressure: customers with high PC2 scores carry substantial debt and are often involved in payment arrangements. In contrast, customers with low PC2 scores show minimal or no active debt and are likely in a healthier financial position.

- **PC3: Payment Cycle Completion**

PC3 was named 'Payment Cycle Completion' because it differentiates between customers who have likely concluded their financial obligations and those still engaged with repayment. 'SEPA' and 'Serviço_Estado_Desativos' both load positively (0.55 and 0.43 respectively), while 'VEP', 'Saldo Total', and all 'PPP' variables load negatively (for example, $VEP = -0.49$; $PPP = -0.17$ to -0.20).

This indicates that high PC3 scores are typical of customers who used direct debit, cleared their debts, and subsequently deactivated their service, suggesting a completed financial cycle. On the other hand, low PC3 scores represent customers who are still repaying debts or have outstanding obligations, often without the structure of automated payments.

After the PCA transformation, a new dataset was created containing only the three PCs as features, replacing the original variables. This step paves the way for the next phase, which involves the clustering of observations.

4.2.4. t-SNE method

In addition to the PCA technique, the t-SNE technique was also tested as a dimensionality reduction method before applying clustering algorithms. Unlike PCA, t-SNE does not require the analysis of eigenvalues or a scree plot to choose the number of components because it focuses on preserving local relationships rather than global variance. Instead of choosing the number of components manually, t-SNE automatically determines the optimal number of dimensions to map the data, by converting similarities between data points into probabilities. Similar points have a high probability of being close to each other in the lower-dimensional space, allowing t-SNE to capture the local structure of the data without needing prior decisions about the number of components [69].

Initially, t-SNE was applied to a subset of 100,000 observations, resulting in a scatter plot represented in Figure C.6 in Appendix C, section C.3. It can be noticed that the observations are highly scattered and unstructured, providing no meaningful insights for clustering.

As an alternative, a smaller subset of 5,000 observations was tested, and the scatter plot, represented in figure C.7 in Appendix C, section C.3, was analyzed, revealing some noticeable agglomerations. However, the number of distinct groups appeared high, and the boundaries between them were not clearly defined.

While t-SNE identifies clusters, it does not seem to be the most suitable approach for this context. Therefore, the decision was made to proceed with the PCA technique, which facilitates the application of clustering algorithms.

4.3. Clustering the data

After deciding to use PCA to reduce data dimensionality to three PCs and naming them, clustering was performed to determine whether clients could be successfully grouped based on their characteristics. The performance of four different methods, namely K-Means, Wards method, Complete Linkage, and DBSCAN, was analyzed to identify the most effective approach.

4.3.1. K-Means

Firstly, the k-means algorithm was applied, and two methods were used, which were the elbow and silhouette methods, in order to determine the optimal number of clusters for K-Means. The elbow method, represented in Figure C.8 in Appendix C, section C.3, plots the Within-Cluster Sum of Squares for different cluster numbers, identifying the elbow point, where adding more clusters has minimal impact [70]. Similarly, the silhouette method, shown in Figure C.9 in Appendix C, section C.3, calculates silhouette scores for clusters within a specific range, selecting the number with the highest score [71].

The Elbow Method shows a noticeable inflection point around $k = 4$, while the Silhouette Method reaches its peak at $k = 6$, but the difference between $k = 5$ and $k = 6$, and even $k = 4$, is minimal. Since no clear conclusion can be drawn, $k = 3, 4, 5$ and 6 were further evaluated using silhouette analysis. These tests are represented in Figures C.10, C.11, C.12 and C.13 in Appendix C, section C.3, with silhouette scores of 0.7088, 0.7632, 0.7938 and 0.8004 respectively. This indicates that 6 is the optimal number of clusters, however 5 is also a good choice because the silhouette values are very similar.

4.3.2. Hierarchical clustering

Ward Method

Secondly, hierarchical clustering was performed using the Ward linkage method. To determine the optimal number of clusters, a dendrogram was generated based on the linkage matrix (Figure C.14 in Appendix C, section C.3), which provides insight into how observations are grouped step by step, enabling the identification of natural divisions in the data. To decide how many clusters to choose, a practical approach is to cut the dendrogram where there is a significant jump in the distance between successive merges. This jump suggests that further merging results in combining highly different groups. In

this dendrogram, a large jump occurs just before forming five main clusters, indicating that cutting at this point is a suitable choice.

In addition, a silhouette analysis was performed to evaluate the optimal number of clusters, considering values of k ranging from 3 to 7, with the corresponding silhouette plots shown in Figures C.15, C.16, C.17, C.18 and C.19 in Appendix C, section C.3. The average silhouette scores obtained for each k were as follows: 0.6657 for $k=3$, 0.7596 for $k=4$, 0.7925 for $k=5$, 0.7980 for $k=6$, and 0.7262 for $k=7$, with $k=6$ obtaining the best score, however, as in the K-Means method, $k = 5$ produces a similar silhouette value.

Complete linkage method

Thirdly, hierarchical clustering was applied using the Complete Linkage method. As with the previous approach, a dendrogram was produced based on the linkage matrix (Figure C.20 in Appendix C, section C.3). In this dendrogram, a clear separation of the data is observed at a dissimilarity level of approximately 1.5 to 2.0. Cutting the dendrogram at this height suggests the presence of three well-defined clusters, as further merges beyond this point involve a substantial increase in dissimilarity.

Furthermore, a silhouette analysis was conducted for a range of cluster values ($k = 3$ to 7). The corresponding silhouette plots are presented in Figures C.21, C.22, C.23, C.24, and C.25 in Appendix C, section C.3, with the average silhouette scores obtained being 0.3618 for $k=3$, 0.5217 for $k=4$, 0.5210 for $k=5$, 0.4373 for $k=6$ and 0.4022 for $k=7$. The highest silhouette score was obtained for $k=4$, however, as in the K-Means and Ward method, $k = 5$ produces a similar silhouette value.

4.3.3. DBSCAN

Lastly, regarding the DBSCAN algorithm, there are two essential parameters that must be manually defined: *eps* (the maximum distance between two points for them to be considered neighbors) and *min_samples* (the minimum number of points required to form a dense region). Due to the importance of these parameters to DBSCAN, a grid search technique - a method that systematically tests combinations of parameter values to find the best outcome - was employed to determine the optimal combination, which returned the optimal values of *eps* = 0.3 and *min_samples* = 11. Subsequently, the DBSCAN

model was executed with these parameters, resulting in 4 clusters that can be visualized in Figure C.26 in Appendix C, section C.3, with a corresponding silhouette score of 0.6651.

4.3.4. Optimal number of clusters

In this chapter, the best clustering method and the optimal number of clusters based on the silhouette scores obtained for each algorithm and number of clusters was determined. Table 4.3 summarizes the experiments conducted for each method with different numbers of clusters. For DBSCAN, there was no need for experiments, as the algorithm automatically selects the optimal number of clusters.

TABLE 4.3. Silhouette scores obtained by algorithm and number of clusters

Algorithm	k = 3	k = 4	k = 5	k = 6	k = 7
K-Means	0.7088	0.7632	0.7938	0.8004	–
Ward Linkage	0.6657	0.7596	0.7925	0.7980	0.7262
Complete Linkage	0.3618	0.5217	0.5210	0.4373	0.4022
DBSCAN	–	0.6651	–	–	–

Looking at the silhouette scores, it was observed that both K-Means and Ward Linkage performed best with 6 clusters, while Complete Linkage and DBSCAN, showed better performance with 4 clusters, although their silhouette scores remained lower compared to K-Means and Ward Linkage. Furthermore, although K-Means and Ward Linkage yielded the highest performance with 6 clusters, the silhouette values for 5 clusters were remarkably similar. Therefore, considering the trade-off between model complexity and minimal information loss, it seemed reasonable to opt for 5 clusters instead of 6, as this resulted in only a marginal decrease in the silhouette score. In terms of the best method, it was found that K-Means slightly outperformed Ward Linkage, making it the best choice for this dataset.

4.3.5. Interpreting and naming the clusters

In this chapter, the 5 clusters formed by applying the K-Means method were interpreted, both with PCt's and the original variables.

Firstly, in order to better understand the characteristics of each cluster, the average values of the three PCt's and the debt fulfillment rate (proportion of good payers - the value

1 in the variable 'Cumprimento da dívida') were calculated for all observations grouped by cluster, which are presented in Table 4.4. The table also includes the percentage distribution of observations across clusters, providing insight into the relative size and prevalence of each group within the dataset.

TABLE 4.4. Average PC scores, debt fulfillment percentage, and observation distribution per cluster

Variable		Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Engagement vs. Churn Risk		-0.0700	0.7393	-0.7248	-0.5025	0.7840
Active Indebtedness		-0.3998	0.1061	0.1136	0.7699	0.9957
Payment Cycle Completion		-0.2424	0.1998	0.3184	-0.5747	-0.5151
Cumprimento da dívida (%)		56.2200%	0.2200%	98.9100%	75.5100%	0.1500%
Observation Percentage (%)		35.11%	30.33%	24.82%	7.46%	2.26%

- Cluster 0: Mixed profile (35.11% of observations)
 - Engagement vs. Churn Risk: -0.0700 (neutral)
 - Active Indebtedness: -0.3998 (low)
 - Payment Cycle Completion: -0.2424 (slightly incomplete)
 - Debt Fulfillment Rate: 56.22%

This cluster represents a neutral or mixed group. Customers show low indebtedness but have not fully completed their payment cycles. With just over half paying on time, these profiles do not exhibit clear risk or reliability patterns.

- Cluster 1: High-risk churners (30.33% of observations)
 - Engagement vs. Churn Risk: 0.7393 (high churn risk)
 - Active Indebtedness: 0.1061 (moderate)
 - Payment Cycle Completion: 0.1998 (slightly complete)
 - Debt Fulfillment Rate: 0.22%

These customers are likely deactivated and show poor payment behavior, despite not having the highest debt levels. Their combination of churn risk and almost

nonexistent payment fulfillment makes them a priority for credit recovery or write-off.

- Cluster 2: Ideal customers (24.82% of observations)
 - Engagement vs. Churn Risk: -0.7248 (high engagement)
 - Active Indebtedness: 0.1136 (low/moderate)
 - Payment Cycle Completion: 0.3184 (mostly complete)
 - Debt Fulfillment Rate: 98.91%

These are highly reliable and engaged customers, with regular SEPA usage and low-to-moderate indebtedness. They have largely completed their financial obligations and represent an ideal profile for retention and commercial campaigns.

- Cluster 3: Indebted but engaged (7.46% of observations)
 - Engagement vs. Churn Risk: -0.5025 (engaged)
 - Active Indebtedness: 0.7699 (high)
 - Payment Cycle Completion: -0.5747 (incomplete)
 - Debt Fulfillment Rate: 75.51%

These customers are still involved and tend to pay, but carry significant debt and have not yet completed their repayment journey. With proper support, this group may migrate toward the ideal profile.

- Cluster 4: Disengaged and heavily indebted (2.26% of observations)
 - Engagement vs. Churn Risk: 0.7840 (high churn risk)
 - Active Indebtedness: 0.9957 (very high)
 - Payment Cycle Completion: -0.5151 (incomplete)
 - Debt Fulfillment Rate: 0.15%

This segment includes the riskiest customers: likely deactivated, with high outstanding balances and poor payment performance. They require immediate attention for loss mitigation or collections.

Cluster Profiles Based on Real Data

Secondly, in order to complement the PCA-based clustering interpretation, the same clusters were analyzed using the original dataset with real, non-normalized values and without the transformation of categorical to numerical variables, which are presented in

Table 4.5. This allowed for a more realistic and intuitive understanding of the characteristics of each group.

TABLE 4.5. Average original variable values per cluster (non-normalized)

Variable	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
'Saldo Total'	61.89	248.50	63.06	333.77	484.65
'30 a 60'	0.16	0.06	0.00	0.14	0.16
'SAS - Ciclo'	9.50	7.55	17.71	14.03	11.12
'VEP - Nº Prestações'	0.04	0.03	0.00	30.23	35.96
'VEP - Nº Prestações Venc'	0.00	0.00	0.00	1.11	13.20
'PPP - Nº Prestações'	0.38	0.76	0.02	0.76	0.01
'PPP - Valor do Plano'	7.90	0.02	0.36	16.24	0.31
'Longevidade'	345.15	801.63	77.89	37.64	993.93
'Data Criação NIF (Diferença Dias)'	4247.28	2881.60	3949.66	3044.93	2046.97
'Cumprimento da dívida' (%)	56.22%	0.22%	98.91%	75.51%	0.15%
'CAT_CONTA_CONTRATO'	PTC	MEO	PTC	PTC	PTC
'Sub-Segmento'	Consumo	Consumo	Consumo	Consumo	Consumo
'Estrangeiro'	N	N	N	N	N
'Dissolução'	N	N	N	N	N
'Insolvência'	N	N	N	N	N
'Insolvência - PER'	N	N	N	N	N
'VEP'	N	N	N	S	S
'PPP'	N	N	N	N	N
'Acompanhamento'	N	N	N	N	N
'Venda Créditos'	N	N	N	N	N
'SEPA'	N	N	S	S	N
'Serviço_Estado'	Ativo	Desativos	Ativo	Ativo	Desativos

Cluster 0 *Previous interpretation:* Mixed group with moderate repayment rate (56.22%) and no clear pattern. *Real data insights:* Customers in this cluster have low average debt ('Saldo Total' 61.89) and almost no active credit plans ('VEP' 0.04, 'PPP'

0.38). Their average tenure is 345 days, with relatively old NIFs (11.6 years). Categorical variables show they are typically 'PTC', non-'SEPA' users, non-'VEP' and their accounts are active. Conclusion: Consistent with the initial interpretation. A diverse and stable group with low debt and moderate repayment behavior, but no strong indicators of either risk or excellence.

Cluster 1 *Previous interpretation:* High-Risk customers with extremely low repayment rate (0.22%). *Real data insights:* These clients have average debt ('Saldo Total' 248.50), no active 'VEP' or 'PPP' plans, and high tenure (801 days), but relatively young NIFs. Most are 'MEO', 'SEPA'-compliant, non-'VEP' and their accounts are deactivated. Conclusion: Fully aligned with the previous interpretation. A highly disengaged and risky segment, despite average debt, these customers exhibit extremely low repayment rates, indicating poor payment behavior.

Cluster 2 *Previous interpretation:* Ideal customers (repayment rate 98.91%). *Real data insights:* These clients have low to moderate average debt ('Saldo Total' 63.06), almost no active credit agreements ('VEP' 0.03, 'PPP' 0.00). Their average tenure is short (77 days), with relatively old NIFs (10.8 years). They are 'PTC', 'SEPA' users, non-'VEP' and accounts are active. Conclusion: Fully consistent with the previous analysis. A reliable and highly engaged group with excellent payment behavior, making them valuable for retention and commercial efforts.

Cluster 3 *Previous interpretation:* Indebted but engaged (repayment rate 75.51%). *Real data insights:* Clients in this cluster have significant average debt ('Saldo Total' 333.77), several active installments ('VEP' 30.23, 'PPP' 0.76), moderate overdue installments (1.11) and very high plan value of 'PPP'. Their average tenure is short (38 days), with NIFs averaging 8.3 years. They are typically 'PTC', 'SEPA'-compliant, with 'VEP' and accounts active. Conclusion: Matches the initial interpretation. High-risk but potentially recoverable customers, indebted but still engaged and regularly paying. With adequate support, they may migrate towards the ideal profile.

Cluster 4 *Previous interpretation:* Heavily indebted and disengaged (repayment rate 0.15%). *Real data insights:* This group has the highest average debt ('Saldo Total' 484.65), a large number of active and overdue installments ('VEP' 35.96, overdue 13.20), and

almost no 'PPP' activity. Their tenure is long (994 days), with relatively young NIFs (5.6 years). They are mainly 'PTC', non-'SEPA', with 'VEP' and accounts deactivated. Conclusion: Fully confirms the initial interpretation. The most critical risk segment, requiring urgent attention for loss mitigation or collections.

The interpretation using real variables complements is consistent with the analysis conducted through PCts, with both approaches identifying similar patterns within the clusters. For example, in Cluster 0, the PCts interpretation pointed to a mixed and moderate profile, which is corroborated by the real variable analysis, showing customers with low debt and moderate payment behavior. In Cluster 2, the combination of a high debt fulfillment rate and low debt in the real variables reinforces the PCts interpretation of an ideal group with excellent payment behavior.

Figure 4.4 shows the distribution of the clusters in space, based on the three principal components. This visualization was generated using a sample of five thousand observations to improve clarity and interpretability. Figures C.27, C.28, and C.29 in Appendix C, section C.3, illustrate the clusters from different viewing angles as the one presented in figure 4.4. This 3D visualization of the clustering results reveals not only the general separation between groups, but also some potential substructures within the clusters themselves. For example, Cluster 3 appears to be spatially divided into two distinct subgroups, despite being labeled with the same cluster ID. This may indicate an internal heterogeneity within the cluster or a possible overlap with another cluster.

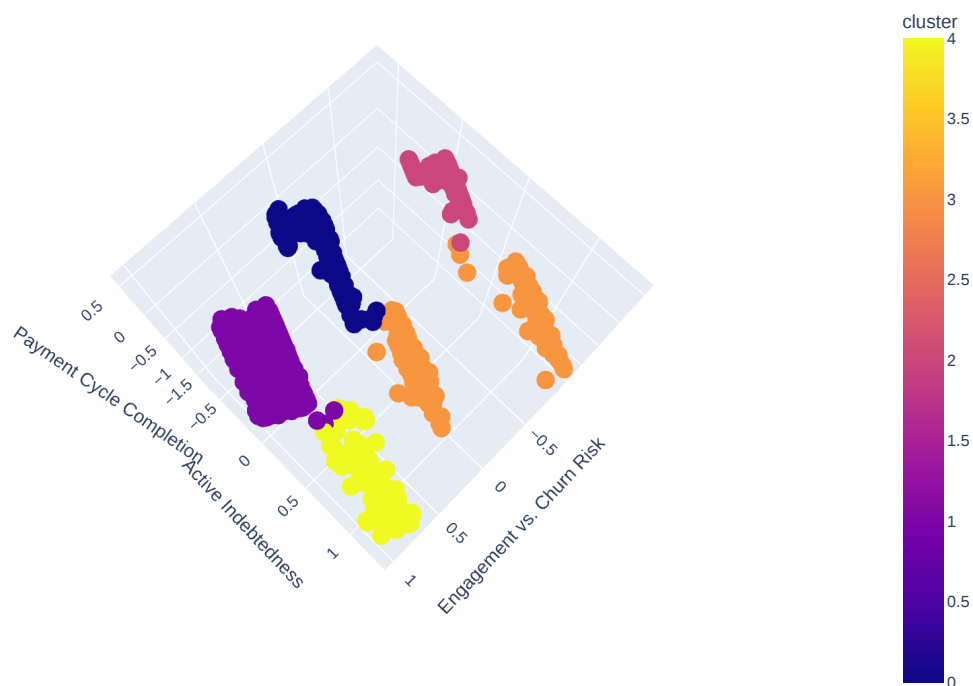


FIGURE 4.4. 3D visualization of K-Means clusters top-down perspective

This visualization is particularly helpful for understanding the spatial distribution of the clusters. It highlights their relative positioning, reveals which ones are close to each other, and reinforces the observation of possible overlaps or internal divisions, such as the one noted in Cluster 3.

4.4. Second experiment

While performing the first experiment, it was noted that, in the analysis of feature importance, some variables consistently emerged as a Top15 feature throughout the supervised learning models, which are listed in table 4.2. Besides this, variables with a high KMO score in unsupervised learning, as discussed in section 4.2.1, can be interpreted as significant for unsupervised learning techniques. When comparing these variables across both supervised and unsupervised learning approaches, the following were found to be common to both: 'Longevidade', '30 a 60', 'Saldo Total', 'SEPA', 'PPP', 'VEP - N° Prestações Venc', 'Data Criação NIF (Diferença Dias)', and 'Serviço_Estado_Desativos'.

This is not a mere coincidence, rather their presence in both analyses suggests that these variables are not only highly informative in predicting debt payment (supervised learning) but also strongly correlated with other key features, effectively capturing underlying structures within the dataset (unsupervised learning).

Hence, a further analysis was conducted to evaluate the performance of all models using only these variables in order to understand if the dataset can be simplified.

4.4.1. Supervised Learning models

First, the correlations were analyzed. These remained consistent with the previous analysis, both in relation to the target variable and between the independent variables themselves. It is important to highlight that no pairwise correlations exceeded 0.7 or fell below -0.7, and therefore no variable was removed on the basis of multicollinearity.

The confusion matrices and classification reports can be found in Tables C.20 to C.33, as well as the accuracy and loss curves from the training of the Neural Network, shown in Figure C.30, all presented in Appendix C, section C.4. Upon analysis and comparison with the results obtained using the full set of features, it can be concluded that the overall performance loss across all models does not exceed 3% in the evaluation metrics. Considering this trade-off - a substantial reduction in dimensionality with only a slight decrease in performance - the use of only the common features proves to be a highly valuable and efficient approach.

4.4.2. Unsupervised Learning models

Since the unsupervised analysis already involved dimensionality reduction through PCA, which reduced the feature space to three principal components, performing a second experiment using a pre-selected subset of features was deemed unnecessary, due to the fact that the results would be very similar. Therefore, further reducing the feature set prior to PCA would not provide additional insight or benefit to the clustering results.

4.5. Comparison with Existing Literature

Supervised Learning

The results obtained from the supervised learning models were compared to previous studies related to churn detection, profitability prediction, and fraud detection in the

telecommunications sector. Although these studies focus on slightly different problems than the one discussed in this dissertation, they provide useful performance benchmarks. As shown in Tables A.1, A.3, and A.2, all presented in Appendix A, the best performing model in this study - Random Forest - achieved comparable performance as the authors of studies [42] and [5]. When it comes to the remaining studies, the predictive performance of the models in this study was generally superior.

Unsupervised Learning

Regarding the literature review on unsupervised learning techniques, Table A.4 in Appendix A presents an overview of the unsupervised learning methods used in previous studies. Regarding dimensionality reduction, the reviewed studies did not apply this technique before clustering, so there is no comparison term. However, the studies found focused on using different clustering methods, and a common pattern observed is the strong performance of K-Means clustering, which aligns with the findings of this research.

Supervised and Unsupervised Learning combined

Several studies in the literature have explored hybrid approaches that combine supervised and unsupervised learning methods. In most of these cases, the workflow involves applying dimensionality reduction techniques, such as PCA or t-SNE, to the dataset prior to the application of supervised learning models. This process reduces data complexity and helps improve model efficiency and generalization. As summarized in Table A.5 in Appendix A, such integrated approaches have generally led to improved model robustness and accuracy.

In contrast to these integrated approaches, the present work applied supervised and unsupervised learning methods independently. Nevertheless, one experiment conducted in this dissertation bears some resemblance to those hybrid methodologies. In a secondary analysis, dimensionality reduction was performed based on feature importance derived from supervised learning models and KMO analysis in the context of unsupervised learning. The performance of supervised models was then compared before and after the reduction. However, this was not the primary focus of this dissertation.

Although most studies reviewed in the literature apply unsupervised learning techniques prior to supervised learning, the approach taken by the authors of [52] is more similar to the one adopted in this dissertation. In their work, XGBoost was first used to predict customer churn, followed by K-Means clustering to segment the predicted churners. This sequential approach enabled the company to target retention strategies more effectively, resulting in a 0.93% reduction in churn.

CHAPTER 5

Discussions and Conclusions

5.1. Overview of the Dissertation

This study aimed to analyze and model customer behavior regarding debt payment within a Portuguese telecommunications company. The process began with comprehensive data preprocessing, including handling missing values, outlier detection and treatment, feature engineering, normalization, categorical variable encoding, target variable creation, and dataset balancing.

Subsequently, various supervised learning algorithms were applied to predict the newly created target variable 'Cumprimento da divida'. The models tested were Logistic Regression, Random Forest, k-NN, and Neural Networks. Among these, the Random Forest model achieved the highest performance, with an accuracy of 99%, making it the most effective model for this task.

Following the supervised learning phase, unsupervised learning techniques were employed in two main steps. First, dimensionality reduction techniques were evaluated, with both PCA and t-SNE being tested, with PCA being selected as the preferred technique due to its superior interpretability and performance, and three principal components being retained and meaningfully named. Second, clustering techniques were applied to the PCA-transformed dataset, including K-Means, Ward's method, Complete Linkage, and DBSCAN. K-Means clustering outperformed the other methods, and the optimal number of clusters was determined to be five. These clusters were interpreted using both the principal components and the original dataset features to gain a deeper understanding of customer behavior.

5.2. Discussion of Results

Although this study did not implement an integrated approach in the traditional sense, it demonstrated how supervised and unsupervised learning methods can be used in a complementary way. For example, after classifying clients as compliant or non-compliant

using the supervised model, one can analyze their corresponding clusters to gain additional insight. If a compliant client falls into a cluster mostly composed of non-compliant clients, this could signal a potential future risk.

This dual-layer analysis, prediction followed by segmentation, offers an enriched perspective, similar to that employed by the authors of article [52]. Furthermore, the use of PCA allowed the identification of underlying patterns and dimensionality reduction, providing interpretable axes for visual cluster exploration and reducing computational complexity, as shown in Figures C.27 to 4.4 in Appendix C, Section C.3.

From a business perspective, this approach enables the company to test hypothetical scenarios by altering specific client features to assess their potential impact on payment behavior. This can be particularly useful for high-risk clients with substantial debt, allowing the company to make informed decisions, such as whether to continue or terminate the contract, or to implement targeted interventions.

For instance, imagine a company receives a business client in March who has debt but is a good payer. Hypothetical scenarios could be tested in order to predict if the client will continue to be a good payer. For example, if the client does not have a '30 a 60' value in March but has one in April, can the company expect the client to remain a good payer? Or, if the client used to pay via 'SEPA' in March but switches to a different payment method in April, will they continue to be a good payer? To test these scenarios, one would simply alter the value of the '30 a 60' variable or the 'SEPA' payment method and observe the supervised model's prediction, keeping other variables constant. As this is a more manual analysis, it would be most useful for clients with the highest debt, as they are causing the most significant financial loss to the company. If such analyses reveal that a good payer might turn into a bad payer if their 'Saldo Total' increases, the company could take preventive actions, such as terminating the contract or encouraging the client to avoid accumulating a higher balance.

Unsupervised analysis also offers value in identifying clients within ambiguous clusters such as clusters 0 and 3, where compliance behavior is less conclusive. By comparing these clients' characteristics with those from more definitive clusters, one can estimate whether they are likely to remain compliant or not. For example, if a client in cluster 0 is a good payer and has characteristics similar to those in cluster 2 (good payers), it is likely that

the client will continue to be a good payer. Another example is a client in cluster 0 who is a non-payer but has characteristics similar to clients in clusters 1 or 4 (non-payers), indicating a higher likelihood of transitioning from a non-payer to a good payer.

5.3. Conclusion

In conclusion, this study successfully demonstrated how both supervised and unsupervised learning techniques can be used to understand and predict client payment behavior in the telecommunications sector. The findings are not only consistent with existing literature but also offer practical applications that can support more informed decision-making of the telecom company. Although initially applied independently, the subsequent joint analysis of both learning paradigms proved to be complementary, providing a robust analytical framework for future business applications and research developments.

5.4. Recommendations for Future Work

Several directions for future research can be proposed based on the findings of this study:

- **Focused Clustering on Non-Compliant Clients:** Conduct clustering exclusively on non-compliant clients to identify subgroups with distinct behavioral profiles, which could enhance the segmentation strategy and risk assessment.
- **Alternative Preprocessing Strategies:** Experiment with different preprocessing techniques (for example, normalization methods, outlier handling) to improve the performance of clustering algorithms, especially with the goal of increasing the silhouette score.
- **Application of Time Series Analysis:** Utilize time series forecasting models to predict the target variable based on historical debt behavior. This could offer an automatic and dynamic alternative to the static approach used in this study.
- **Simulation of Feature Changes:** Expand on the idea of testing client profiles by simulating changes in individual features to evaluate their impact on predicted compliance, supporting proactive risk management strategies.

Bibliography

- [1] A. K. Ahmad, A. Jafar, and K. Aljoumaa, "Customer churn prediction in telecom using machine learning in big data platform," *J Big Data*, vol. 6, no. 1, Dec. 2019. DOI: 10.1186/s40537-019-0191-6.
- [2] H. Jain, A. Khunteta, and S. Srivastava, "Churn prediction in telecommunication using logistic regression and logit boost," in *Procedia Computer Science*, Elsevier B.V., 2020, pp. 101–112. DOI: 10.1016/j.procs.2020.03.187.
- [3] J. Gupta, *Data analytics in telecommunications*, <https://www.quantexa.com/resources/data-analytics-in-telecommunications/>, Accessed: Dec. 23, 2024, 2024.
- [4] N. Nazareth and Y. V. R. Reddy, "Financial applications of machine learning: A literature review," *Elsevier Ltd.*, Jun. 2023. DOI: 10.1016/j.eswa.2023.119640.
- [5] S. K. Wagh, A. A. Andhale, K. S. Wagh, J. R. Pansare, S. P. Ambadekar, and S. H. Gawande, "Customer churn prediction in telecom sector using machine learning techniques," *Results in Control and Optimization*, vol. 14, Mar. 2024. DOI: 10.1016/j.rico.2023.100342.
- [6] C. Wang, C. Rao, F. Hu, X. Xiao, and M. Goh, "Risk assessment of customer churn in telco using felcnn-lstm model," *Expert Syst Appl*, vol. 248, Aug. 2024. DOI: 10.1016/j.eswa.2024.123352.
- [7] F. Naz and F. Popowich, *Mining retail telecommunication data to predict profitability*, Unpublished, 2019.
- [8] S. Naeem, A. Ali, S. Anam, and M. M. Ahmed, "An unsupervised machine learning algorithms: Comprehensive review," *International Journal of Computing and Digital Systems*, vol. 13, no. 1, pp. 911–921, 2023, Cited by: 49; All Open Access, Gold Open Access. DOI: 10.12785/ijcds/130172. [Online]. Available: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85161879817&doi=10.12785%2fijcds%2f130172&partnerID=40&md5=4c8f17e978de881b785605ff53d8e635>.

- [9] V. Renuka Devi, G. Bharathi, and G. Prasad, "Prediction of customer churn in telecom sector using clustering technique," *International Journal of Engineering and Advanced Technology*, vol. 8, no. 6 Special Issue 2, pp. 826–832, 2019, Cited by: 0; All Open Access, Bronze Open Access. DOI: 10.35940/ijeat.F1207.0886S219. [Online]. Available: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85073727612&doi=10.35940%2fijeat.F1207.0886S219&partnerID=40&md5=ec4e8302e0601fb9f7bb76f0534ce10b>.
- [10] H. Wibowo and K. P. Sinaga, "Telecommunication analytics based on customer segmentation using unsupervised algorithms," Cited by: 1, 2021. DOI: 10.1109/ICORIS52787.2021.9649598. [Online]. Available: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85124365145&doi=10.1109%2fICORIS52787.2021.9649598&partnerID=40&md5=e0f24f3e74cd869fc3ac525ac1b38c90>.
- [11] J. S. Saltz, "Crisp-dm for data science: Strengths, weaknesses and potential next steps," pp. 2337–2344, 2021. DOI: 10.1109/BigData52589.2021.9671634.
- [12] D. S. Central, *Crisp-dm: A standard methodology to ensure a good outcome*, Accessed: 2025-03-03, 2020. [Online]. Available: <https://www.datasciencecentral.com/crisp-dm-a-standard-methodology-to-ensure-a-good-outcome/>.
- [13] M. A. Akasheh, E. F. Malik, O. Hujran, and N. Zaki, "A decade of research on machine learning techniques for predicting employee turnover: A systematic literature review," *Expert Syst Appl*, vol. 238, Mar. 2024. DOI: 10.1016/j.eswa.2023.121794.
- [14] H. Gao et al., "Machine learning in business and finance: A literature review and research opportunities," *Financial Innovation*, vol. 10, no. 1, p. 86, 2024. DOI: 10.1186/s40854-024-00629-z.
- [15] X. Tian, Z. Y. Tian, S. F. A. Khatib, and Y. Wang, "Machine learning in internet financial risk management: A systematic literature review," *Public Library of Science*, Apr. 2024. DOI: 10.1371/journal.pone.0300195.
- [16] V. Nasteski, "An overview of the supervised machine learning methods," *HORIZONS.B*, vol. 4, pp. 51–62, 2017. DOI: 10.20544/horizons.b.04.1.17.p05.
- [17] GeeksforGeeks, *Supervised machine learning*, Accessed: 2025-02-26, 2025. [Online]. Available: <https://www.geeksforgeeks.org/supervised-machine-learning/>.

- [18] H. K. Gianey and R. Choudhary, "Comprehensive review on supervised machine learning algorithms," in *Proceedings - 2017 International Conference on Machine Learning and Data Science, MLDS 2017*, IEEE, 2017, pp. 38–43. DOI: 10.1109/MLDS.2017.11.
- [19] F. J. Bargagli-Stoffi, J. Niederreiter, and M. Riccaboni, *Supervised learning for the prediction of firm dynamics*. Springer, 2020. DOI: 10.1007/978-3-030-66891-4_2.
- [20] GeeksforGeeks, *Unsupervised learning*, Accessed: 2025-02-26, 2023. [Online]. Available: <https://www.geeksforgeeks.org/unsupervised-learning/>.
- [21] S. Hidayati, A. Darmaliana, and R. Riski, "Comparison of k-means, fuzzy c-means, fuzzy gustafson kessel, and dbscan for village grouping in surabaya based on poverty indicators," *Jurnal Pendidikan Matematika (Kudus)*, vol. 5, p. 185, Dec. 2022. DOI: 10.21043/jpmk.v5i2.16552.
- [22] A. F. Jadama, B. Jobarteh, M. M. Islam, and M. K. Toray, "Ensemble learning: Methods, techniques, application," *ResearchGate*, 2024. [Online]. Available: https://www.researchgate.net/publication/381773312_Ensemble_Learning_Methods_Techniques_Application.
- [23] S. F. Bilal, A. A. Almazroi, S. Bashir, F. H. Khan, and A. A. Almazroi, "An ensemble based approach using a combination of clustering and classification algorithms to enhance customer churn prediction in telecom industry," *PeerJ Computer Science*, vol. 8, 2022, Cited by: 25; All Open Access, Gold Open Access, Green Open Access. DOI: 10.7717/PEERJ-CS.854. [Online]. Available: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85125836092&doi=10.7717%2fPEERJ-CS.854&partnerID=40&md5=78c11c86bab29151d6a57850bacc5929>.
- [24] M. Bagri, J. K. Singh, M. Abhilash, R. Sunitha, and S. Kumar, "Churn analysis in telecommunication industry," Cited by: 2, 2018, pp. 126–132. DOI: 10.1109/ICACE.2018.8686852. [Online]. Available: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85064883375&doi=10.1109%2fICACE.2018.8686852&partnerID=40&md5=a934853ddac1e65ab5b84850d19b0297>.
- [25] M. ukanovi, L. Kaelan, S. Vukovi, I. Martinovi, and M. alasan, "A machine learning approach for time series forecasting with application to debt risk of the montenegrin

- electricity industry,” *Energy Reports*, vol. 9, pp. 362–369, Sep. 2023. DOI: 10.1016/j.egy.2023.05.240.
- [26] M. Abdoli, M. Akbari, and J. Shahrabi, “Bagging supervised autoencoder classifier for credit scoring,” Aug. 2021, Available: <http://arxiv.org/abs/2108.07800>.
- [27] S. Tyagi, “Analyzing machine learning models for credit scoring with explainable ai and optimizing investment decisions,” 2022.
- [28] G. Li, H. D. Ma, R. Y. Liu, M. D. Shen, and K. X. Zhang, “A two-stage hybrid default discriminant model based on deep forest,” *Entropy*, vol. 23, no. 5, 2021. DOI: 10.3390/e23050582.
- [29] V. L. N. Gorle and S. Panigrahi, “A semi-supervised anti-fraud model based on integrated xgboost and bigru with self-attention network: An application to internet loan fraud detection,” *Multimed Tools Appl*, vol. 83, no. 19, pp. 56 939–56 964, 2024. DOI: 10.1007/s11042-023-17681-z.
- [30] R. Najem, M. F. Amr, A. Bahasse, and M. Talea, “Advancements in artificial intelligence and machine learning for stock market prediction: A comprehensive analysis of techniques and case studies,” *Procedia Computer Science*, pp. 198–204, 2024. DOI: 10.1016/j.procs.2023.12.193.
- [31] C. Manteigas and N. António, “Understanding and predicting lapses in mortgage life insurance using a machine learning approach,” *Expert Syst Appl*, vol. 255, 2024. DOI: 10.1016/j.eswa.2024.124753.
- [32] M. A. Khan, M. A. I. Khan, M. Aref, and S. F. Khan, “Cluster & rough set theory based approach to find the reason for customer churn,” *International Journal of Applied Business and Economic Research*, vol. 14, no. 1, pp. 439–455, 2016, Cited by: 6. [Online]. Available: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-84971264795&partnerID=40&md5=e11059c47a9363faa40261de98666166>.
- [33] M. Gök, T. Özyer, and J. Jida, “A case study for the churn prediction in turksat internet service subscription,” Cited by: 4, 2015, pp. 1220–1224. DOI: 10.1145/2808797.2808821. [Online]. Available: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-84962585766&doi=10.1145%2f2808797.2808821&partnerID=40&md5=16e7fa95875ab126ebd663d1e5c4da04>.

- [34] M. J. P. et al., “The prisma 2020 statement: An updated guideline for reporting systematic reviews,” *J Clin Epidemiol*, vol. 134, pp. 178–189, Jun. 2021. DOI: 10.1016/j.jclinepi.2021.03.001.
- [35] D. V. Parums, “Editorial: Review articles, systematic reviews, meta-analysis, and the updated preferred reporting items for systematic reviews and meta-analyses (prisma) 2020 guidelines,” Aug. 2021. DOI: 10.12659/MSM.934475.
- [36] P. K. S. P. Sahoo and R. Uchil, “Identification of critical success factors for leveraging industry 4.0 technology and research agenda: A systematic literature review using prisma protocol,” May 2024. DOI: 10.1108/APJBA-03-2022-0105.
- [37] K. R. Felizardo, E. Y. Nakagawa, S. G. MacDonell, and J. C. Maldonado, “A visual analysis approach to update systematic reviews,” in *ACM International Conference Proceeding Series*, Association for Computing Machinery, 2014. DOI: 10.1145/2601248.2601252.
- [38] A. Abdallah, M. A. Maarof, and A. Zainal, “Fraud detection system: A survey,” *Journal of Network and Computer Applications*, vol. 68, pp. 90–113, Jun. 2016. DOI: 10.1016/j.jnca.2016.04.007.
- [39] B. Garimella, G. V. S. N. R. V. Prasad, and M. H. M. K. Prasad, “Churn prediction using optimized deep learning classifier on huge telecom data,” *Journal of Ambient Intelligence and Humanized Computing*, vol. 14, no. 3, pp. 2007–2028, Mar. 2023. DOI: 10.1007/s12652-021-03413-4.
- [40] S. S. Poudel, S. Pokharel, and M. Timilsina, “Explaining customer churn prediction in telecom industry using tabular machine learning models,” *Machine Learning with Applications*, vol. 17, p. 100567, Sep. 2024. DOI: 10.1016/j.mlwa.2024.100567.
- [41] S. M. Shrestha and A. Shakya, “A customer churn prediction model using xgboost for the telecommunication industry in nepal,” in *Procedia Computer Science*, Elsevier B.V., 2022, pp. 652–661. DOI: 10.1016/j.procs.2022.12.067.
- [42] J. K. Vieri, T. A. Munandar, D. B. Srisulistiwati, D. Handayani, A. No, and T. S. Lestari, “Comparative study of classification algorithms for customer decisions on telecommunication products using supervised learning,” 2023.

- [43] E. Diaz-Aviles et al., *Towards real-time customer experience prediction for telecommunication operators*, Online, Aug. 2015. [Online]. Available: <http://arxiv.org/abs/1508.02884>.
- [44] A. A. et al., “Customer churn prediction in the telecommunication sector using a rough set approach,” *Neurocomputing*, vol. 237, pp. 242–254, May 2017. DOI: 10.1016/j.neucom.2016.12.009.
- [45] L. D. et al., “Modeling approaches for silent attrition prediction in payment networks,” in *Proceedings - 20th IEEE International Conference on Machine Learning and Applications, ICMLA 2021*, Institute of Electrical and Electronics Engineers Inc., 2021, pp. 409–414. DOI: 10.1109/ICMLA52953.2021.00070.
- [46] A. Chouiekh and E. H. I. E. Haj, “Convnets for fraud detection analysis,” in *Procedia Computer Science*, Elsevier B.V., 2018, pp. 133–138. DOI: 10.1016/j.procs.2018.01.107.
- [47] X. Hu, H. Chen, S. Liu, H. Jiang, G. Chu, and R. Li, “Btg: A bridge to graph machine learning in telecommunications fraud detection,” *Future Generation Computer Systems*, vol. 137, pp. 274–287, Dec. 2022. DOI: 10.1016/j.future.2022.07.020.
- [48] A. Gaikwad, T. Ghodke, A. Jadhav, M. Pande, and S. Mirchandani, “Profit prediction for businesses using machine learning algorithms,” in *Proceedings of the 8th International Conference on Communication and Electronics Systems, ICCES 2023*, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 1039–1044. DOI: 10.1109/ICCES57224.2023.10192797.
- [49] K. K. Sharma, A. Seal, E. Herrera-Viedma, and O. Krejcar, “An enhanced spectral clustering algorithm with s-distance,” *Symmetry*, vol. 13, no. 4, 2021, Cited by: 22; All Open Access, Gold Open Access. DOI: 10.3390/sym13040596. [Online]. Available: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85104180766&doi=10.3390%2fsym13040596&partnerID=40&md5=ab8493bac9ae480e0b7d38afbceca2a5>.
- [50] K. Shahid and S. Chaudhury, “Scalable clustering and applications,” Cited by: 0, 2016. DOI: 10.1145/3009977.3010073. [Online]. Available: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85014856715&doi=10.1145%2f3009977.3010073&partnerID=40&md5=aa9e6d31608b37f8462d4f119593fc5d>.

- [51] S. M. Kosti, M. I. Simi, and M. V. Kosti, "Social network analysis and churn prediction in telecommunications using graph theory," *Entropy*, vol. 22, no. 7, 2020, Cited by: 30; All Open Access, Gold Open Access, Green Open Access. DOI: 10.3390/e22070753. [Online]. Available: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85088565953&doi=10.3390%2fe22070753&partnerID=40&md5=18119c498a2915d2a410af4b1c19d062>.
- [52] Y. Gu, A. R. Malicdem, J. S. Dela Cruz, and T. D. Palaoag, "Using big data analysis to retain customers for telecom industry," Cited by: 0, 2019, pp. 38–43. DOI: 10.1145/3330482.3330510. [Online]. Available: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85071128980&doi=10.1145%2f3330482.3330510&partnerID=40&md5=fb5b1d286ab7611b24283314e56f1449>.
- [53] H. Nutalapati, A. Hayat, R. Zheng, C. H. Li, N. Prakoso, and N. M. Tiglao, "Machine learning-based predictive analytics for customer churn in the telecom industry," Cited by: 0, 2024. DOI: 10.1109/ISNCC62547.2024.10758995. [Online]. Available: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85212228512&doi=10.1109%2fISNCC62547.2024.10758995&partnerID=40&md5=c6d2b84bf471c9d1024c26a30e3be74c>.
- [54] S. Preetha and R. Rayapeddi, "Predicting customer churn in the telecom industry using data analytics," Cited by: 3, 2018, pp. 38–43. DOI: 10.1109/ICGCIoT.2018.8753096. [Online]. Available: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85069518962&doi=10.1109%2fICGCIoT.2018.8753096&partnerID=40&md5=a97635ff406fb7bc2f5c7f98bdb192fc>.
- [55] K. F. Googerdchi, S. Asadi, and S. M. Jafari, "Customer churn modeling in telecommunication using a novel multiobjective evolutionary clustering-based ensemble learning," *PLoS ONE*, vol. 19, no. 6 June, 2024, Cited by: 1; All Open Access, Gold Open Access, Green Open Access. DOI: 10.1371/journal.pone.0303881. [Online]. Available: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85195438770&doi=10.1371%2fjournal.pone.0303881&partnerID=40&md5=95a03e3becbfe9d75e28b60edd5840c3>.
- [56] E. Sivasankar and J. Vijaya, "Hybrid ppfcm-ann model: An efficient system for customer churn prediction through probabilistic possibilistic fuzzy clustering and

- artificial neural network,” *Neural Computing and Applications*, vol. 31, no. 11, pp. 7181–7200, 2019, Cited by: 34. DOI: 10.1007/s00521-018-3548-4. [Online]. Available: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85047660042&doi=10.1007%2fs00521-018-3548-4&partnerID=40&md5=d4769472159ee4956b1cdf6e6e26b5bad>.
- [57] R. A. Moral et al., “Profiling television watching behavior using bayesian hierarchical joint models for time-to-event and count data,” *IEEE Access*, vol. 10, pp. 113 018–113 027, 2022, Cited by: 1; All Open Access, Gold Open Access, Green Open Access. DOI: 10.1109/ACCESS.2022.3215682. [Online]. Available: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85140712174&doi=10.1109%2fACCESS.2022.3215682&partnerID=40&md5=6e4d5547e48c8f237cb24b3fe5561b2a>.
- [58] J. Quinn and L. Sparks, “Research frontiers in wholesale distribution (editorial),” *The International Review of Retail Distribution and Consumer Research*, vol. 17, Sep. 2007. DOI: 10.1080/09593960701505736.
- [59] T. Enderle, “On the impact of billing cycles on compensations in network slas,” 2021. DOI: 10.23919/ONDM51796.2021.9492378.
- [60] C.-C. Chou, N.-C. Hwang, G. Schneider, T. Wang, C.-W. Li, and W. Wei, “Using smart contracts to establish decentralized accounting contracts: An example of revenue recognition,” *Journal of Information Systems*, vol. 35, pp. 17–52, Sep. 2021. DOI: 10.2308/ISYS-19-009.
- [61] M. Boutros, N. Clara, and F. Gomes, “Borrow now, pay even later: A quantitative analysis of student debt payment plans,” *Journal of Financial Economics*, vol. 159, p. 103 898, 2024, ISSN: 0304-405X. DOI: <https://doi.org/10.1016/j.jfineco.2024.103898>.
- [62] S. Sharma and S. Chatterjee, “Winsorization for robust bayesian neural networks,” *Entropy*, vol. 23, no. 11, 2021, Cited by: 22. DOI: 10.3390/e23111546.
- [63] P. J. Muhammad Ali, “Investigating the impact of min-max data normalization on the regression performance of k-nearest neighbor with different similarity measurements,” *ARO-The Scientific Journal of Koya University*, vol. 10, no. 1, pp. 85–91, 2022. DOI: 10.14500/aro.10955.

- [64] M. K. Dahouda and I. Joe, "A deep-learned embedding technique for categorical features encoding," *IEEE Access*, vol. 9, pp. 114 381–114 391, 2021. DOI: 10.1109/ACCESS.2021.3104357. [Online]. Available: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85113332296&doi=10.1109%2fACCESS.2021.3104357&partnerID=40&md5=a2d4f4f7f816b483a6fb9707c7fca952>.
- [65] R. Mohammed, J. Rawashdeh, and M. Abdullah, "Machine learning with oversampling and undersampling techniques: Overview study and experimental results," in *2020 11th International Conference on Information and Communication Systems (ICICS)*, 2020, pp. 243–248. DOI: 10.1109/ICICS49469.2020.239556.
- [66] R. Wilcox, "Multicollinearity and ridge regression: Results on type i errors, power and heteroscedasticity," *Journal of Applied Statistics*, vol. 46, pp. 1–12, Sep. 2018. DOI: 10.1080/02664763.2018.1526891.
- [67] D. H. Arsham and M. Lovric, "Bartlett's test," *International Encyclopedia of Statistical Science*, vol. 2, pp. 20–23, Mar. 2011. DOI: 10.1007/978-3-642-04898-2_132.
- [68] A. Shkeer and Z. Awang, "Exploring the items for measuring the marketing information system construct: An exploratory factor analysis," *International Review of Management and Marketing*, vol. 9, pp. 87–97, Oct. 2019. DOI: 10.32479/irmm.8622.
- [69] H. Liu et al., "Using t-distributed stochastic neighbor embedding (t-sne) for cluster analysis and spatial zone delineation of groundwater geochemistry data," *Journal of Hydrology*, vol. 597, p. 126 146, 2021. DOI: <https://doi.org/10.1016/j.jhydrol.2021.126146>.
- [70] D. Marutho, S. Hendra Handaka, E. Wijaya, and Muljono, "The determination of cluster number at k-mean using elbow method and purity evaluation on headline news," 2018, pp. 533–538. DOI: 10.1109/ISEMANTIC.2018.8549751. [Online]. Available: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85060025914&doi=10.1109%2fISEMANTIC.2018.8549751&partnerID=40&md5=134ea0b43e1b964257aeaa6a53ba54a6>.
- [71] I. M. S. Bimantara and I. Widiartha, "Optimization of k-means clustering using particle swarm optimization algorithm for grouping traveler reviews data on tripadvisor sites," *Jurnal Ilmiah Kursor*, vol. 12, pp. 1–10, Jun. 2023. DOI: 10.21107/kursor.v12i01.269.

- [72] F. Sarac, H. Seker, M. Lisowski, and A. Timothy, "A hybrid two-level support vector machine-based method for churn analysis," Cited by: 2; All Open Access, Green Open Access, 2021, pp. 77–81. DOI: 10.1145/3481646.3481665. [Online]. Available: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85120944115&doi=10.1145%2f3481646.3481665&partnerID=40&md5=ff0528b497375b6b647d0a58b170771e>.
- [73] N. Aravinda, R. A. Reddy, V. Bala Dhandayuthapani, J. P. R. Nayak, and A. Hussein, "Customer churn prediction in the telecom sector with machine learning and adaptive k-means cluster using imbalance data," Cited by: 1, 2024. DOI: 10.1109/ICDCOT61034.2024.10515832. [Online]. Available: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85193230833&doi=10.1109%2fICDCOT61034.2024.10515832&partnerID=40&md5=ec24c751489621c898168622d567b588>.

APPENDIX A

Literature Review

TABLE A.1. Overview of articles on churn prediction using supervised learning in financial data in telecommunications

Articles	Techniques	Metrics
Towards Real-time Customer Experience Prediction for Telecommunication Operators [43]	Restricted Random Forest (RRF)	Transaction-level: F1-score: 85% improvement; User-level: F1-score: 57% improvement
A Customer Churn Prediction Model using XGBoost for the Telecommunication Industry in Nepal [41]	XGBoost	Native Dataset: Accuracy: 97%, F1-score: 88%; Public Dataset: Accuracy: 96.25%, F1-score: 86.34%
Churn Prediction in Telecommunication using Logistic Regression and Logit Boost [2]	Logistic regression, Logit Boost	Logit Boost: MAE: 22.61%, RMSE: 33.57%, Accuracy: 85.18%, Precision: 80.2%, Recall: 85.2%, F1-score: 80.6%; Logistic Regression: MAE: 22.37%, RMSE: 33.52%, Accuracy: 85.24%, Precision: 80.6%, Recall: 85.2%, F1-score: 81%
Churn prediction using optimized deep learning classifier on huge telecom data [39]	Deep convolutional neural network using Firefly-Spider Optimization	Dice coefficient: 94.61%; Jaccard coefficient: 94.8%; Accuracy: 94.76%
Comparative Study of Classification Algorithms for Customer Decisions on Telecommunication Products Using Supervised Learning [42]	Logistic regression, SVM, DT, RF	Logistic Regression: Accuracy: 80%, F1-score: 80%; SVM: Accuracy: 82%, F1-score: 82%; Decision Tree: Accuracy: 99%, F1-score: 99%; Random Forest: Accuracy: 99%, F1-score: 99%
Customer churn prediction in the telecommunication sector using a rough set approach[44]	RST	GA: Coverage: 100%, Accuracy: 98.2%, F1-score: 92.5%; EA: Coverage: 100%, Accuracy: 92.6%, F1-score: 72.6%; LA: Coverage: 66.8%, Accuracy: 99.3%, F1-score: 69.8%; CA: Coverage: 64%, Accuracy: 87.8%, F1-score: 48.7%
Customer churn prediction in telecom sector using machine learning techniques [5]	DT, RF	DT: Accuracy: 93.85%, F1-score: 93%; RF: Accuracy: 99.09%, F1-score: 99%
Customer churn prediction in telecom using machine learning in big data platform [1]	XGBoost, DT, RF, GBM	GBM: AUC: 90.8%; DT: AUC: 83%; RF: AUC: 87.76%; XGBoost: AUC: 93.3%
Explaining customer churn prediction in telecom industry using tabular machine learning models [40]	GBM	Accuracy: 81%, Precision: 67%, Recall: 55%, F1-score: 60%, ROC-score: 86%, PR-score: 68%
Modeling approaches for Silent Attrition prediction in Payment networks [45]	Logistic regression, SVM, RF, Ensemble	Logistic Regression: F1-score: 73.9%; SVM: F1-score: 77.7%; RF: F1-score: 80%; Ensemble: F1-score: 85%
Risk assessment of customer churn in telco using FCLCNN-LSTM model [6]	FCLCNN-LSTM	Accuracy: 94.15%, Precision: 84.59%, Recall: 51.23%, F1-score: 50.93%, AUC: 51.23%

TABLE A.2. Overview of articles on fraud detection using supervised learning in financial data in telecommunications

Articles	Techniques	Metrics
BTG: A Bridge to Graph machine learning in telecommunications fraud Detection [47]	BTG	AUC: 92.45%; Recall: 85.21%; Precision: 89.90%; Accuracy: 89.28%; F1: 87.01%
ConvNets for Fraud Detection analysis [46]	DCNN; SVM; RF; Gradient Boosting Classifier	- DCNN: Accuracy: 82%; SVM: Accuracy: 77%; RF: Accuracy: 72%; Gradient Boosting Classifier: Accuracy: 79%
Fraud detection system: A survey [38]	DT; ANN; k-NN; SVM; Logistic Regression	No numerical metrics presented. Superior results achieved with ANN.

TABLE A.3. Overview of articles on profitability prediction using supervised learning in financial data in telecommunications

Articles	Techniques	Metrics
Mining Retail Telecommunication Data to Predict Profitability [7]	BTG	AUC: 92.45%; Recall: 85.21%; Precision: 89.90%; Accuracy: 89.28%; F1: 87.01%

TABLE A.4. Overview of unsupervised learning applications in financial data for telecommunications

Articles	Techniques	Metrics
"An Enhanced Spectral Clustering Algorithm with S-Distance" [49]	Spectral Clustering with S-Distance	Jaccard Index: 92.45%, F-score: 87.01%, Recall: 85.21%, Precision: 89.90%, Accuracy: 89.28%
"Cluster & Rough Set Theory Based Approach to Find the Reason for Customer Churn" [32]	Hierarchical Clustering + Rough Set Theory	No metrics. The technique was very effective in identifying churn reasons
"Prediction of customer churn in telecom sector using clustering technique" [9]	K-means	No metrics. K-means was effective in identifying customer segments with higher churn probability
"Scalable Clustering and Applications" [50]	Spectral Clustering with Nyström and Lanczos methods	No metrics. The approach reduces computational complexity, making it efficient for big data
"Social Network Analysis and Churn Prediction in Telecommunications Using Graph Theory" [51]	Wards Minimum-Variance Clustering + Graph Theory	Top decile lift: 2.5-10 (depending on the model used); clustering was effective in identifying influencers and churn propagation
"Telecommunication Analytics Based on Customer Segmentation Using Unsupervised Algorithms" [10]	K-means, K-medoids, FCM, GK, Hierarchical Clustering	K-means: AR: 53.41%, RI: 50.50%, FMI: 61.09% ; K-Medoids: AR: 53.30%, RI: 50.48%, FMI: 61.03% ; FCM: AR: 55.15%, RI: 50.15%, FMI: 59.99% ; GK: AR: 50.24%, RI: 49.99%, FMI: 55.20%

TABLE A.5. Overview of supervised and unsupervised learning techniques combined in financial data for telecommunications

Articles	Techniques	Metrics
"A Case Study for the Churn Prediction in Turksat Internet Service Subscription" [33]	k-means + RPART, SVM	SVM performed 3% better than RPART when using distances to cluster centers. RPART achieved better performance with binary categorical data.
"A Hybrid Two-Level Support Vector Machine-Based Method for Churn Analysis" [72]	Multi-Cluster Feature Selection (MCFS) + SVM	Accuracy: 81.5%; AUC: 85.6
"An ensemble based approach using a combination of clustering and classification algorithms" [23]	K-means, K-medoids, X-means, Random Clustering + DT, RF, Gradient Boosting Trees, k-NN, Deep Learning, Naïve Bayes, Ensemble Methods (Voting, Bagging, Stacking, AdaBoost)	Best result is using k-medoids + Gradient Boosting Trees, DT, Deep Learning, Adaboost. Precision: 94.7%, Precision: 87.04%, Recall: 75.1%, F1-score: 80.63%
"Churn Analysis in Telecommunication Industry" [24]	K-means + RF, SVM	RF: Accuracy: 95.44%, F1-score: 97% (not churn) / 84% (churn). SVM: Accuracy: 93.68%, F1-score: 96% (not churn) / 78% (churn).
"Customer Churn Prediction in the Telecom Sector with Machine Learning and Adaptive k-Means Cluster using Imbalance Data" [73]	Adaptive k-means + SVM	Accuracy: 95.70%; Precision: 56.01%; Recall: 55.05%; F1-score: 96.05%
"Customer Churn Modeling in Telecommunication Using a Novel Multi-Objective Evolutionary Clustering-Based Ensemble Learning" [55]	MOEEC (MOEEC-1, MOEEC-2)	MOEEC-1: Accuracy: 97.30%, Recall: 88.66%, Precision: 93.48%, F1-score: 91.01%, AUC: 93.76%. MOEEC-2: Accuracy: 96.35%, Recall: 92.78%, Precision: 84.9%, F1-score: 88.67% AUC: 94.89%.
"Hybrid PPFCM-ANN model: an efficient system for customer churn prediction through probabilistic possibilistic fuzzy clustering and artificial neural network" [56]	PPFCM-ANNLM (Probabilistic Possibilistic Fuzzy C-Means Artificial Neural Network Levenberg-Marquardt)	CA (Classification accuracy): 94,62%
"Machine Learning-Based Predictive Analytics for Customer Churn in the Telecom Industry" [53]	PCA + Logistic regression, SVM, RF, XGBoost, Gradient Boosted Decision Trees	Best result achieved with RF. Accuracy: 99.97%, F1-score: 99.8%
"Predicting customer churn in the telecom industry using data analytics" [54]	K-Means, Logistic Regression, RF	Logistic Regression: Accuracy: 85,9%. Random Forest: 94.4%. K-Means: No metrics, the more the number of clusters, the higher the accuracy
"Profiling Television Watching Behavior Using Bayesian Hierarchical Joint Models for Time-to-Event and Count Data" [57]	Bayesian hierarchical joint model + SVM	Accuracy: 92%, True Positive rates: 100%, False Positive Rates: 14%
"Using big data analysis to retain customers for telecom industry" [52]	XGBoost + K-Means	XGBoost: Accuracy: 85.1%. K-Means: 3 clusters, and churn rate reduced by 93%

APPENDIX B

Data Analysis

B.1. Variables description

TABLE B.1. Datasets variables description

Variables	Description
'CAT _CONTA _CON- TRATO'	Company or entity associated with the customer's contract
'Sub-Segmento'	The company's area to which the customer belongs (for example, consumer, business, public, wholesale)
'Estrangeiro'	Binary variable indicating whether the customer is foreign or not
'Tipo_Cliente'	Binary variable indicating whether the customer is national or international. Complementary to the 'Estrangeiro' variable
'Saldo Total'	Total outstanding debt of the customer
'Não Vencido'	Debt amount that has not yet matured
'0 a 30'	Indicates whether the customers debt is between 0 and 30 days overdue
'30 a 60'	Indicates whether the customers debt is between 30 and 60 days overdue
'60 a 90'	Indicates whether the customers debt is between 60 and 90 days overdue
'Dissolução'	Date of business or entity closure
'Insolvência'	Customer unable to pay their debts
'Insolvência - PER'	Customer who is insolvent but still has the possibility of repayment
'Data Criação NIF'	Date when the customer signed a contract with the company
'Data Criação Conta'	Date when the customer joined the company
'Data Fim Conta'	Date when the customer left the company
'SAS - Ciclo'	Billing cycles that occur throughout the month
'VEP'	Indicates whether there are installment sales for equipment
'VEP - N° Prestações'	Number of installments under 'VEP'
'VEP - N° Prestações Venc'	Number of overdue installments (already paid by the customer)
'PPP'	Indicates whether there is an installment payment plan for out-standing debts
'PPP - N° Prestações'	Number of installments under 'PPP'
'PPP - Valor do Plano'	Total amount of the 'PPP'
'Acompanhamento'	Debt manager responsible for monitoring the customers debt
'Venda Créditos'	Indicates whether this contract account has been written off as uncollectable
'SEPA'	Indicates whether the customer pays via direct debit
'Serviço_Estado'	Indicates whether the customers account is active, deactivated, or interrupted. Complementary to the 'Tipo_Conta' variable
'Tipo_Conta'	Indicates whether the customers account is CN (normal account), CT (litigious account), or CA (suspended account)

TABLE B.2. Final dataset variables description

Variable	Type	Description
'Estrangeiro'	Binary	Indicates whether the customer is foreign
'Saldo Total'	Numeric	Total outstanding debt of the customer
'30 a 60'	Numeric	Debt overdue between 30 and 60 days
'Dissolução'	Binary	Indicates if the company has been dissolved
'Insolvência'	Binary	Indicates if the customer is insolvent
'Insolvência - PER'	Binary	Indicates if the customer is in insolvency proceedings
'VEP'	Binary	Indicates whether the customer is in the VEP program
'VEP - Nº Prestações'	Numeric	Number of installment payments under VEP
'VEP - Nº Prestações Venc'	Numeric	Number of overdue installment payments under VEP
'PPP'	Binary	Indicates whether the customer is in a PPP plan
'PPP - Nº Prestações'	Numeric	Number of installment payments under PPP plan
'PPP - Valor do Plano'	Numeric	Total amount of the PPP installment plan
'Acompanhamento'	Binary	Indicates if the customer is being monitored
'Venda Créditos'	Binary	Indicates if the customer has credit sales
'SEPA'	Binary	Indicates whether the customer pays by direct debit
'Longevidade'	Numeric	Duration in days since the customer registered
'Data Criação NIF (Diferença Dias)'	Numeric	Days since the creation of the customer's tax ID
'CAT_CONTA __CONTRATO (various)'	Binary	One-hot encoded variables representing different contract entities
'Sub_Segmento (various)'	Binary	One-hot encoded variables representing different company areas
'Serviço_Estado (various)'	Binary	One-hot encoded variables indicating service status (for example, deactivated, interrupted)
'SAS - Ciclo (various)'	Binary	One-hot encoded variables representing different billing cycle days
'Cumprimento da dívida'	Binary	Target variable (1 = Paid, 0 = Not paid)

B.2. Variable '60 a 90' relevance

TABLE B.3. Logistic regression - Confusion Matrix for Train Set ('with 60 a 90')

	Predicted 0	Predicted 1
True 0	396833	38249
True 1	12304	422789

TABLE B.4. Logistic regression - Classification Report for Train Set ('with 60 a 90')

Class	Precision	Recall	F1-Score
0	0.97	0.91	0.94
1	0.92	0.97	0.94
Accuracy	-	-	0.94
Macro avg	0.94	0.94	0.94
Weighted avg	0.94	0.94	0.94

TABLE B.5. Logistic regression - Confusion Matrix for Test Set ('with 60 a 90')

	Predicted 0	Predicted 1
True 0	99237	9526
True 1	3156	105634

TABLE B.6. Logistic regression - Classification Report for Test Set ('with 60 a 90')

Class	Precision	Recall	F1-Score
0	0.97	0.91	0.94
1	0.92	0.97	0.94
Accuracy	-	-	0.94
Macro avg	0.94	0.94	0.94
Weighted avg	0.94	0.94	0.94

TABLE B.7. Random Forest - Confusion Matrix for Train Set ('with 60 a 90')

	Predicted 0	Predicted 1
True 0	426133	8920
True 1	4060	431010

TABLE B.8. Random Forest - Classification Report for Train Set ('with 60 a 90')

Class	Precision	Recall	F1-Score
0	0.99	0.98	0.98
1	0.98	0.99	0.98
Accuracy	-	-	0.99
Macro avg	0.99	0.99	0.99
Weighted avg	0.99	0.99	0.99

TABLE B.9. Random Forest - Confusion Matrix for Test Set ('with 60 a 90')

	Predicted 0	Predicted 1
True 0	106607	2160
True 1	1084	107686

TABLE B.10. Random Forest - Classification Report for Test Set ('with 60 a 90')

Class	Precision	Recall	F1-Score
0	0.99	0.98	0.98
1	0.98	0.99	0.98
Accuracy	-	-	0.99
Macro avg	0.99	0.99	0.99
Weighted avg	0.99	0.99	0.99

TABLE B.11. k-NN - Confusion Matrix for Train Set ('with 60 a 90')

	Predicted 0	Predicted 1
True 0	422654	12597
True 1	11583	423483

TABLE B.12. k-NN - Classification Report for Train Set ('with 60 a 90')

Class	Precision	Recall	F1-Score
0	0.97	0.97	0.97
1	0.97	0.97	0.97
Accuracy	-	-	0.97
Macro avg	0.97	0.97	0.97
Weighted avg	0.97	0.97	0.97

TABLE B.13. k-NN - Confusion Matrix for Test Set ('with 60 a 90')

	Predicted 0	Predicted 1
True 0	104878	3889
True 1	3890	104876

TABLE B.14. k-NN - Classification Report for Test Set ('with 60 a 90')

Class	Precision	Recall	F1-Score
0	0.96	0.96	0.96
1	0.96	0.96	0.96
Accuracy	-	-	0.96
Macro avg	0.96	0.96	0.96
Weighted avg	0.96	0.96	0.96

TABLE B.15. Neural Network - Confusion Matrix for Test Set ('with 60 a 90')

	Predicted 0	Predicted 1
True 0	105870	2897
True 1	3126	105641

TABLE B.16. Neural Network - Classification Report for Test Set ('with 60 a 90')

Class	Precision	Recall	F1-Score
0	0.97	0.97	0.97
1	0.97	0.97	0.97
Accuracy	-	-	0.97
Macro avg	0.97	0.97	0.97
Weighted avg	0.97	0.97	0.97

APPENDIX C

Modeling and Results

C.1. Correlations

TABLE C.1. Pairs of variables with correlations greater than 0.7 or less than -0.7

Variable 1	Variable 2	Correlation
'VEP - N° Prestações'	'VEP'	80.55%
'VEP'	'VEP - N° Prestações'	80.55%
'PPP - N° Prestações'	'PPP'	93.78%
'PPP - Valor do Plano'	'PPP'	94.18%
'PPP'	'PPP - N° Prestações'	93.78%
'PPP - Valor do Plano'	'PPP - N° Prestações'	98.06%
'PPP'	'PPP - Valor do Plano'	94.18%
'PPP - N° Prestações'	'PPP - Valor do Plano'	98.06%
'CAT_CONTA_CONTRATO_PTC'	'CAT_CONTA_CONTRATO_MEO'	-90.45%
'SAS - Ciclo_1.0'	'CAT_CONTA_CONTRATO_MEO'	87.42%
'CAT_CONTA_CONTRATO_MEO'	'CAT_CONTA_CONTRATO_PTC'	-90.45%
'SAS - Ciclo_1.0'	'CAT_CONTA_CONTRATO_PTC'	-94.44%
'Sub-Segmento_Outras_PTP_Consumo'	'CAT_CONTA_CONTRATO_SAP'	97.64%
'Serviço_Estado_Desativos'	'CAT_CONTA_CONTRATO_SAP'	-81.49%
'Sub-Segmento_DVE'	'Sub-Segmento_Consumo'	-85.90%
'Sub-Segmento_Consumo'	'Sub-Segmento_DVE'	-85.90%
'CAT_CONTA_CONTRATO_SAP'	'Sub-Segmento_Outras_PTP_Consumo'	97.64%
'Serviço_Estado_Desativos'	'Sub-Segmento_Outras_PTP_Consumo'	-81.01%
'CAT_CONTA_CONTRATO_SAP'	'Serviço_Estado_Desativos'	-81.49%
'Sub-Segmento_Outras_PTP_Consumo'	'Serviço_Estado_Desativos'	-81.01%
'CAT_CONTA_CONTRATO_MEO'	'SAS - Ciclo_1.0'	87.42%
'CAT_CONTA_CONTRATO_PTC'	'SAS - Ciclo_1.0'	-94.44%

TABLE C.2. Correlation of variables with the target variable

Variable	Correlation with Target Variable
'SEPA'	62.67%
'Data Criação NIF (Diferença Dias)'	17.81%
'SAS - Ciclo_13.0'	16.61%
'SAS - Ciclo_16.0'	15.77%
'SAS - Ciclo_22.0'	13.69%
'SAS - Ciclo_19.0'	13.52%
'SAS - Ciclo_25.0'	11.85%
'VEP'	5.09%
'Sub-Segmento_DVE'	4.54%
'SAS - Ciclo_28.0'	3.95%
'CAT_CONTA_CONTRATO_PTPPrime'	3.01%
'Sub-Segmento_Outras PTP - Consumo'	2.15%
'Venda Créditos'	1.64%
'Sub-Segmento_DCA'	0.97%
'CAT_CONTA_CONTRATO_PTCom'	0.93%
'Sub-Segmento_DCG'	0.90%
'Sub-Segmento_Outras PTP - Outros'	0.10%
'Sub-Segmento_Instaladores'	0.00%
'Sub-Segmento_Outras PTP - DCA'	-0.10%
'Sub-Segmento_Outras PTP - DVE'	-0.15%
'Sub-Segmento_Outras PTP - DCM'	-0.15%
'Sub-Segmento_Outras PTP - IntraGrupo'	-0.15%
'Sub-Segmento_DCM'	-0.23%
'Sub-Segmento_Outras PTP - DCG'	-0.30%
'CAT_CONTA_CONTRATO_SAPSDGRUPO'	-0.32%
'Sub-Segmento_Wholesale Nacional'	-0.32%
'CAT_CONTA_CONTRATO_PrepaidRoaming'	-0.34%
'Sub-Segmento_Outras PTP - DCT'	-0.37%
'Sub-Segmento_Outros'	-0.68%
'Sub-Segmento_IntraGrupo'	-0.82%
'Sub-Segmento_SAPO'	-0.99%
'Insolvência - PER'	-1.39%

Table C.2 continued from previous page

Variable	Correlation with Target Variable
'Sub-Segmento_PP'	-1.41%
'Sub-Segmento_Wholesale Internacional'	-1.48%
'Estrangeiro'	-1.85%
'Sub-Segmento_DCT'	-2.28%
'CAT_CONTA_CONTRATO_PTC/VozEmpresarial'	-2.67%
'CAT_CONTA_CONTRATO_SAPSD'	-3.16%
'Acompanhamento'	-3.40%
'SAS - Ciclo_7.0'	-3.91%
'SAS - Ciclo_4.0'	-4.19%
'SAS - Ciclo_10.0'	-5.49%
'Serviço_Estado_Interrupção'	-6.36%
'Dissolução'	-6.93%
'Insolvência'	-9.74%
'VEP - Nř Prestações Venc'	-14.77%
'PPP'	-15.76%
'Saldo Total'	-32.85%
'30 a 60'	-33.23%
'Longevidade'	-35.21%
'CAT_CONTA_CONTRATO_MEO'	-44.18%
'Serviço_Estado_Desativos'	-69.26%

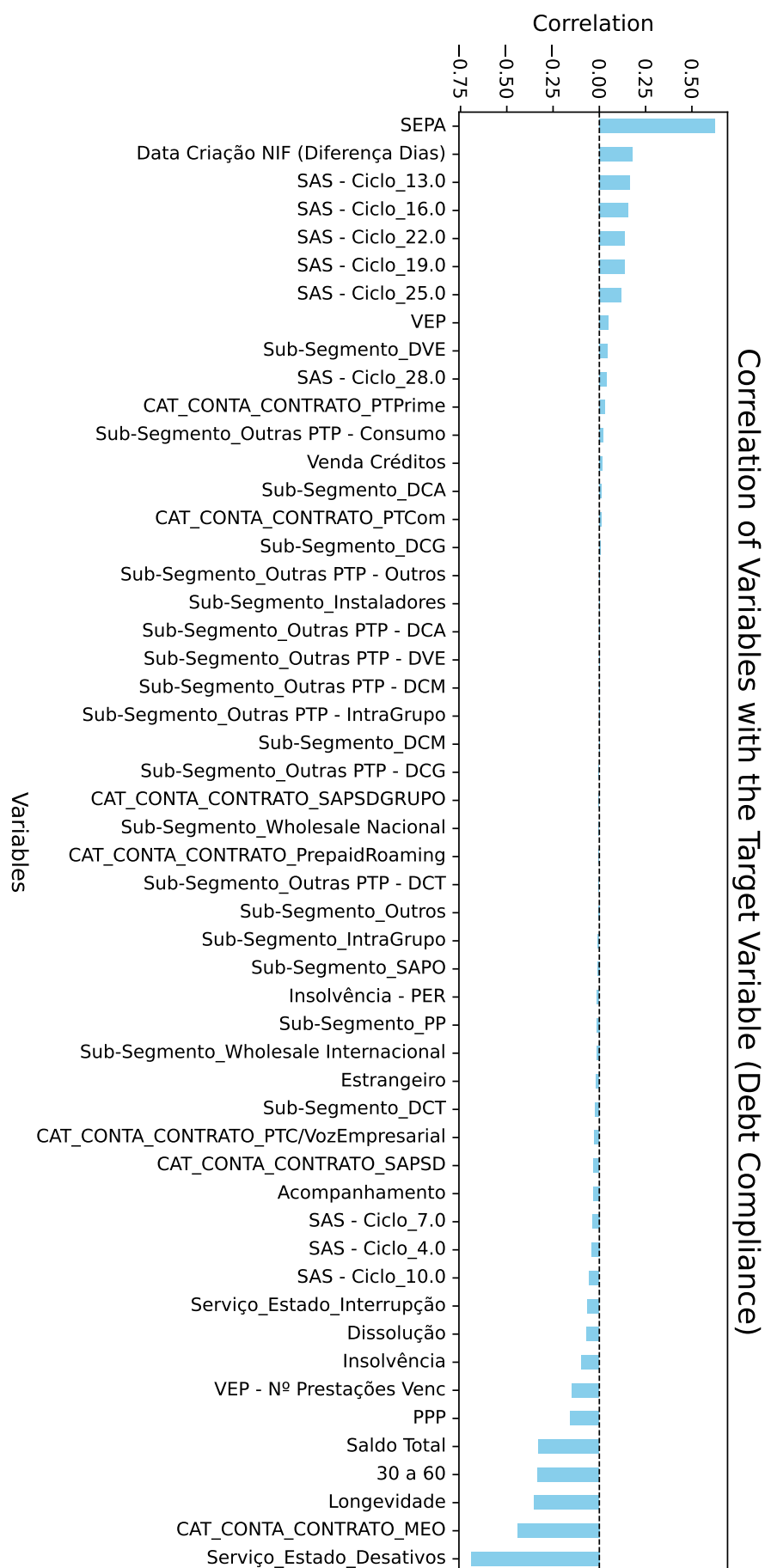


FIGURE C.1. Correlation between the variables and the target variable

C.2. Supervised Learning

TABLE C.3. Logistic regression - Confusion Matrix for Train Set

	Predicted 0	Predicted 1
True 0	396857	38209
True 1	12324	422742

TABLE C.4. Logistic regression - Classification Report for Train Set

Class	Precision	Recall	F1-Score
0	0.97	0.91	0.94
1	0.92	0.97	0.94
Accuracy	-	-	0.94
Macro avg	0.94	0.94	0.94
Weighted avg	0.94	0.94	0.94

TABLE C.5. Logistic regression - Confusion Matrix for Test Set

	Predicted 0	Predicted 1
True 0	99231	9536
True 1	3148	105619

TABLE C.6. Logistic regression - Classification Report for Test Set

Class	Precision	Recall	F1-Score
0	0.97	0.91	0.94
1	0.92	0.97	0.94
Accuracy	-	-	0.94
Macro avg	0.94	0.94	0.94
Weighted avg	0.94	0.94	0.94

TABLE C.7. Random Forest - Confusion Matrix for Train Set

	Predicted 0	Predicted 1
True 0	426139	8927
True 1	4059	431007

TABLE C.8. Random Forest - Classification Report for Train Set

Class	Precision	Recall	F1-Score
0	0.99	0.98	0.98
1	0.98	0.99	0.99
Accuracy	-	-	0.99
Macro avg	0.99	0.99	0.99
Weighted avg	0.99	0.99	0.99

TABLE C.9. Random Forest - Confusion Matrix for Test Set

	Predicted 0	Predicted 1
True 0	106602	2165
True 1	1079	107688

TABLE C.10. Random Forest - Classification Report for Test Set

Class	Precision	Recall	F1-Score
0	0.99	0.98	0.99
1	0.98	0.99	0.99
Accuracy	-	-	0.99
Macro avg	0.99	0.99	0.99
Weighted avg	0.99	0.99	0.99

TABLE C.11. k-NN - Confusion Matrix for Train Set

	Predicted 0	Predicted 1
True 0	422567	12599
True 1	11581	423485

TABLE C.12. k-NN - Classification Report for Train Set

Class	Precision	Recall	F1-Score
0	0.97	0.97	0.97
1	0.97	0.97	0.97
Accuracy	-	-	0.97
Macro avg	0.97	0.97	0.97
Weighted avg	0.97	0.97	0.97

TABLE C.13. k-NN - Confusion Matrix for Test Set

	Predicted 0	Predicted 1
True 0	104887	3880
True 1	3891	104876

TABLE C.14. k-NN - Classification Report for Test Set

Class	Precision	Recall	F1-Score
0	0.96	0.96	0.96
1	0.96	0.96	0.96
Accuracy	-	-	0.96
Macro avg	0.96	0.96	0.96
Weighted avg	0.96	0.96	0.96

TABLE C.15. Neural Network - Confusion Matrix for Test Set

	Predicted 0	Predicted 1
True 0	105704	3063
True 1	2766	106001

TABLE C.16. Neural Network - Classification Report for Test Set

Class	Precision	Recall	F1-Score
0	0.97	0.97	0.97
1	0.97	0.97	0.97
Accuracy	-	-	0.97
Macro avg	0.97	0.97	0.97
Weighted avg	0.97	0.97	0.97

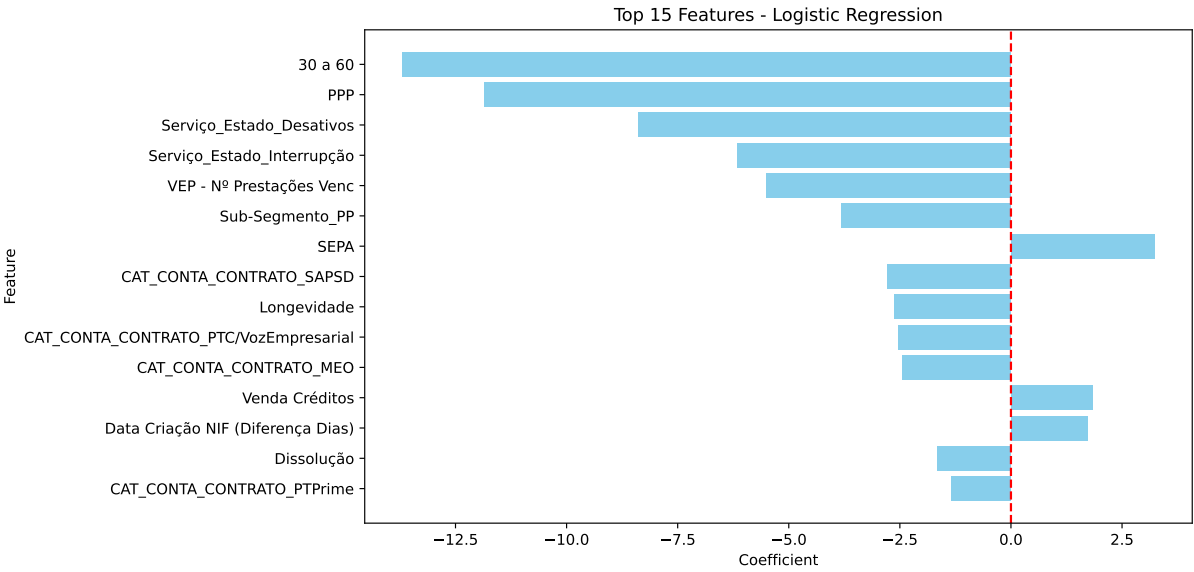


FIGURE C.2. Top15 feature importance of logistic regression model

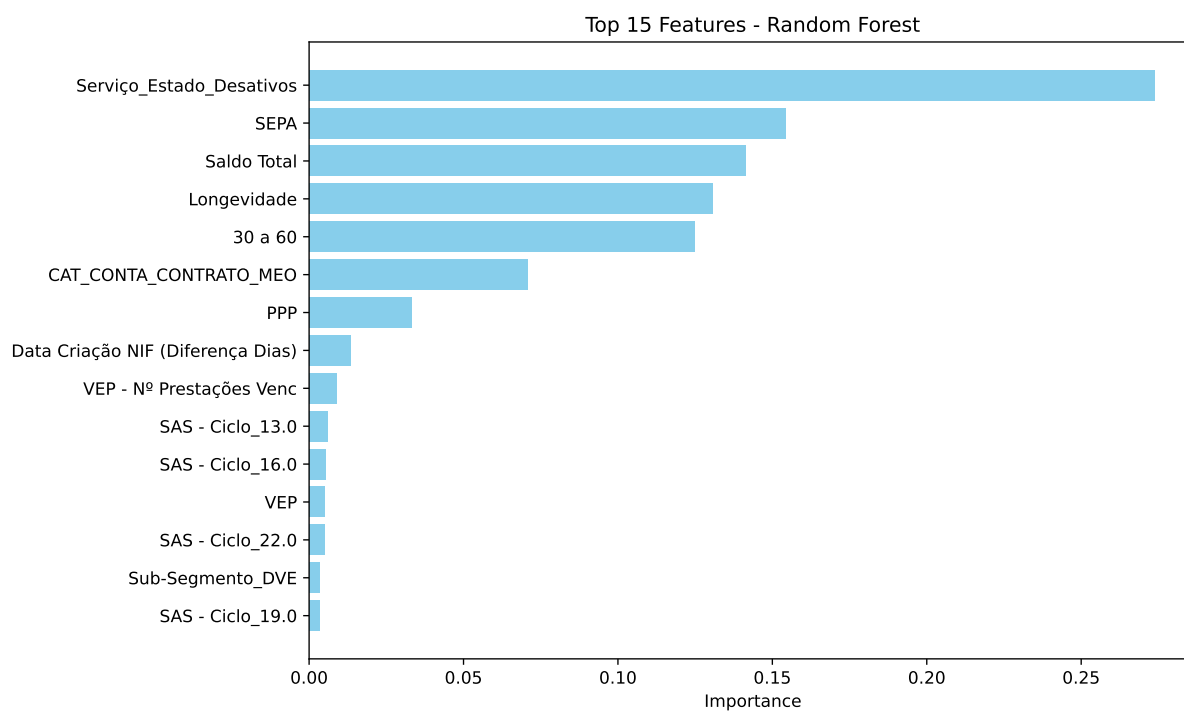


FIGURE C.3. Top15 feature importance of random forest model

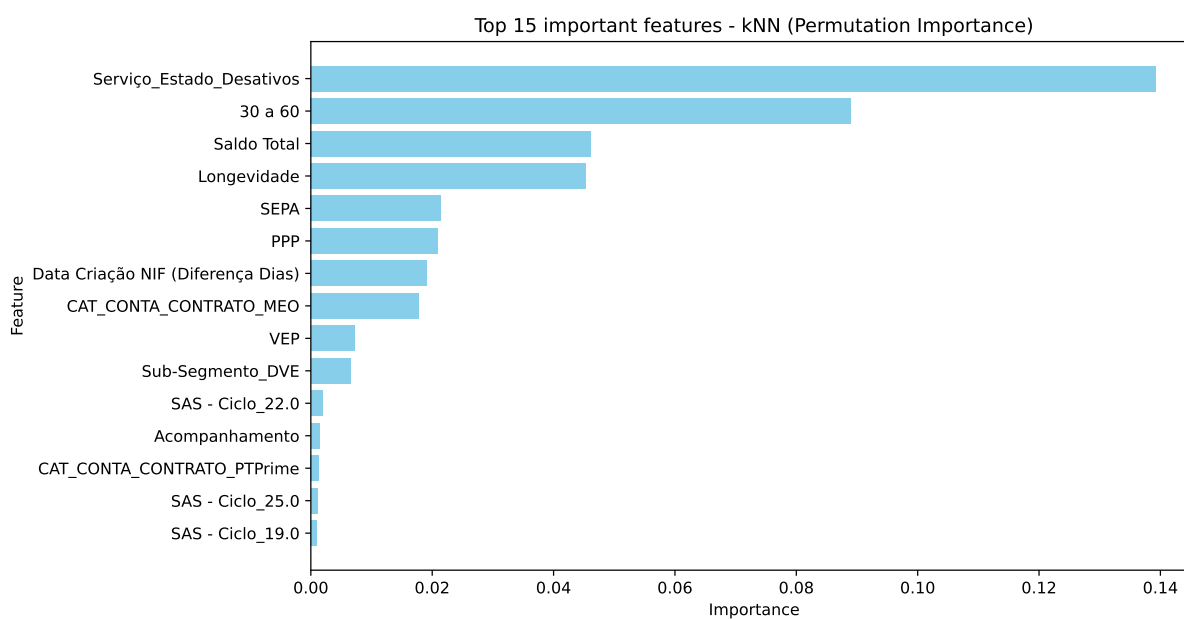


FIGURE C.4. Top15 feature importance of k-NN model

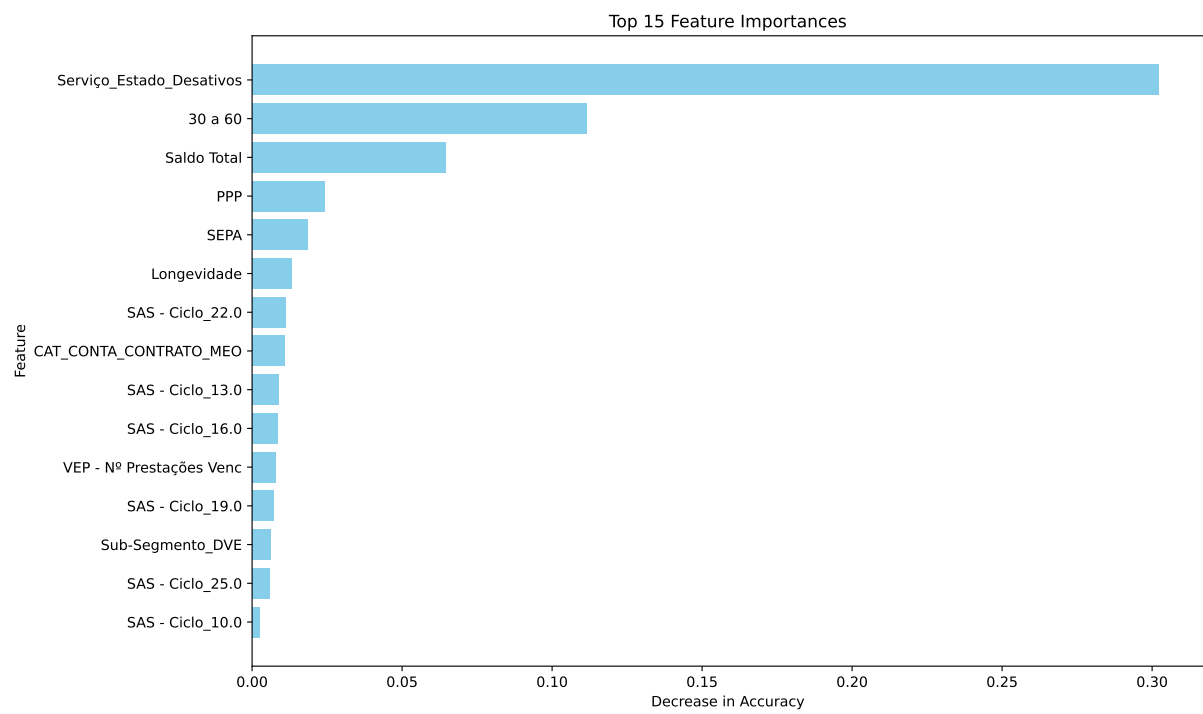


FIGURE C.5. Top15 feature importance of neural network model

C.3. Unsupervised learning

TABLE C.17. Eigenvalues of Principal Components

Component	Eigenvalue
Component 1	0.228
Component 2	0.132
Component 3	0.083
Component 4	0.053
Component 5	0.051
Component 6	0.038
Component 7	0.018
Component 8	0.016
Component 9	0.011
Component 10	0.009
Component 11	0.009
Component 12	0.006
Component 13	0.006
Component 14	0.002
Component 15	0.001
Component 16	0.001
Component 17	0.000
Component 18	0.000
Component 19	0.000

TABLE C.18. Cumulative explained variance by component

Component	Cumulative Explained Variance
Component 1	39.17%
Component 2	55.97%
Component 3	68.88%
Component 4	77.73%
Component 5	83.70%
Component 6	88.31%
Component 7	91.79%
Component 8	93.79%
Component 9	95.52%
Component 10	96.79%
Component 11	97.82%
Component 12	98.49%
Component 13	99.13%
Component 14	99.57%
Component 15	99.74%
Component 16	99.84%
Component 17	99.92%
Component 18	99.96%
Component 19	100.00%

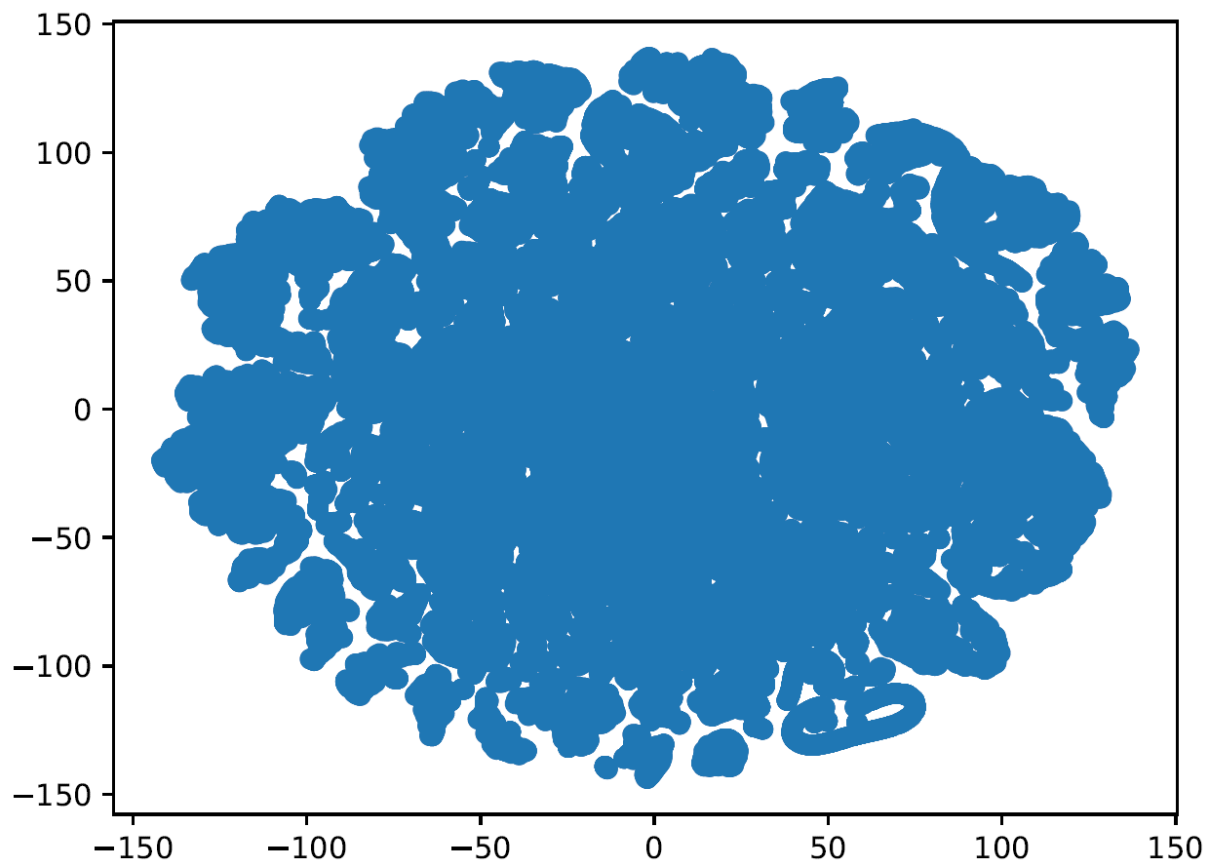


FIGURE C.6. t-SNE - Scatter plot with 100,000 observations

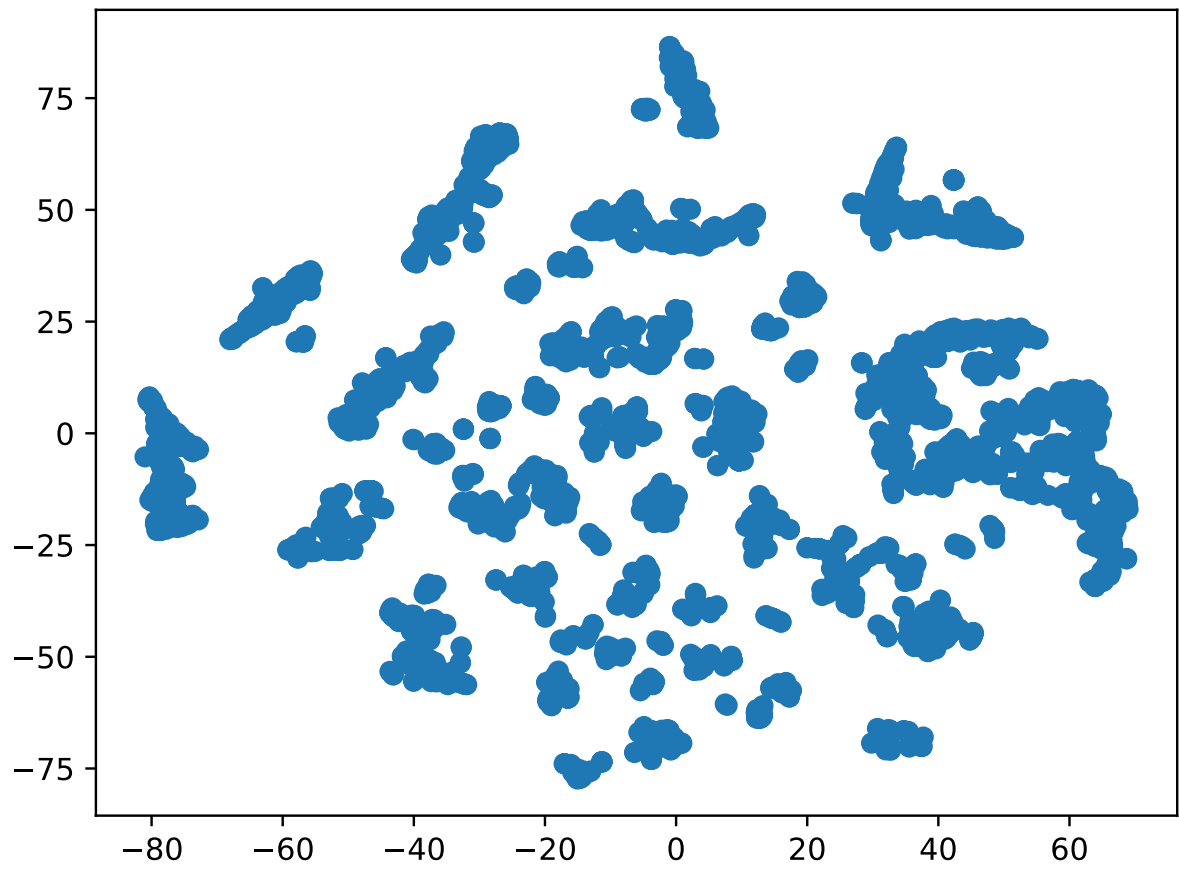


FIGURE C.7. t-SNE - Scatter plot with 5,000 observations

TABLE C.19. PC loadings

Variável	PC1	PC2	PC3
Cliente Estrangeiro	0.0004	-0.0001	-0.0014
Saldo Total	0.1887	0.4795	-0.2170
30 a 60 Dias Vencido	0.0106	-0.0287	-0.1259
Dissolução	0.0108	0.0020	0.0071
Insolvência	0.0287	0.0121	0.0106
Insolvência - PER	0.0007	0.0001	-0.0002
VEP	-0.0564	0.5453	-0.4853
VEP - Nř Prestações	-0.0142	0.2163	-0.2059
VEP - Nř Prestações Vencidas	0.0339	0.1303	-0.1176
PPP	-0.0070	-0.0046	-0.2007
PPP - Nř Prestações	-0.0059	0.0004	-0.1671
PPP - Valor do Plano	-0.0059	0.0013	-0.1692
Monitorização	-0.0018	-0.0180	-0.0233
Venda a Crédito	0.0000	-0.0011	-0.0003
SEPA (Pagamento Débito Direto)	-0.6282	0.5031	0.5543
Longevidade	0.2536	-0.0496	0.2216
Data Criação NIF (Diferença Dias)	-0.0427	-0.1154	0.0429
Estado do Serviço - Desativado	0.7056	0.3721	0.4337
Estado do Serviço - Interrupção	-0.0005	0.0140	-0.0389

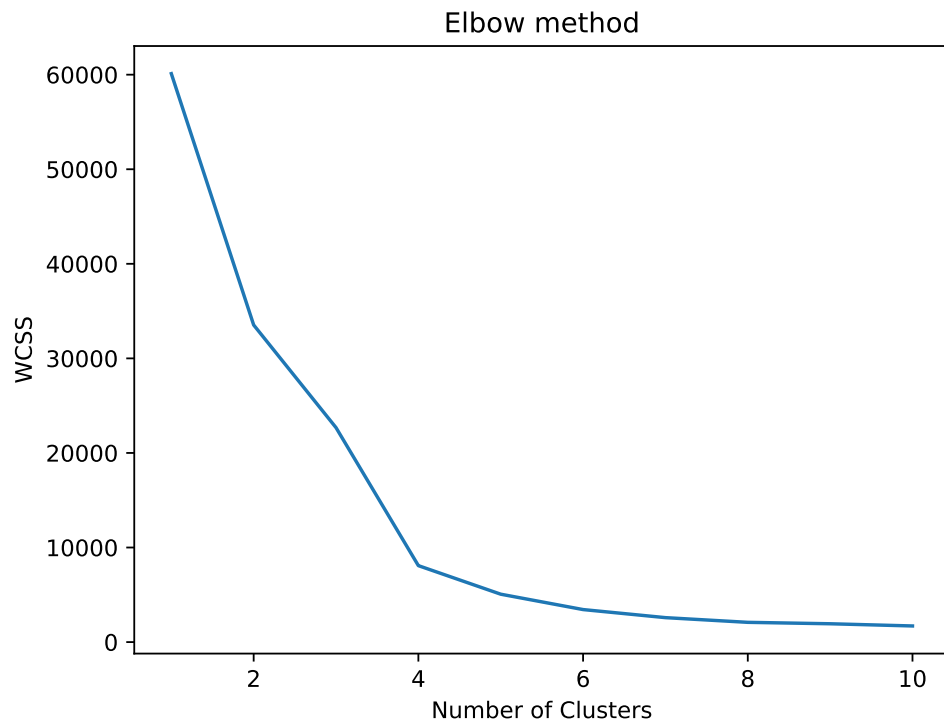


FIGURE C.8. Elbow method

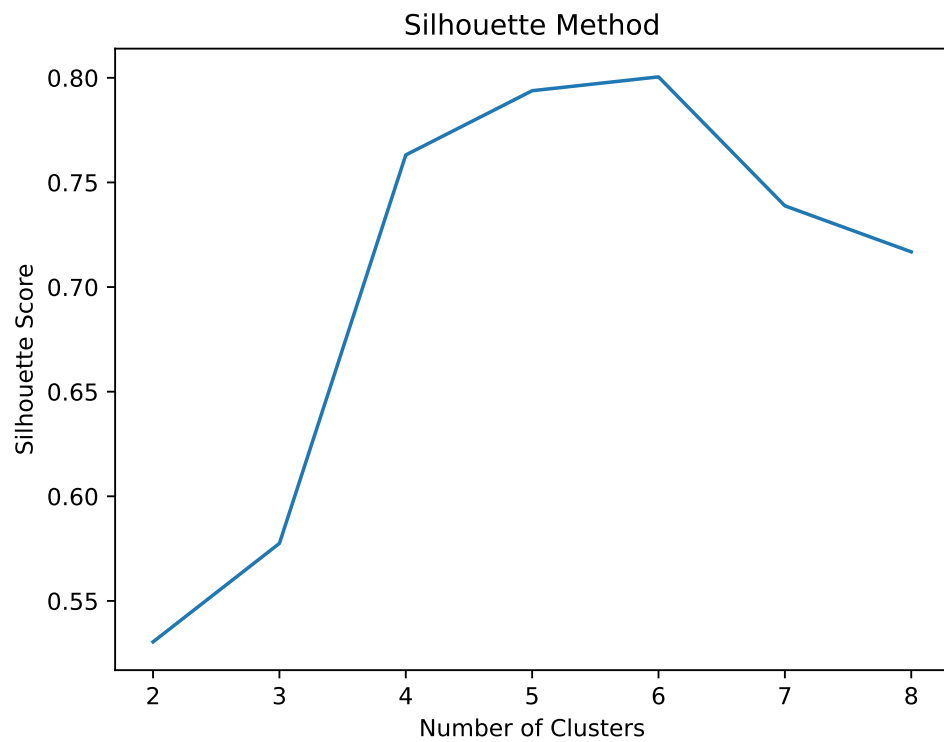


FIGURE C.9. Silhouette method

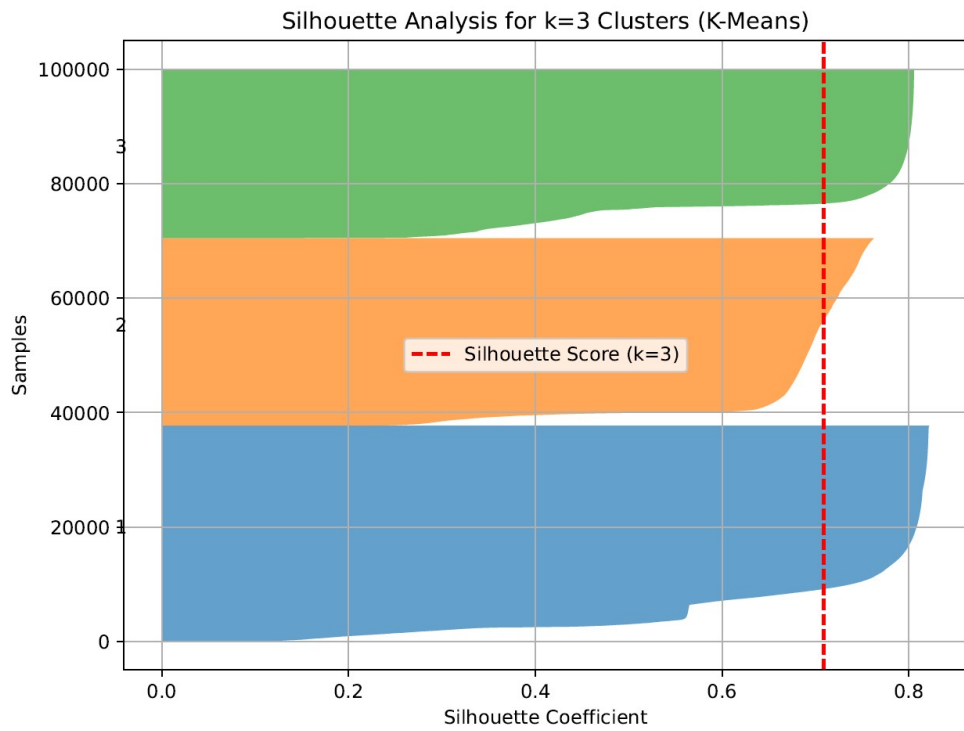


FIGURE C.10. Silhouette Plot for k-means with k=3 Clusters

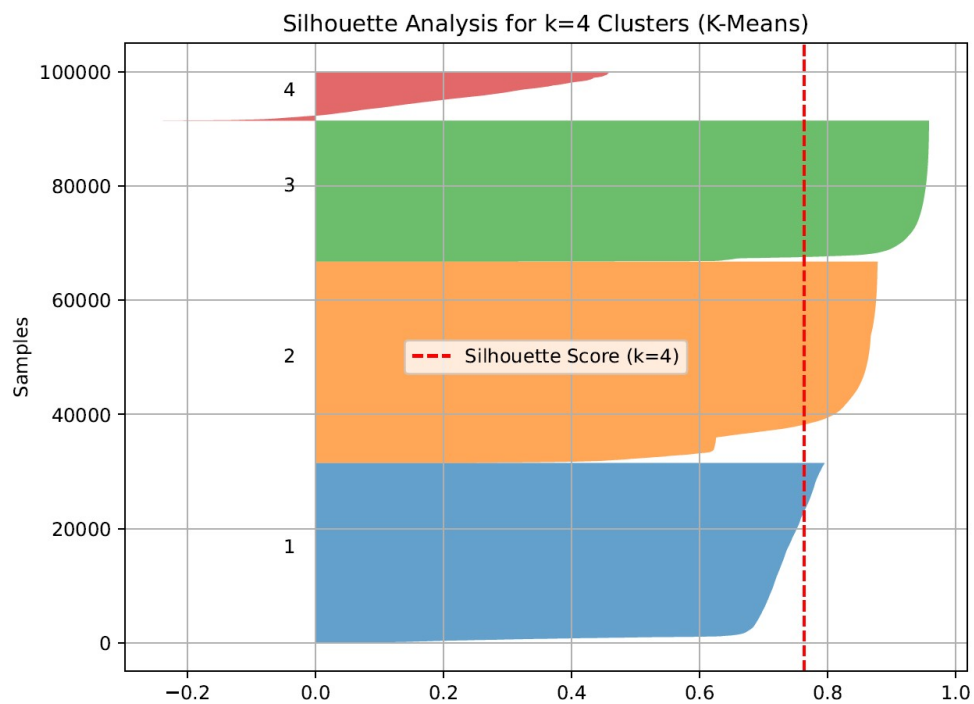


FIGURE C.11. Silhouette Plot for k-means with k=4 Clusters

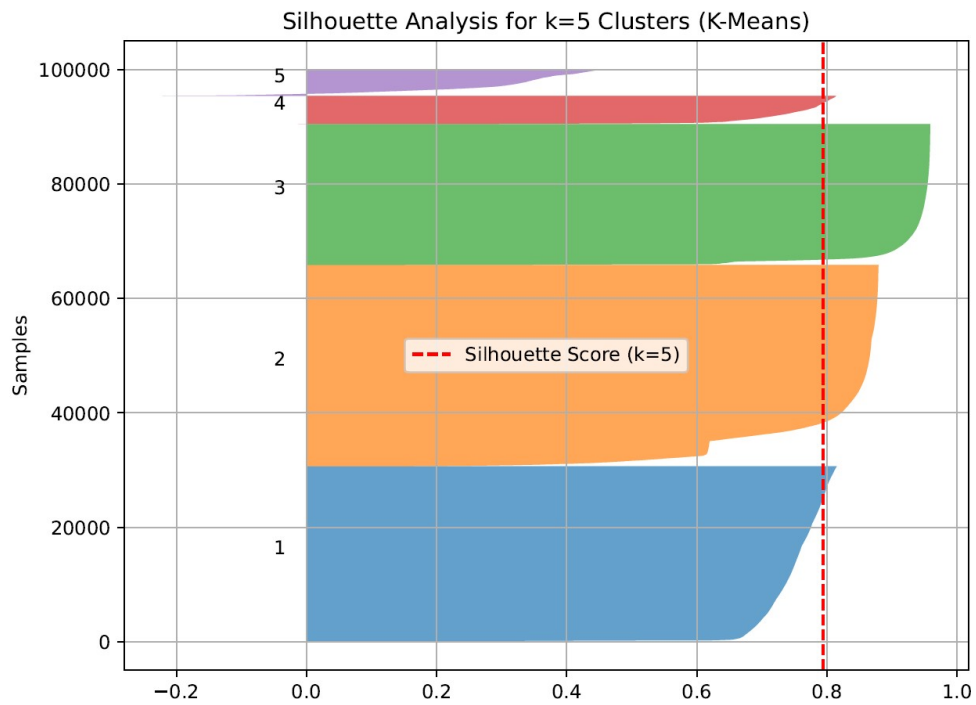


FIGURE C.12. Silhouette Plot for k-means with k=5 Clusters

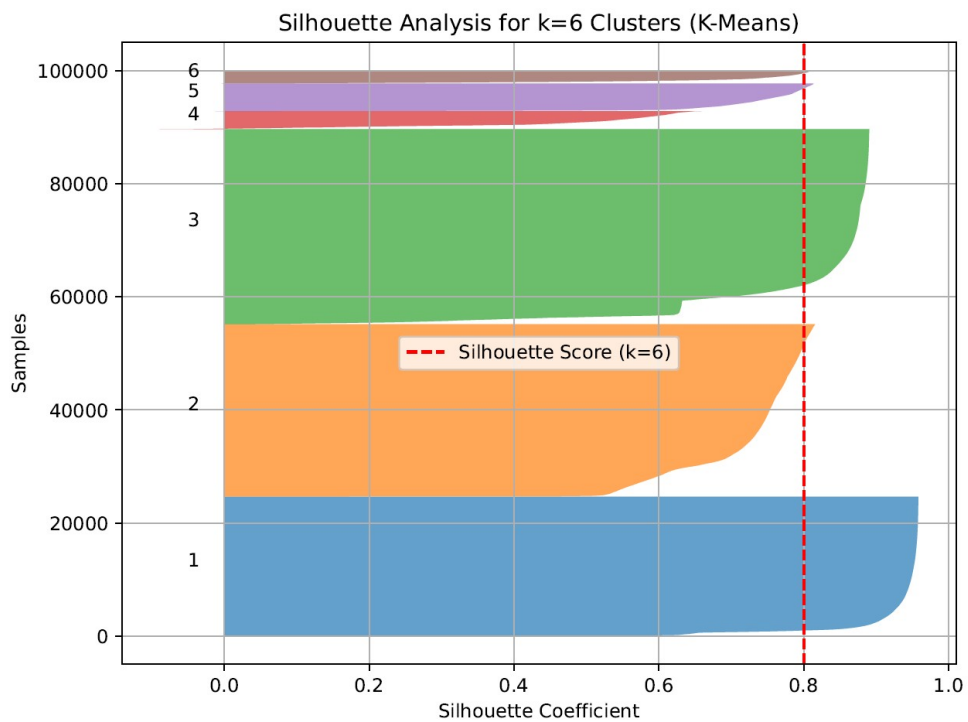


FIGURE C.13. Silhouette Plot for k-means with k=6 Clusters

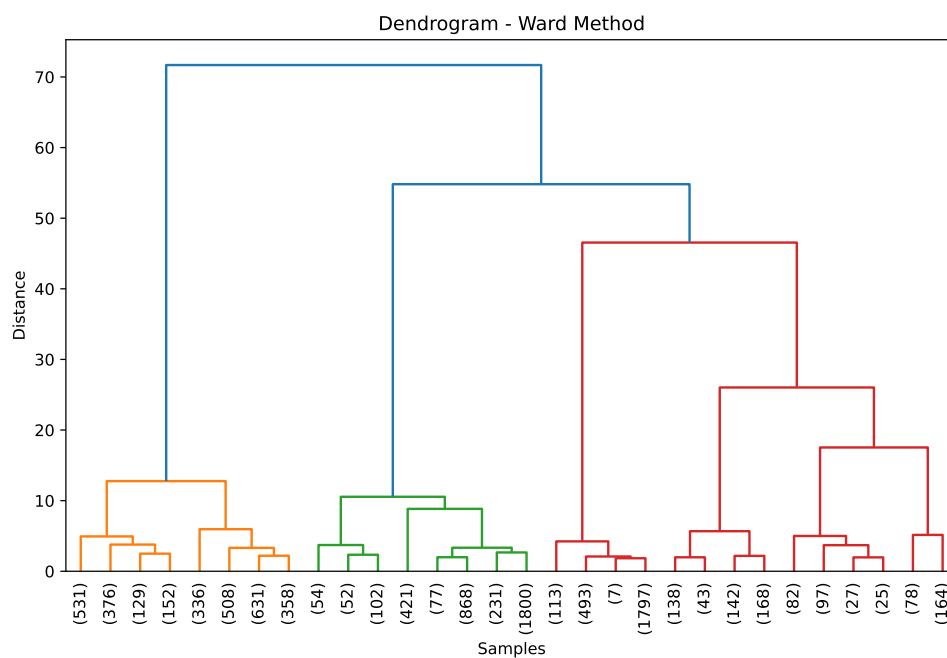


FIGURE C.14. Dendrogram of the Ward linkage method

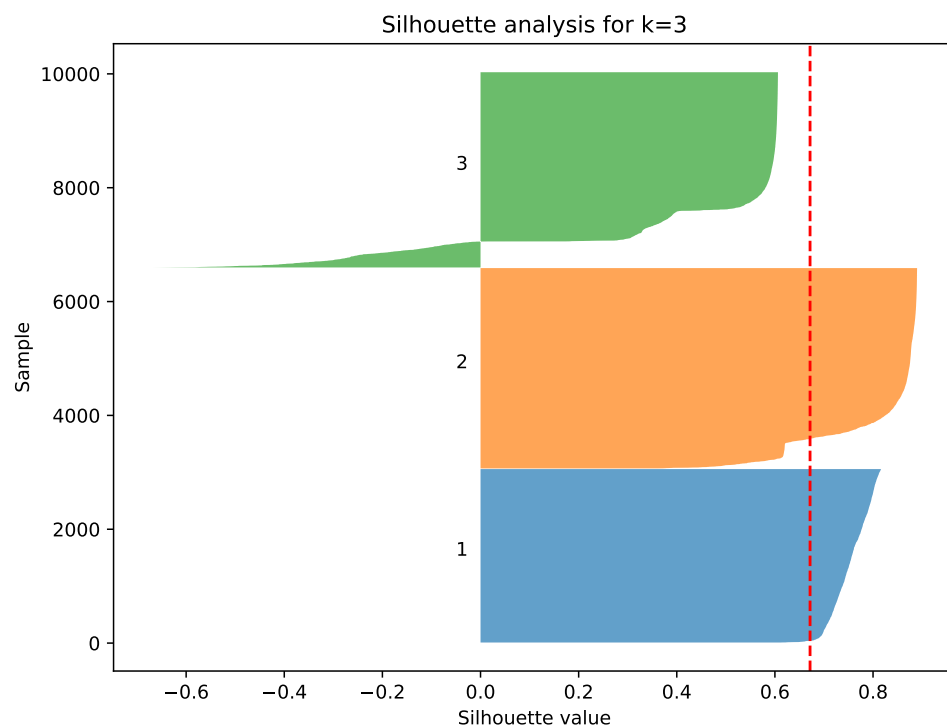


FIGURE C.15. Silhouette plot for Ward method with k=3 Clusters

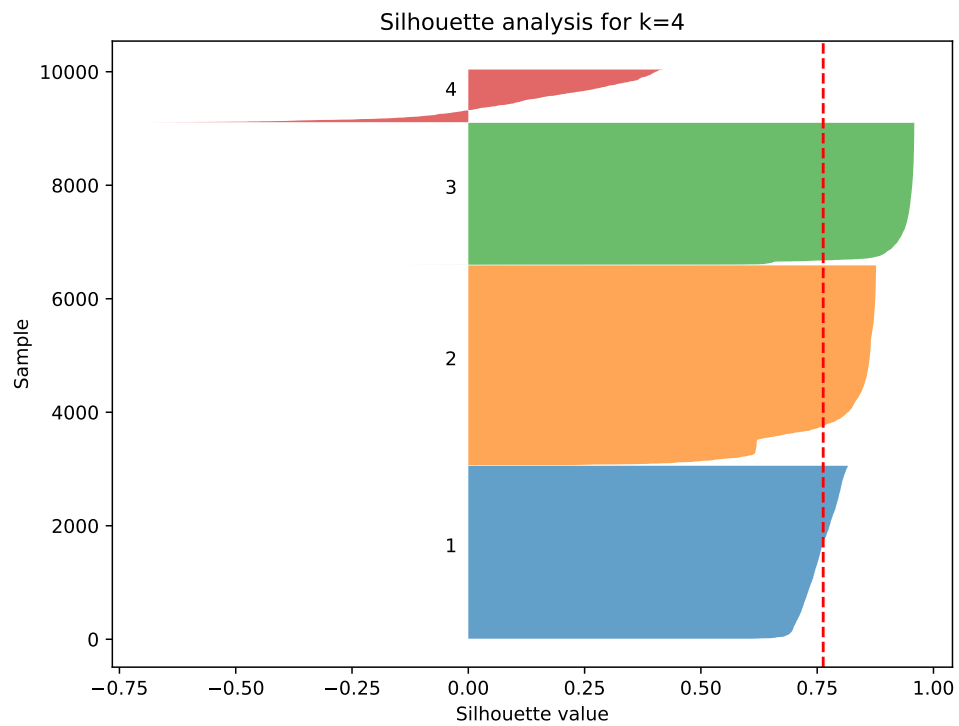


FIGURE C.16. Silhouette plot for Ward method with k=4 Clusters

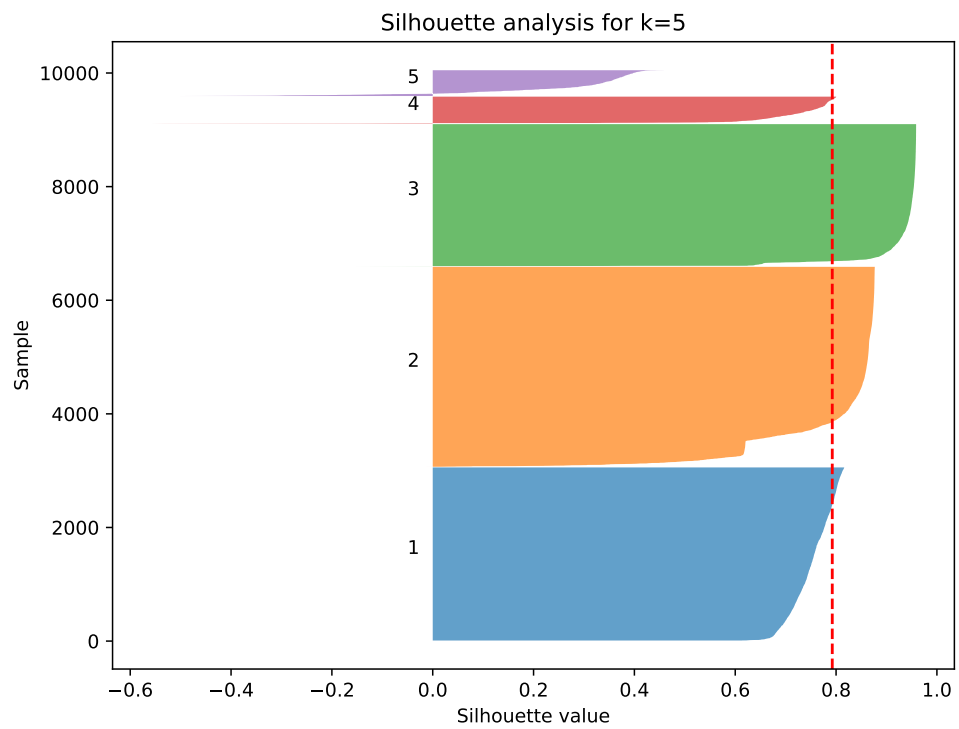


FIGURE C.17. Silhouette plot for Ward method with k=5 Clusters

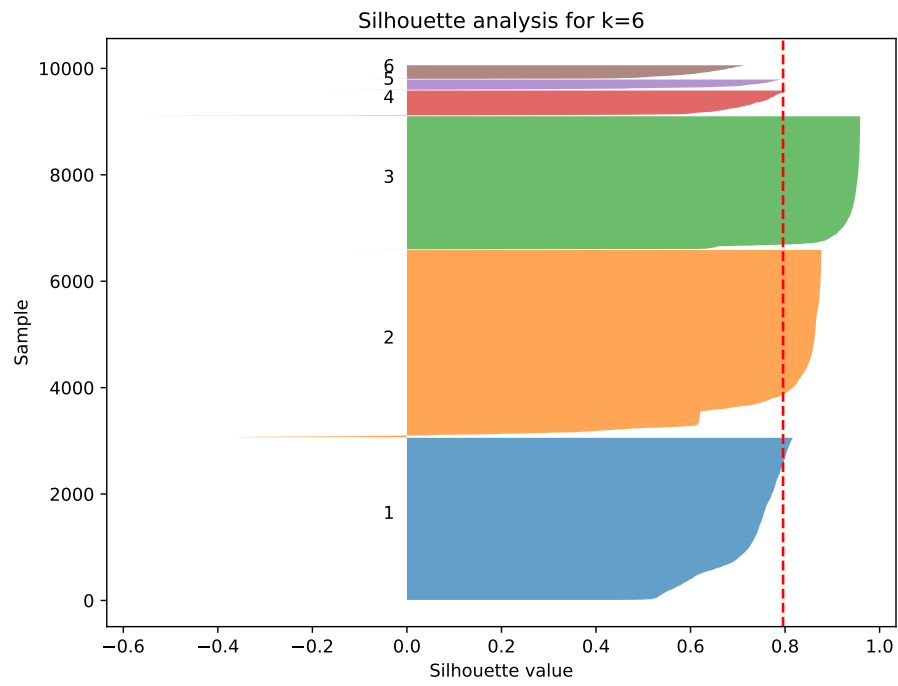


FIGURE C.18. Silhouette plot for Ward method with k=6 Clusters

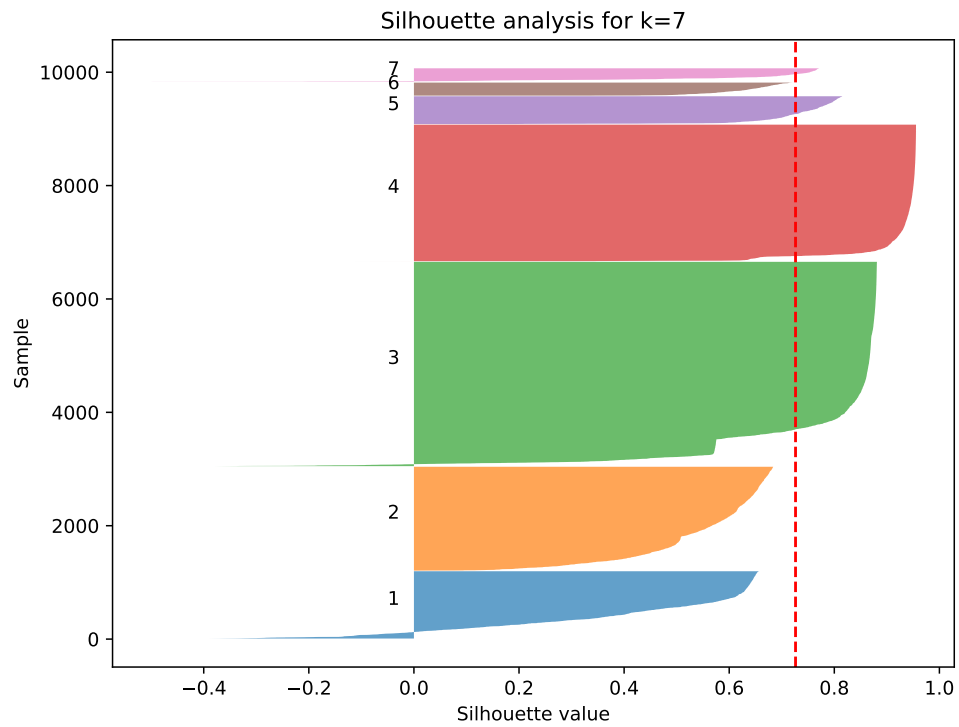


FIGURE C.19. Silhouette plot for Ward method with k=7 Clusters

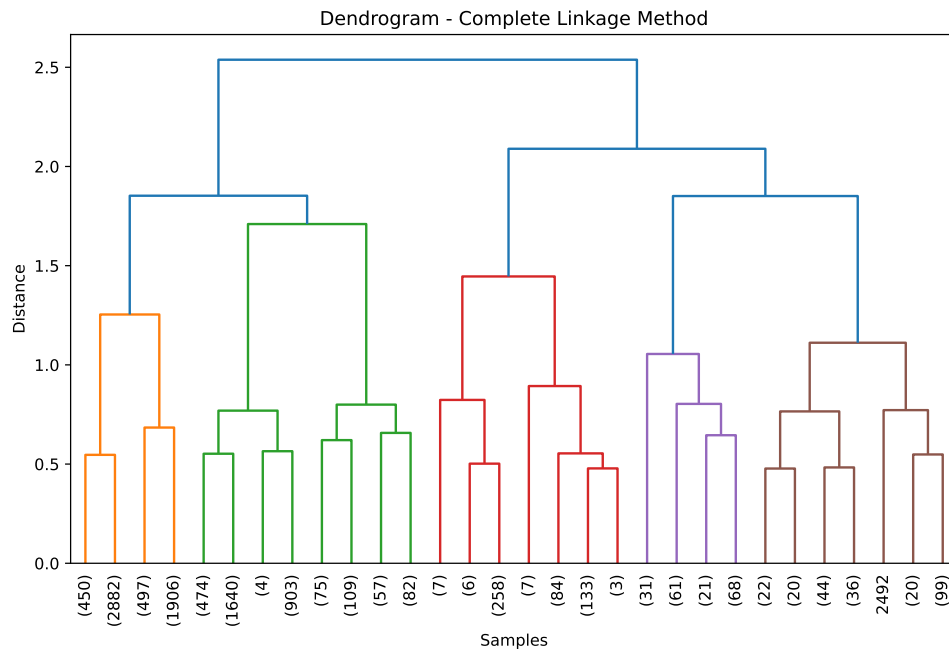


FIGURE C.20. Dendrogram of the Complete linkage method

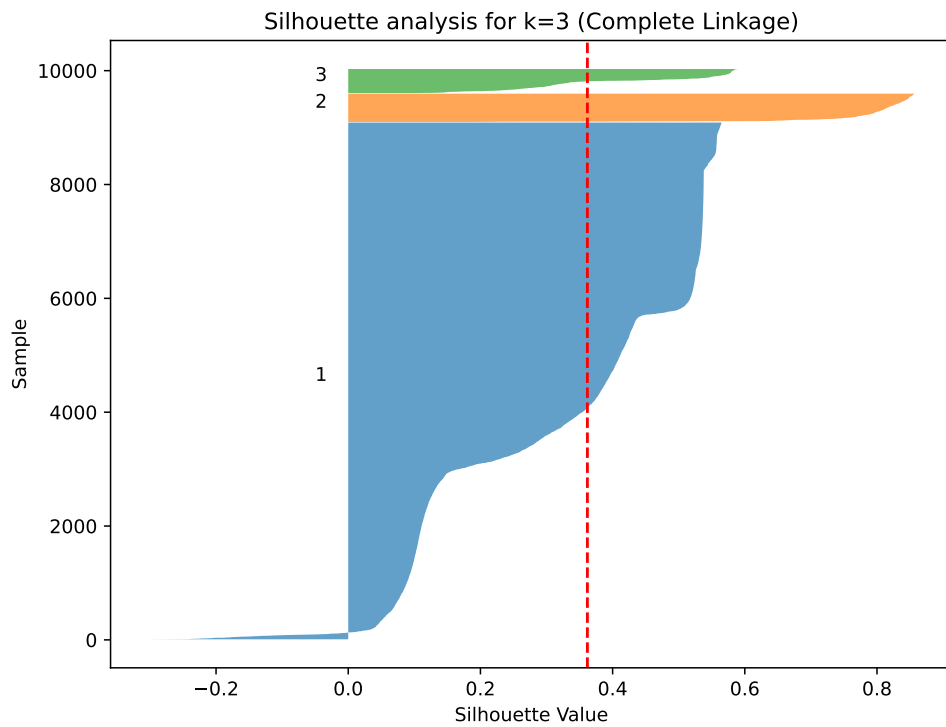


FIGURE C.21. Silhouette plot for Complete linkage method with $k=3$ Clusters

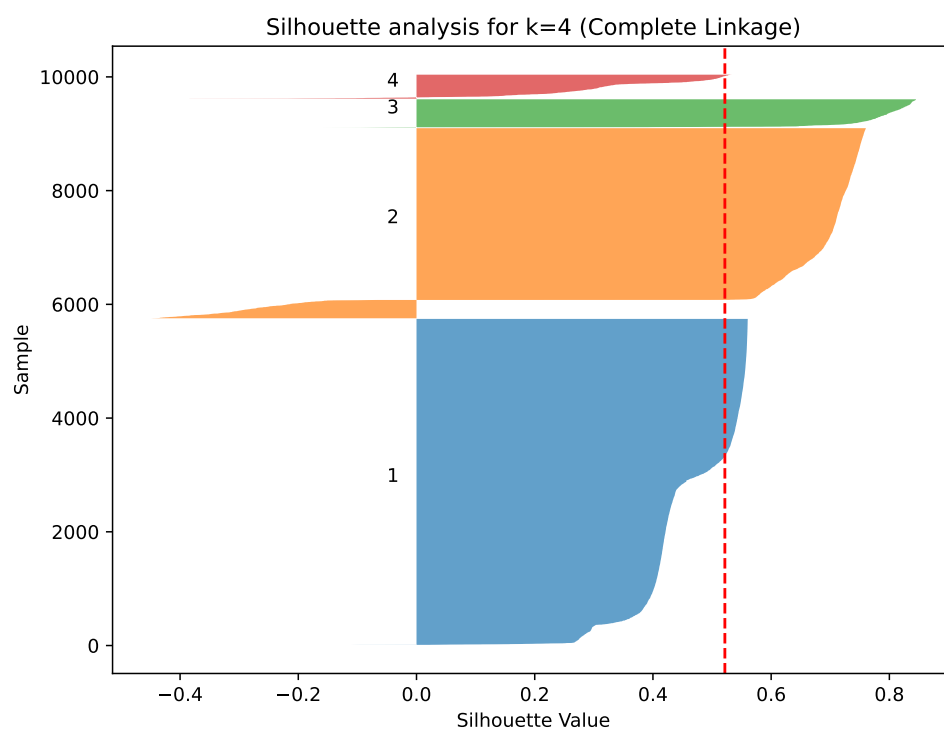


FIGURE C.22. Silhouette plot for Complete linkage method with k=4 Clusters

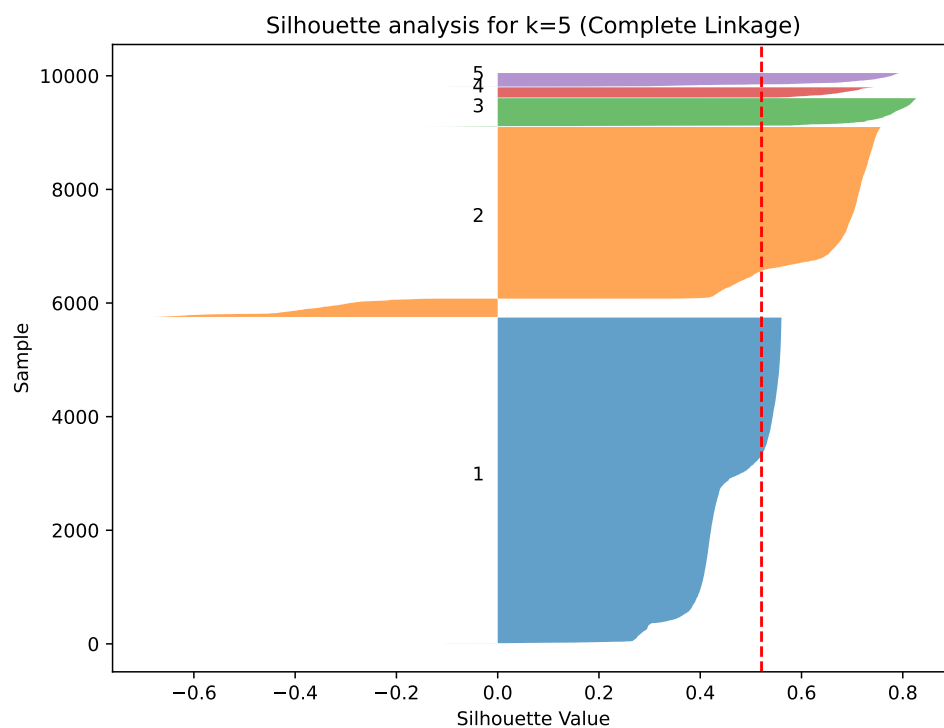


FIGURE C.23. Silhouette plot for Complete linkage method with k=5 Clusters

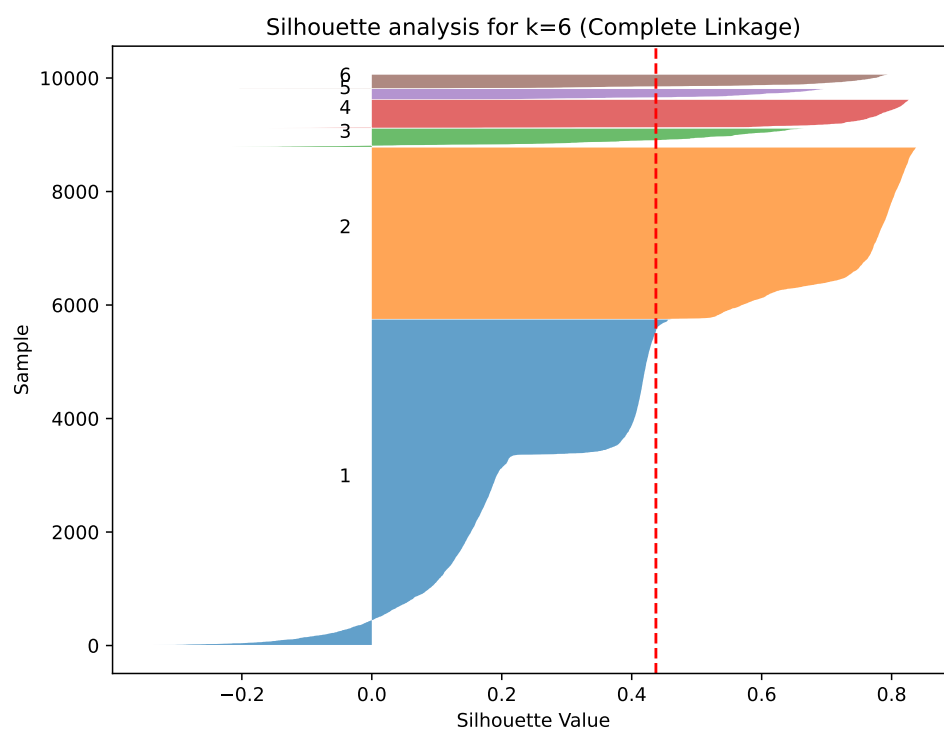


FIGURE C.24. Silhouette plot for Complete linkage method with k=6 Clusters

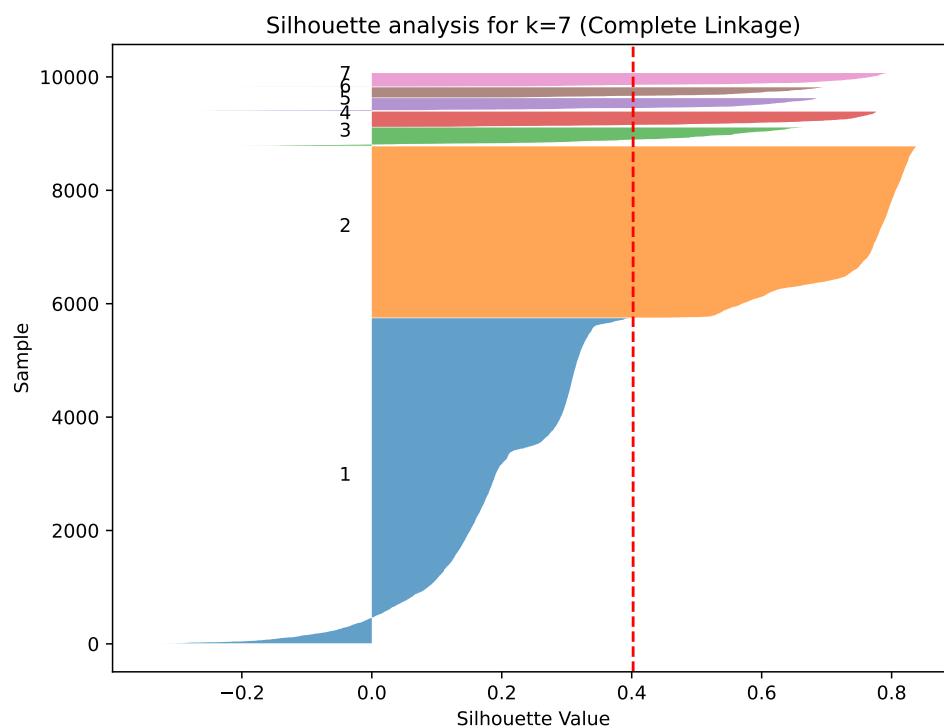


FIGURE C.25. Silhouette plot for Complete linkage method with k=7 Clusters

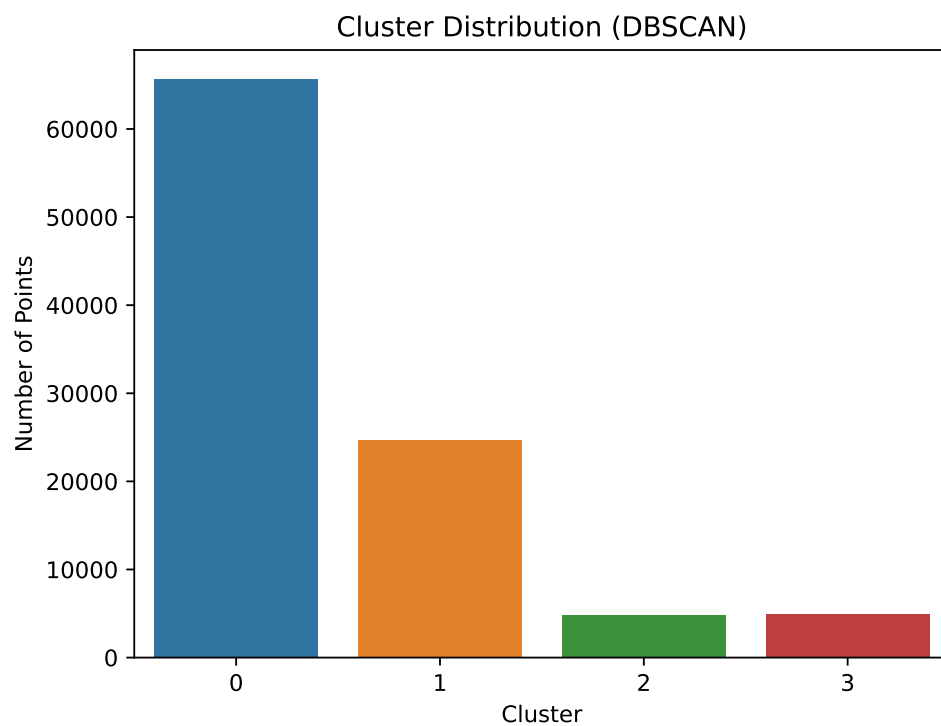


FIGURE C.26. Cluster formation with the DB-SCAN method

K-Means Clustering 3D

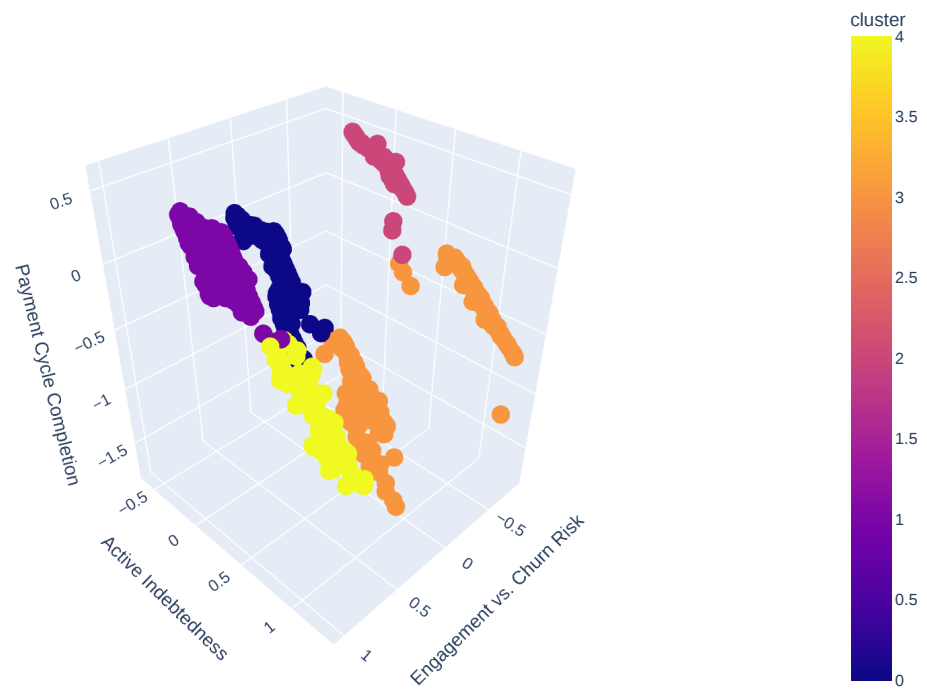


FIGURE C.27. 3D visualization of K-Means clusters diagonal perspective

K-Means Clustering 3D

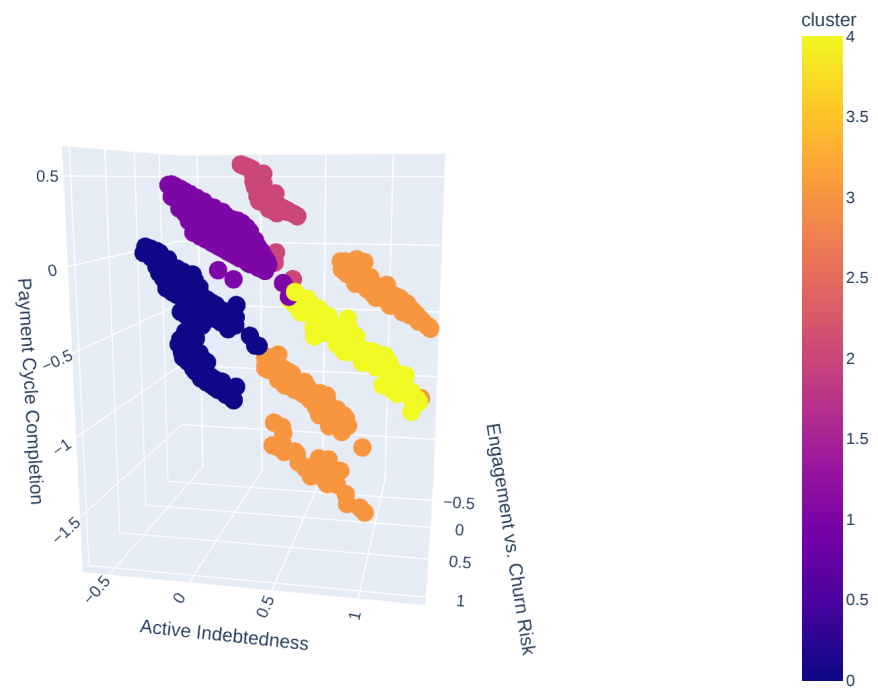


FIGURE C.28. 3D visualization of K-Means clusters side perspective

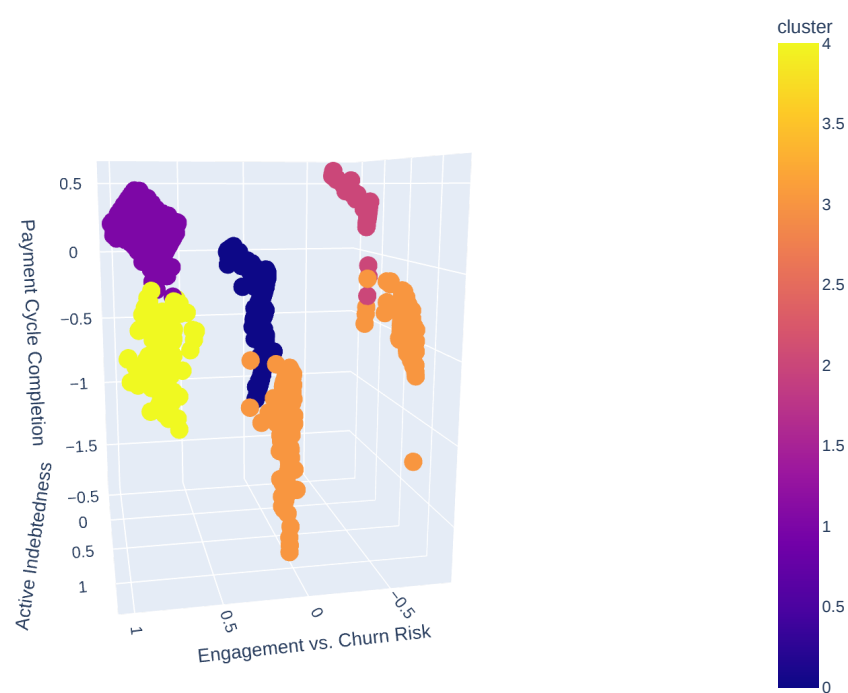


FIGURE C.29. 3D visualization of K-Means clusters opposite side perspective

C.4. Second experiment

TABLE C.20. Logistic Regression - Confusion Matrix for Train Set using common features only

	Predicted 0	Predicted 1
True 0	384849	50217
True 1	13342	421724

TABLE C.21. Logistic Regression - Classification Report for Train Set using common features only

Class	Precision	Recall	F1-Score
0	0.97	0.88	0.92
1	0.89	0.97	0.93
Accuracy	-	-	0.93
Macro avg	0.93	0.93	0.93
Weighted avg	0.93	0.93	0.93

TABLE C.22. Logistic Regression - Confusion Matrix for Test Set using common features only

	Predicted 0	Predicted 1
True 0	96240	12527
True 1	3334	105433

TABLE C.23. Logistic Regression - Classification Report for Test Set using common features only

Class	Precision	Recall	F1-Score
0	0.97	0.88	0.92
1	0.89	0.97	0.93
Accuracy	-	-	0.93
Macro avg	0.93	0.93	0.93
Weighted avg	0.93	0.93	0.93

TABLE C.24. Random Forest - Confusion Matrix for Train Set using common features only

	Predicted 0	Predicted 1
True 0	425116	9950
True 1	4665	430401

TABLE C.25. Random Forest - Classification Report for Train Set using common features only

Class	Precision	Recall	F1-Score
0	0.99	0.98	0.98
1	0.98	0.99	0.98
Accuracy	-	-	0.98
Macro avg	0.98	0.98	0.98
Weighted avg	0.98	0.98	0.98

TABLE C.26. Random Forest - Confusion Matrix for Test Set using common features only

	Predicted 0	Predicted 1
True 0	106280	2487
True 1	1247	107520

TABLE C.27. Random Forest - Classification Report for Test Set using common features only

Class	Precision	Recall	F1-Score
0	0.99	0.98	0.98
1	0.98	0.99	0.98
Accuracy	-	-	0.98
Macro avg	0.98	0.98	0.98
Weighted avg	0.98	0.98	0.98

TABLE C.28. k-NN - Confusion Matrix for Train Set using common features only

	Predicted 0	Predicted 1
True 0	47762	2238
True 1	1662	48338

TABLE C.29. k-NN - Classification Report for Train Set using common features only

Class	Precision	Recall	F1-Score
0	0.97	0.96	0.96
1	0.96	0.97	0.96
Accuracy	-	-	0.96
Macro avg	0.96	0.96	0.96
Weighted avg	0.96	0.96	0.96

TABLE C.30. k-NN - Confusion Matrix for Test Set using common features only

	Predicted 0	Predicted 1
True 0	102782	5985
True 1	5036	103731

TABLE C.31. k-NN - Classification Report for Test Set using common features only

Class	Precision	Recall	F1-Score
0	0.95	0.94	0.95
1	0.95	0.95	0.95
Accuracy	-	-	0.95
Macro avg	0.95	0.95	0.95
Weighted avg	0.95	0.95	0.95

TABLE C.32. Neural Network - Confusion Matrix for Test Set using common features only

	Predicted 0	Predicted 1
True 0	102291	6476
True 1	2539	106228

TABLE C.33. Neural Network - Classification Report for Test Set using common features only

Class	Precision	Recall	F1-Score
0	0.98	0.94	0.96
1	0.94	0.98	0.96
Accuracy	-	-	0.96
Macro avg	0.96	0.96	0.96
Weighted avg	0.96	0.96	0.96

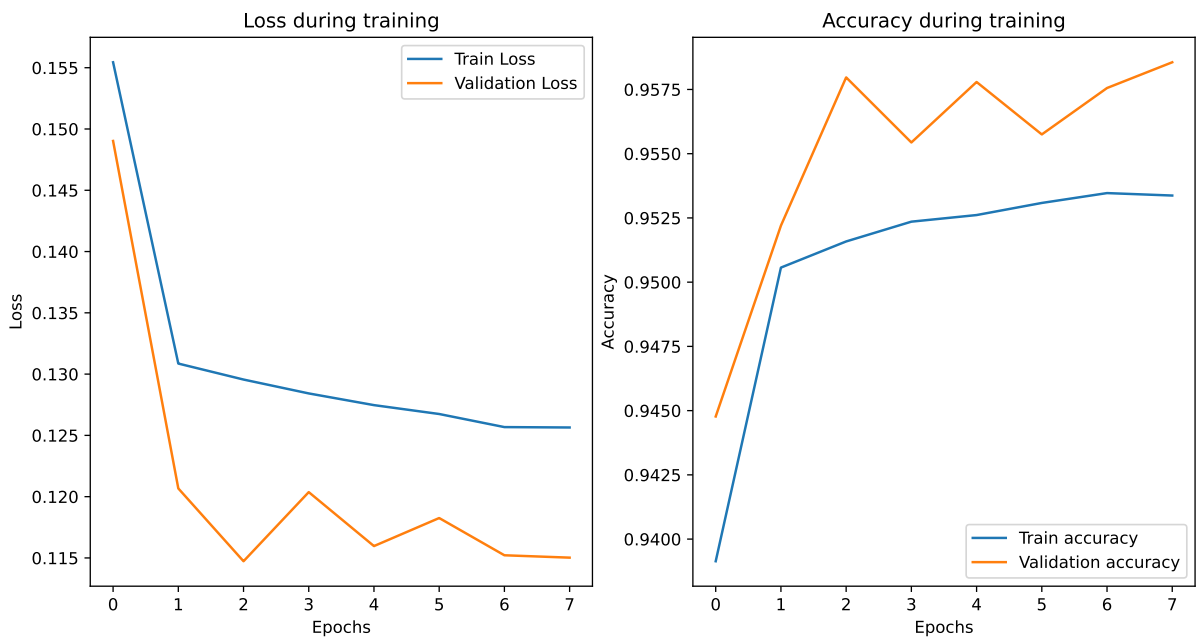


FIGURE C.30. Neural Network - accuracy (right) and loss (left) curves using common features only