NOVA IMS Information Management School

MDSAA

Master Degree Program in

Data Science and Advanced Analytics

Predicting customer churn in telecommunications

A machine learning approach using tensor representations

Duarte Pereira Barbosa

Master Thesis

presented as partial requirement for obtaining a Master's Degree in Data Science and Advanced Analytics

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

Predicting customer churn in telecommunications

A machine learning approach using tensor representations

by

Duarte Pereira Barbosa

Master Thesis presented as partial requirement for obtaining the Master's degree in Data Science and Advanced Analytics, with a specialization in Business Analytics

Supervised by

Augusto Santos, PhD, Nova IMS

July, 2025

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledged the Rules of Conduct and Code of Honor from the NOVA Information Management School.

Lisbon, 15th of July 2025

Duarte Barbosa

ACKNOWLEDGEMENTS

Firstly, I want to thank my supervisor, Professor Augusto Santos, for the encouragement and guidance. Your help has been essential for the completion of this thesis.

I also wish to express my gratitude to KPMG, especially to Luís Duarte, for the valuable insights, discussions and resources shared with me during this work. Your contribution helped me gain a deeper understanding of the topic.

I want to express my deepest appreciation for my family. To my parents, Susana and José, and to my brother João, thank you for the support you have provided throughout this journey.

Finally, I am also grateful to my colleagues and friends from my degree. Thank you for the discussions and shared ideas. To the friends who supported me outside the university, thank you for your encouragement.

ABSTRACT

Customer churn has become an increasingly important concern in the telecommunications sector, where high operational costs and market competition emphasize the need to retain customers. This study aimed to evaluate if advanced methods can improve the identification of customers at risk of churning. Conventional machine learning models revealed limitations in recognizing churners, mainly due to class imbalance, despite achieving a good overall accuracy. To address these challenges, a new method leveraging covariance structures is presented, capturing complex relationships and patterns. In addition, the use of sentiment analysis on customer interactions was explored, highlighting its potential for early churn prediction. While churn prediction remains a difficult task influenced by evolving customer behaviours, this analysis demonstrated that supervised machine learning along with proper feature engineering can enhance churn detection, enabling telecommunication companies to act proactively to improve retention, reduce customer losses and maintain income stability. This work contributes to the adoption of data-driven strategies in customer management, supporting telecommunications companies in addressing churn challenges with greater efficiency.

KEYWORDS

Churn Prediction; Telecommunications; Imbalanced Data; Powers of the Covariance Matrices; Sentiment Analysis

Sustainable Development Goals (SDG):

TABLE OF CONTENTS

1. Introduction 6
2. Literature Review 9
   2.1. Churn Prediction in Telecommunications 9
   2.2. Imbalanced Data in Machine Learning 11
   2.3. Limitations in Existing Literature 12
3. Methodology 13
   3.1. Dataset Description 13
   3.2. Dataset Exploration 14
   3.3. Data Preprocessing 15
   3.4. Feature Selection 16
      3.4.1. Filter Methods 17
      3.4.2. Wrapper Methods 17
      3.4.3. Embedded Methods 17
      3.4.4. Final Insights 18
   3.5. Modelling 19
      3.5.1. Machine Learning models 19
      3.5.2. Addressing Imbalanced Data 21
      3.5.3. Voting Classifier 22
      3.5.4. Proposed Method 23
4. Results and Discussion 30
   4.1. Performance of Traditional Models 30
   4.2. Impact of Oversampling Techniques 31
   4.3. Ensemble Learning Techniques 31
   4.4. Proposed Method 31
5. Future Work 34
   5.1. Sentiment Analysis for Churn Detection 34
6. Conclusions 36
7. References 38

LIST OF FIGURES

Figure 1 – Customer Journey in Telecommunications 6
Figure 2 – Churn proportion in the dataset 14
Figure 3 – Feature selection scores using Lasso 18
Figure 4 – Logistic Regression model results 20
Figure 5 – Gradient Boosting model results 20
Figure 6 – AdaBoost model results 21
Figure 7 – AdaBoost model results with SMOTTENN technique correctly applied 22
Figure 8 – AdaBoost model results with SMOTEENN technique incorrectly applied 22
Figure 9 – Voting Classifier method results 23
Figure 10 – Proposed method pipeline 24
Figure 11 – Tensor representation example 25
Figure 12 – Confusion matrix of the CNN model 26
Figure 13 – CNN model results 27
Figure 14 – Model accuracy and loss over epochs 28
Figure 15 – Confusion matrix of the Proposed Method 29
Figure 16 – Proposed Method results 29
Figure 17 – Comparison of results 30
Figure 18 – Histograms of predicted probabilities (validation and test) 32

1. INTRODUCTION

The telecommunications sector is facing increasing difficulties. Growth has slowed while operational costs continue to rise. According to KPMG International (2024), in 2022, business-to-consumer (B2C) revenues increased by only 1.6%, with an average annual growth rate of only 0.8% between 2018 and 2022. This slow growth can be attributed to factors such as the rise of low-cost competitors and increased price competition. At the same time, the costs of maintaining and expanding telecom networks are rising. Companies are investing in advanced technologies such as 5G and Fiber-to-the-home (FTTH) to meet demand and remain competitive. In 2022, capital expenditures rose by 4.2%, a trend that should persist as new network technologies emerge, according to KPMG International (2024).

In such a challenging context, retaining customers has been a key focus for telecom operators. This thesis aims to explore different methods of identifying churn risk and highlight the role of predictive tools, and AI-powered solutions in increasing the retention rate in these companies. To do this, it is also necessary to consider the customer journey and understand the importance of the retention phase.

The customer journey, shown in Fig. 1, includes several key moments for engagement, support and finally retention.

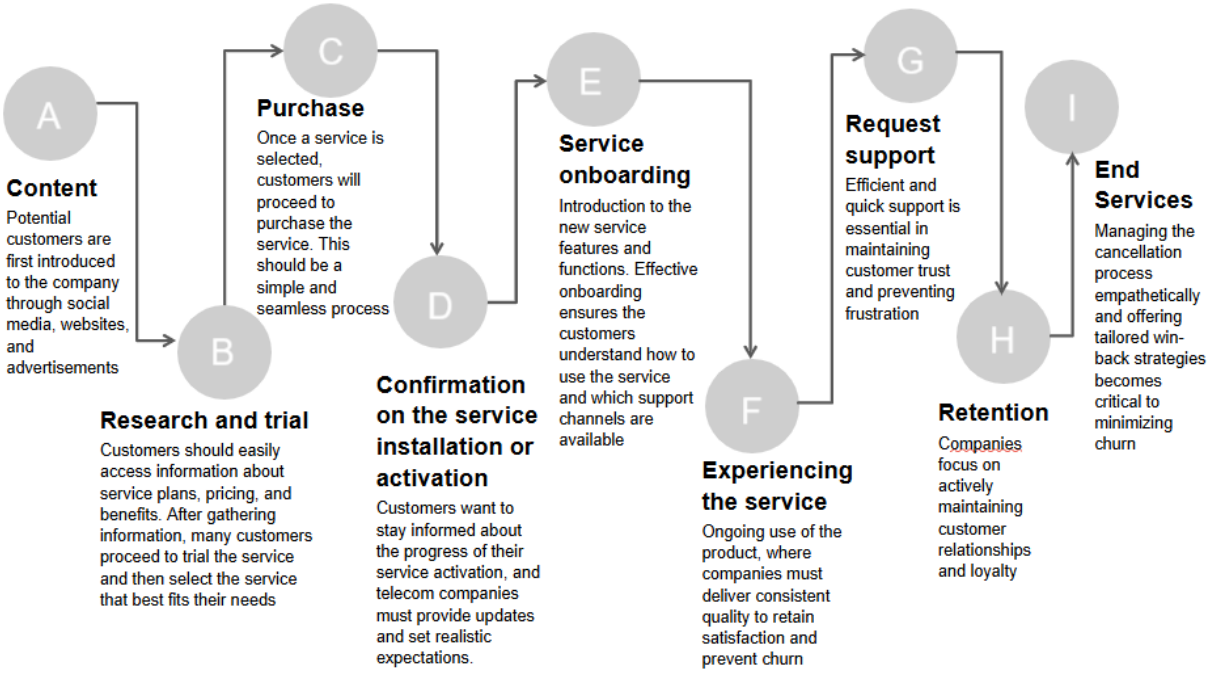

Figure 1 – Customer Journey in Telecommunications

A. It begins with engaging with content, where potential customers are first introduced to the company through social media, websites, and advertisements.

B. Following this, customers move on to browsing service information and researching various services. Customers should easily access information about service plans, pricing, and benefits. After gathering information, many customers proceed to trial service and then select the service that best fits their needs. Telecom companies must offer personalized recommendations during this phase, helping customers to make an informed decision based on their preferences and needs.

C. Once a service is selected, customers will proceed to purchase it. This should be a simple and seamless process.

D. After purchase, the next phase is tracking and getting confirmation on the service installation or activation. Customers want to stay informed about the progress of their service activation, and telecom companies must provide updates and set realistic expectations.

E. Following activation, customers enter the service onboarding phase, where they are introduced to the new service features and functions. Effective onboarding ensures that customers understand how to use the service and which support channels are available.

F. As customers continue to use the service, they enter the phase of experiencing it. This is the ongoing use of the product, where companies must deliver consistent quality to retain satisfaction and prevent churn. However, it is common for customers to inquire about charges or services. A transparent and easy-to-understand billing system can help with concerns.

G. At some point in the customer's journey, they will need to request support. Efficient and quick support is essential in maintaining customer trust and preventing frustration. As customer needs evolve, they may want to upgrade or manage their plans or services. Offering flexible options such as personalized plans or promotions can encourage customers to stay rather than seeking new alternatives.

H. Then there is a retention phase where companies focus on actively maintaining customer relationships and loyalty. During this phase, companies must engage with proactive support, especially when signs of dissatisfaction arise.

I. However, if retention efforts are unsuccessful, customers may proceed to the end services stage. In such cases, managing the cancellation process empathetically and offering tailored win-back strategies becomes critical to minimizing churn.

Customer retention is a top priority for telecom operators, as customer churn represents a threat to the company's revenue stability. Identifying at-risk customers and implementing strategies to improve retention through predictive tools and insights can have a major impact.

By leveraging AI driven solutions, telecom companies can develop customized services such as dynamic pricing and personalized customer offers to meet customer needs.

One of the main causes of customer churn is the way retention and cancellation processes are handled. They often fail to meet the customer's needs and concerns, leading customers to think that the company does not value customer loyalty. Added to this is the perception that customer feedback is often not valued, which affects trust. In addition, many telecom companies struggle to create a good experience for customers who want to cancel a service, which increases dissatisfaction. Customers often go through very slow and complicated processes, increasing frustration.

Looking ahead, there are several opportunities to use advanced technologies to address these pain points. Artificial intelligence can play an important role by analysing feedback in real time. Predictive analytics can identify customers at risk of churn, enabling companies to act proactively and try to resolve concerns before they reach the point of cancellation. These retention strategies can offer customers incentives based on their customer profile or specific reasons for churn, such as offers from competitors or dissatisfaction with a service.

Some possible solutions for the future could include systems such as Next Best Action (NBA), which creates personalized retention offers that consider the customer's value to the company and the customer's reason for cancelling the service. Systems such as Customer Journey Management (CJM) and Business Process Management (BPM) can be helpful to have a more effective retention process, with a customer-based approach. Enhanced Feedback Management Systems (EFMS) can also ensure that customer insights are continually being applied to improve service offerings.

By developing predictive models to identify customers at risk of churn and automating proactive support, telecom companies can significantly enhance their retention efforts. These strategies improve the feedback process, enable personalized offers, and make the cancellation experience less frustrating. As competition intensifies, such improvements position telecom operators to respond more effectively to customer churn challenges.

These insights were made possible with the collaboration of KPMG, which provided access to internal materials, allowing a deeper understanding of different challenges that telecommunication companies are currently facing and possible future solutions.

2. LITERATURE REVIEW

This section provides an overview of the current literature on customer churn prediction, particularly in the telecommunications sector, highlighting existing methods and challenges in predicting churn.

2.1. CHURN PREDICTION IN TELECOMMUNICATIONS

Prabadevi et al. (2023) have analysed customer churn prediction using various machine learning techniques, including Stochastic Gradient Boosting, Random Forest, K-Nearest Neighbors (KNN), and Logistic Regression. Their study evaluates the effectiveness of these algorithms in identifying potential churners based on historical customer data. By analysing metrics such as Receiver Operating Characteristics (ROC) and Area Under the Curve (AUC), the authors determined that the Stochastic Gradient Booster model was the best performer, obtaining an AUC of 0.84, while K-Nearest Neighbors had the lowest performance with an AUC of 0.781. This study highlights the importance of parameter optimization techniques such as Grid Search CV and Randomized Search CV to improve overall accuracy. The study also highlights the need for future research, especially into preprocessing and hyperparameter tuning methods, to improve the performance of churn prediction models.

Wagh et al. (2024) developed methods for customer churn prediction using machine learning techniques. The authors highlight the challenge of identifying churn factors, highlighting that some reasons for churn are obvious, but others can be difficult to detect. The system proposed in this paper applies Decision Trees and Random Forest classifiers before and after performing upsampling. For this purpose, they used SMOTEENN, which combines SMOTE (Synthetic Minority Over-sampling Technique) and ENN (Edited Nearest Neighbor) techniques. The results indicate that the Decision Trees classifier initially performed below expectations due to the dataset being imbalanced. With the use of upsampling techniques, the model reached an overall accuracy of 99% and a recall and precision of 99% as well. The authors suggest, for future improvements, the inclusion of deep learning techniques such as neural networks to capture nonlinear relationships in the data.

Sikri et al. (2024) explored the topic of customer churn prediction using different machine learning techniques, with a special focus on addressing the imbalance in the data. This study introduced a data balancing technique based on a ratio, which effectively acts on balancing the data and improves the model's accuracy. The research compared machine learning algorithms with ensemble techniques such as Gradient Boosting and Extreme Gradient Boosting (XGBoost). The results demonstrated that ensemble techniques achieved better

results than machine learning algorithms, particularly when applied to balanced datasets. Among the different ratios tested, the 75:25 ratio combined with XGBoost achieved an overall accuracy of 89.6%, making it the most efficient model in customer churn prediction. Finally, the paper highlights the importance of balancing techniques in optimizing customer churn prediction.

Xia et al. (2018) presented a churn prediction model for the telecommunications sector using a Hidden Markov Model (HMM), a time-series model designed to predict churners based on patterns. Their analysis highlighted that patterns such as communication quality and activeness are related to churn behaviour. The authors implemented the HMM where the observation states represent consumer behaviour and the hidden states correspond to customer churn. The HMM model was compared with other commonly used models such as Random Forest (RF) for example. The HMM model reached an AUC of 0.855 while the RF model got 0.843. The authors concluded that the quality of features extracted from the dataset is of greater importance than the choice of the classification model. This study also emphasizes the importance of feature engineering in churn prediction models, especially in large-scale data environments.

Edwine et al. (2022) explored the impact of hyperparameter optimization techniques on churn prediction models in the telecommunications sector. The study compared various techniques such as Grid Search (GS), Random Search (RS), and Genetic Algorithm (GA), with a particular focus on improving the performance of machine learning models, such as Random Forest (RF). The results demonstrate that the RF model optimized with the GS technique achieved the best results across various metrics such as accuracy, recall, and precision. The study also concluded that an undersampling of low ratio produced better results when dealing with imbalanced datasets. It was concluded that the RF algorithm optimized with the Grid Search technique is particularly effective in customer churn prediction, overcoming the challenges of class imbalance and obtaining insights into customer retention strategies.

Saha et al. (2024) addressed the challenges of customer churn prediction, noticing that machine learning models struggle to deal with non-linear and complex relationships in the data. To improve performance, the authors proposed a deep learning model, Churn Net, which integrates 1D Convolutional Neural Networks (CNN) and spatial attention modules to get key features and estimate such complex relationships in the data. The model also uses the ability of CNNs to identify important features and patterns in the data. Additionally, the authors also apply data balancing techniques such as SMOTEENN to deal with data imbalance in the dataset. The model proposed achieved better results than typical machine learning models and deep learning models. Finally, the paper also highlights the importance of

addressing data imbalance with data balancing techniques to enhance the model and achieve better results in customer churn prediction.

Liu et al. (2024) introduced a new hybrid deep learning model to improve customer churn prediction by addressing the challenge of complex multimodal data. The model integrates CNN, BiLSTM, and Multi-Head Self-Attention to enhance predictive accuracy. The CNN efficiently extracts local features while the attention mechanism is capable of modelling complex customer behaviour patterns. A systematic hyperparameter tuning and a 10-fold cross validation were applied to ensure model's reliability. The authors concluded that the used method outperformed the existing deep learning models, demonstrating a better capacity for generalization. Finally, the importance of addressing class imbalance in the data, in this case using ADASYN, to improve predictive performance was also highlighted in this paper.

2.2. IMBALANCED DATA IN MACHINE LEARNING

Altalhan et al. (2025) provides an overview of techniques for dealing with data imbalance in machine learning. The study highlights how the upsampling technique can increase the representation of the minority class but, at the same time, may lead to overfitting. Meanwhile, the undersampling technique can reduce the bias in the model, but there is also a risk of losing important data information for model performance. The authors also accentuate the limitations of current techniques in terms of model generalization and robustness. Additionally, they recommend the use of different evaluation metrics instead of just using accuracy, which can be misleading in datasets that present class imbalance. The paper concludes with the suggestion that future work should focus on deep learning-based approaches to improve model performance.

Chen et al. (2024) provided a review on imbalanced training, discussing traditional classification tasks but also regression, clustering, and deep learning. The study investigates both data-level and algorithm-level methods, as well as hybrid techniques, evaluating their efficiency in different scenarios. The authors emphasize that there are many solutions, but challenges regarding scalability and weak adaptability to real-world scenarios remain present and significant. The authors further advocate the need for more robust models capable of dealing with data that is constantly undergoing changes, as seen in real life. To guide future work, the paper outlines some important improvements, including continuous learning methods to address the imbalance in the data and improving evaluation strategies.

Carriero et al. (2025) explore how methods of correcting class imbalance affect the performance of risk prediction models, not only customer churn predictions. Through different simulations and using real-world data, techniques such as oversampling, undersampling, Smote, and ensemble methods were evaluated. Their findings demonstrate that despite some methods improving performance, they often compromised calibration, leading to inflated and misleading results. The authors end up concluding that in many risk prediction scenarios, imbalance correction methods can reduce reliability and should be used with caution, especially when probability estimates are critical for decision-making.

2.3. LIMITATIONS IN EXISTING LITERATURE

The previous section outlined different papers related to this field of study. Although the reviewed studies provide important insights into customer churn prediction, it is possible to find some limitations. A recurring problem is reproducibility, i.e., the lack of detailed descriptions of the different phases of the model pipeline. In some cases, aspects such as data preprocessing or model configuration are presented in a general way and without the needed detail. The lack of specificity may limit the replicability of the results, rendering it harder for other authors to benchmark and apply the knowledge gained from these methods in future work.

Additionally, a few studies use data balancing techniques, such as upsampling, to deal with imbalanced datasets. However, it is not always clear whether these techniques are applied exclusively to the training data or if they are applied to the entire dataset (training and test). In the latter case, there would be data leakage in the model, where information from the test set influences the model training, thereby inflating performance metrics. Although it is not explicitly stated in any paper that this happens, the lack of clarity leaves room for concern.

3. METHODOLOGY

This section outlines our proposed methodology to explore churn prediction in telecommunications. To ensure reproducibility, the data source, preprocessing and modelling employed to obtain the results will be explained.

3.1. DATASET DESCRIPTION

The dataset used for experiments in this paper was obtained from Kaggle, and it is also known as the IBM Watson dataset, released in 2015. This dataset consists of 7043 instances, where each one refers to a customer. Each instance has 21 attributes, where 20 consist of the information related to the customer and the 21st attribute is the target variable "Churn", which determines whether the customer has churned or not.

A detailed description of the dataset and the respective attributes is given below:

- **customerID**: contains customer ID
- **gender**: whether the customer is female or male
- **SeniorCitizen**: whether the customer is a senior citizen or not (0,1)
- **Partner**: whether the customer has a partner or not (Yes, No)
- **Dependents**: whether the customer has dependents or not (Yes, No)
- **tenure**: number of months the customer has stayed with the company
- **PhoneService**: whether the customer has a phone service or not (Yes, No)
- **MultipleLines**: whether the customer has multiple lines or not (Yes, No, No phone service)
- **InternetService**: customer's internet service provider (DSL, Fiber optic, No)
- **OnlineSecurity**: whether the customer has online security or not (Yes, No, No internet service)
- **OnlineBackup**: whether the customer has online backup or not (Yes, No, No internet service)
- **DeviceProtection**: whether the customer has device protection or not (Yes, No, No phone service)
- **TechSupport**: whether the customer has tech support or not (Yes, No, No internet service)
- **StreamingTV**: whether the customer has streaming TV or not (Yes, No, No internet service)
- **StreamingMovies**: whether the customer has streaming movies or not (Yes, No, No internet service)
- **Contract**: the contract term of the customer (Month-to-month, One year, Two year)
- **PaperlessBilling**: whether the customer has paperless billing or not (Yes, No)
- **PaymentMethod**: the customer's payment method (Electronic check, Mailed check, Bank transfer (automatic), Credit card (automatic))
- **MonthlyCharges**: the amount charged to the customer monthly

- **TotalCharges**: the total amount charged to the customer
- **Churn**: whether the customer churned or not (Yes, No)

3.2. DATASET EXPLORATION

In the dataset under analysis, there is a significant class imbalance between churners and non-churners. Specifically, there are 1869 churners and 5163 non-churners, representing approximately 26.6% and 73.4% of the dataset, respectively. This disproportion in class distribution can have a big impact on the performance of predictive models, namely, it can naturally bias supervised methods toward the majority class.

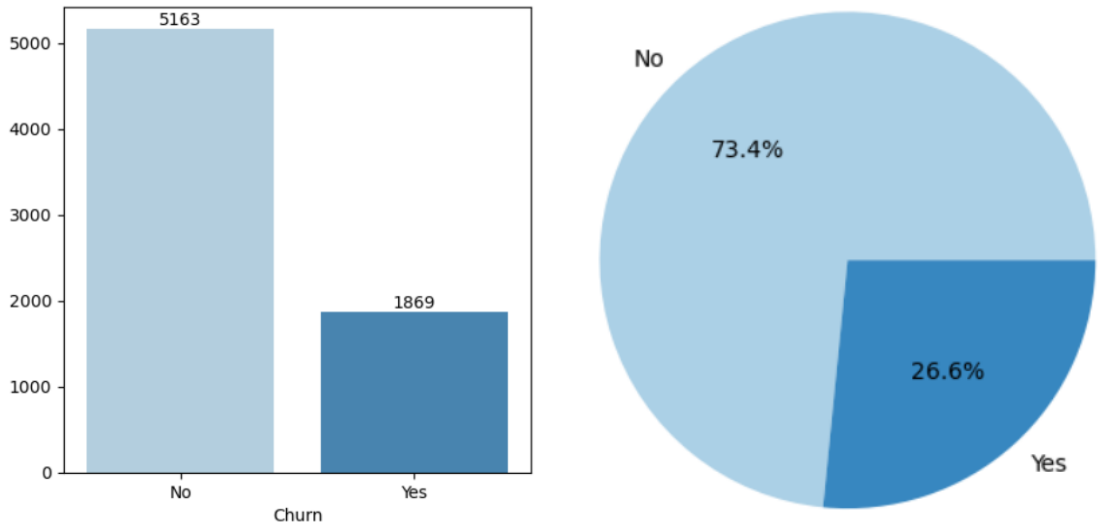

Figure 2 – Churn proportion in the dataset

Class imbalance remains a challenge in classification tasks because most machine learning algorithms are made to reach high overall accuracy. When one class dominates the dataset, the model could become biased toward predicting the majority class (in this case, non-churners). This represents a problem in churn prediction tasks, where identifying customers who are likely to churn is more important than getting a high overall accuracy.

To deal with this problem, resampling techniques can be used to rebalance the data and improve the model's sensitivity to the minority class. As described by Altalhan et al. (2025), oversampling and undersampling are two approaches for balancing the class distribution across the dataset. Oversampling is the increase of representation of the minority class by replicating samples or creating synthetic ones using algorithms like SMOTE (Synthetic Minority

Oversampling Technique). While this method attempts to preserve minority class information, it may also increase the risk of overfitting. The model might also learn about the statistics of the synthetic data that may differ from the underlying real-world samples. This can happen when the model learns the training data too well, including noise, leading to poor generalization. On the other hand, undersampling decreases the number of majority class samples to get a balanced training set. This approach can lower the training time and model bias but could cause the loss of important data and lead to underfitting. This can happen when the model is too simple to capture the patterns in the data, causing poor performance on both training data and testing data (unseen data).

3.3. DATA PREPROCESSING

This section describes the preprocessing steps used to prepare the dataset for machine learning. Raw datasets are often incomplete, inconsistent, and could hold redundant information. Without handling them correctly, models can end up being biased or fail to generalize well to unseen data.

To ensure data consistency, the column 'TotalCharges' was converted to numerical values, and so 11 invalid values were replaced by NaN (which stands for missing values), ensuring that the column only contains numeric data. Since these represented a very small fraction of the dataset, they were considered insignificant and were dropped from the dataset. The 'customerID' was dropped because it represented a unique identifier for each customer and did not contribute to the predictive model. Additionally, a new feature, 'Number_AdditionalServices', was created to have a new column with the total number of add-on services a customer subscribes to. This feature contributes for a better representation of customer engagement and includes services such as 'OnlineSecurity', 'DeviceProtection', 'StreamingMovies', 'TechSupport', 'StreamingTV', and 'OnlineBackup'.

The dataset holds both numerical and categorical features, each requiring appropriate preprocessing to be used in machine learning models. Three different numerical columns were identified: 'tenure', 'MonthlyCharges' and 'TotalCharges', which were preserved in their original form as numerical values. Before further processing, outliers in these numeric columns were capped using the Interquartile Range (IQR) method to limit the influence of extreme values. As highlighted by Zhu et al. (2024), categorical variables, however, need to be encoded into a numerical format to be processed by machine learning models. Label encoding was applied to binary categorical features: 'Churn', 'gender', 'Partner', 'Dependents', 'PhoneService', and 'PaperlessBilling'. In label encoding, each category is replaced by a unique integer, 0 and 1 for binary variables. This approach is appropriate for binary variables because it does not introduce ordinal relationship. For categorical features with more than two possible values ('MultipleLines', 'InternetService', 'OnlineSecurity', 'OnlineBackup', 'DeviceProtection', 'TechSupport', 'StreamingTV', 'StreamingMovies', 'Contract',

'PaymentMethod', and 'Number_AdditionalServices') one-hot-encoding was used. As explained by Poslavskaya et al. (2023), one-hot-encoding represents each category as a sparse vector containing mostly zeros, with a single 1 indicating the active category. This method created new binary columns for each category in a feature. A value of 1 in one of these columns indicates the presence of this category, while the rest of the columns for that feature are set to 0. To avoid the dummy variable trap, which introduces multicollinearity by making one column a perfect combination of others, the first category in each feature was dropped during encoding. For example, the 'Contract' column, which has three categories (Month-to-month, One year, Two year), is encoded into only two binary columns, as the third can be deduced.

After handling the categorical features, the next step is to partition the dataset into training and test sets. This step ensures that the model is evaluated on unseen data, which is crucial to get a realistic understanding of how well the model performs in real-world scenarios. The dataset was split into two subsets: a training set (80% of the data), which will be further split internally during model training to create a validation set for tuning, and a test set (20% of the data). It is important to maintain the representation of the target variable ('Churn') in both subsets, which can be done using a stratified split. This guarantees that both training and test sets have the same proportion of the target variable, avoiding bias during evaluation.

Since the dataset has numerical features with significantly different scales, it is important to normalize these values, so each feature contributes equally to the model's learning process. Without normalization, machine learning algorithms that rely on distance calculations, like logistic regression or neural networks, might perform badly because they could give more weight to features with bigger ranges. To address this, normalization was done using StandardScaler, which transforms the features to have a mean of zero and a standard deviation of one. However, it is important to handle the normalization process with caution to prevent data leakage. This can happen when information from the test set is used in training the model, leading to an optimistic view of its performance. To avoid data leakage, normalization was performed separately for the training and test sets, namely, first the data splitting (into training and test) was performed, and then the data was independently pre-processed at each subset. After completing the preprocessing steps, the dataset now contains 36 encoded and normalized features. With these transformations in place, the dataset is ready for the next steps in the modelling process.

3.4. FEATURE SELECTION

Selecting the right features is an important step in creating an effective model. It helps to reach a higher predictive performance, reducing overfitting and simplifying the model by eliminating irrelevant or redundant information. As indicated by X. Cheng [15], there are three

main groups of feature selection methods. To identify these features, filter, wrapper, and embedded methods were applied.

3.4.1. Filter Methods

Filter methods were used to evaluate features based on their intrinsic statistical properties. Variance analysis was carried out first, which highlighted low-variance features like 'gender' that exhibited the lowest variability, rendering them the main candidates for elimination in the feature selection phase.

Correlation was also used to assess linear relationships between features or variables, helping to indicate variables that may present redundancy.

Additionally, Mutual Information (MI) was also used to estimate the dependency between individual features and the target variable. Features such as 'PhoneService', 'OnlineBackup_Yes', and 'gender' got zero MI scores, implying that these features have no informative value for predicting churn, and the target variable is independent of these features. Conditional Mutual Information (CMI), which considers feature dependencies, indicated features like 'StreamingMovies_Yes', 'DeviceProtection_Yes', and 'gender' as redundant when conditioned on other variables. This means that the target variable is conditionally independent of these variables given another subset of variables.

3.4.2. Wrapper Methods

Wrapper methods evaluate subsets of features based on the performance of a chosen model. Recursive Feature Elimination (RFE) was applied using a classifier to repeatedly remove the least significant features. With this process, 'MultipleLines_No phone service', 'PaymentMethod_Mailed Check' and 'Partner' were identified as having lower contribution and were indicated for removal.

3.4.3. Embedded Methods

Embedded methods integrate feature selection as part of the model training process itself, allowing the model to remove less important features during training. As X. Cheng [15] indicated, regularization techniques like Lasso (L1 regularization) are examples of this approach, striking a balance between efficiency and performance, making them a common choice for many applications. In this work, Lasso was performed for this purpose and features like 'Number_AdditionalServices_1', 'StreamingMovies_No internet service', and once again 'MultipleLines_No phone service' were identified as less important compared to other features (with a Lasso score of 0).

Figure 3 – Feature selection scores using Lasso

3.4.4. Final Insights

To decide which features should be removed from the dataset, only those indicated as irrelevant by both RFE and Lasso methods were dropped. In this case, 'MultipleLines_No phone service' was the only feature consistently flagged by these two methods. Although Mutual Information (MI) and Conditional Mutual Information (CMI) identified other low-relevance features ('gender', for example), they were not included for removal because the other performed methods did not identify those features as less important. This approach guarantees that the feature selection process is based on agreement between two different techniques, reducing the risk of losing important information.

While the features identified by MI and CMI were not used in the final removal step, it is still valuable to consider these methods and the features indicated for removal by these methods, as they provide complementary perspectives on feature relevance and can also offer additional insights for future refinement.

3.5. MODELLING

In this section, different machine learning models were tested as first approaches to the churn prediction task. These commonly used supervised machine learning models were applied to assess their performance on the dataset and to understand the relationship between model complexity, data balance, and overall prediction accuracy. Additionally, a new method aiming to improve upon other approaches was developed and will be explained in this section.

3.5.1. Machine Learning models

A variety of learning algorithms were tested for their effectiveness in predicting customer churn. These models were not only selected due to their popularity and performance in classification tasks, but also to evaluate how linear and non-linear models perform on the dataset. For each model, hyperparameter tuning was applied using grid search combined with k-fold cross-validation to improve model's overall performance.

Hyperparameter tuning is an important part of the modelling process, as choosing the correct parameters can have a big impact on the model's performance. In these experiments, Grid Search was used to evaluate a broad range of hyperparameter values. For each combination of parameters, the chosen model was trained using a 5-fold cross-validation, a technique that separates the training data into five new subsets. As indicated by Lalwani et al. (2022), k-fold cross-validation shuffles the dataset randomly and splits the training set into k groups. One group is then randomly chosen as a validation set while the rest serve as training sets. In this case, the model was trained in four of these subsets and validated on the fifth one. This process guarantees robust performance estimation by mitigating the variance associated with a single train and test split.

The first model tested was Logistic Regression, a linear classifier that models the probability of a binary target variable using the logistic function. As indicated by Sikri et al. (2024), this model is one of the most famous supervised machine learning algorithms and can be used to predict the dependent variable by using the independent variables. It estimates the relationship between features and the target variable. In this case, the regularization strength and solver algorithm were tuned. The best performing configuration achieved an accuracy of 81% and a class 1 (churn class) recall of 57% on the test set. The classification report showed

high precision and recall for the non-churn class but more moderate performance on the churn class, which is common when dealing with imbalanced data.

```
--- Logistic Regression Results ---
Best Parameters: {'C': 0.1, 'solver': 'lbfgs'}
Accuracy: 0.8123667377398721
Classification Report:
              precision    recall  f1-score   support

           0       0.85      0.90      0.88      1033
           1       0.68      0.57      0.62       374

    accuracy                           0.81      1407
   macro avg       0.76      0.73      0.75      1407
weighted avg       0.80      0.81      0.81      1407
```

Figure 4 – Logistic Regression model results

The next algorithm explored was Gradient Boosting, an ensemble method that builds models sequentially. As explained by Sikri et al. (2024), this method first implements a model, then sequentially builds additional models to correct its inaccuracies, aligning the models in a sequential manner. Gradient boosting can recognize patterns and interactions between features, an effective method when dealing with structured data. The hyperparameters optimized in this model were the learning rate, the number of estimators, and the maximum tree depth. With its best configuration, Gradient Boosting reached an overall test accuracy of 81% and a class 1 recall of 53%, demonstrating slightly more flexibility but facing similar limitations in churn recall.

```
--- Gradient Boosting Results ---
Best Parameters: {'learning_rate': 0.1, 'max_depth': 5, 'n_estimators': 50}
Accuracy: 0.8052594171997157
Classification Report:
              precision    recall  f1-score   support

           0       0.84      0.90      0.87      1033
           1       0.67      0.53      0.59       374

    accuracy                           0.81      1407
   macro avg       0.75      0.72      0.73      1407
weighted avg       0.80      0.81      0.80      1407
```

Figure 5 – Gradient Boosting model results

Following this, AdaBoost was tested. AdaBoost also builds an ensemble of weak learners by placing greater emphasis on misclassified instances in subsequent iterations. As highlighted by Lalwani et al. (2022), it retrains iteratively by choosing the training set based on the accuracy of the previous training. This allows the algorithm to focus on the more difficult examples in the dataset. The grid search covered the number of estimators, the learning rate, and the boosting algorithm variant. AdaBoost obtained a test accuracy of 81% and a class 1 recall of 55%, again showing high performance for the dominant class, but limited recall for churners.

```
--- AdaBoost Results ---
Best Parameters: {'algorithm': 'SAMME.R', 'learning_rate': 1.0, 'n_estimators': 50}
Accuracy: 0.8137882018479033
Classification Report:
              precision    recall  f1-score   support

           0       0.85      0.91      0.88      1033
           1       0.69      0.55      0.61       374

    accuracy                           0.81      1407
   macro avg       0.77      0.73      0.74      1407
weighted avg       0.81      0.81      0.81      1407
```

Figure 6 – AdaBoost model results

3.5.2. Addressing Imbalanced Data

To deal with the imbalance observed in the dataset, where churners are underrepresented (73,4% of the data are non-churners), resampling methods were applied to face this challenge. The approach used was the method SMOTEENN, a hybrid approach that combines oversampling and undersampling techniques. SMOTE (Synthetic Minority Over-sampling Technique) creates synthetic samples of the minority class, while ENN (Edited Nearest Neighbours) eliminates ambiguous examples from the majority class.

When SMOTEEN was applied only to the training data, followed by retraining the previously tuned AdaBoost model, performance improved for the minority class. The model's recall for churners increased from 55% to 87%, although overall accuracy dropped from 81% to 72%. This trade-off is expected, as identifying churners is often more valuable than overall accuracy.

```
Accuracy: 0.720682302771855
Classification Report:
              precision    recall  f1-score   support

           0       0.93      0.67      0.78      1033
           1       0.49      0.87      0.62       374

    accuracy                           0.72      1407
   macro avg       0.71      0.77      0.70      1407
weighted avg       0.81      0.72      0.74      1407
```

Figure 7 – AdaBoost model results with SMOTTENN technique correctly applied

An additional approach was guided where SMOTEENN was applied to both training and test sets. This improper use of this method led to data leakage, causing an artificially high overall test accuracy of 92% and a class 1 recall of 94%. The classification report (which shows the metrics for each class) in this case was misleadingly optimistic, emphasizing the importance of using resampling techniques strictly within the training phase to avoid contaminating evaluation results.

```
Accuracy (data leakage): 0.9243281471004243
Classification Report:
              precision    recall  f1-score   support

           0       0.92      0.90      0.91       587
           1       0.93      0.94      0.94       827

    accuracy                           0.92      1414
   macro avg       0.92      0.92      0.92      1414
weighted avg       0.92      0.92      0.92      1414
```

Figure 8 – AdaBoost model results with SMOTEENN technique incorrectly applied

3.5.3. Voting Classifier

To improve predictive performance, an ensemble technique was performed using a Voting Classifier. Ensemble training methods support the diversity of different models to improve generalization. In this case, a voting method was used, which combines the class probabilities from different models and attributes the final prediction based on the average probabilities.

The ensemble combined the best performing versions of Logistic Regression, Gradient Boosting, and AdaBoost. After training on the scaled training set, the Voting Classifier achieved a test accuracy of 81% and a class 1 recall of 56%. More importantly, it achieved a more balanced classification performance, improving the recall for the churn class compared to individual models. The coordination between the linear characteristics of Logistic Regression

and the non-linear, tree-based capabilities of the boosting methods contributed to the ensemble's robustness.

Classification Report on Test Set:

	precision	recall	f1-score	support
0	0.85	0.91	0.88	1033
1	0.69	0.56	0.61	374
accuracy			0.81	1407
macro avg	0.77	0.73	0.75	1407
weighted avg	0.81	0.81	0.81	1407

Figure 9 – Voting Classifier method results

These modelling experiments offered a comprehensive view of how various machine learning algorithms perform on the churn prediction task, both with and without addressing data imbalance. The experiments revealed that while standard classifiers like Logistic Regression, Gradient Boosting, and AdaBoost performed reasonably well, their default behaviour leaned toward the majority class. Only after addressing data imbalance with resampling techniques, without any data leakage, did the model show improvements in identifying churners. However, this came at the cost of reduced overall accuracy, as the model completely compromised the majority class to improve the minority class detection. This trade-off emphasizes the necessity of new approaches when working with imbalanced datasets, where correctly identifying at-risk customers is more valuable than simply reaching a high overall accuracy. Telecommunications churn prediction is an example of this situation, as retaining existing customers is often more cost-effective than acquiring new ones, making accurate identification of churners a business priority.

3.5.4. Proposed Method

While traditional machine learning models such as Logistic Regression, Gradient Boosting, and AdaBoost reached good results for overall accuracy, they exhibited limitations in consistently capturing the minority churn class. Specifically, these models struggled with low recall for churners, often misclassifying many churner cases due to the class imbalance and subtle patterns. This represents a critical problem in churn prediction, where failing to identify potential churners can have significant business consequences for the company.

To overcome the limitations observed in traditional models, particularly the low recall for the minority churn class, a new method was developed. It consists of two main phases: a tensor group classification phase followed by a signature-based classification.

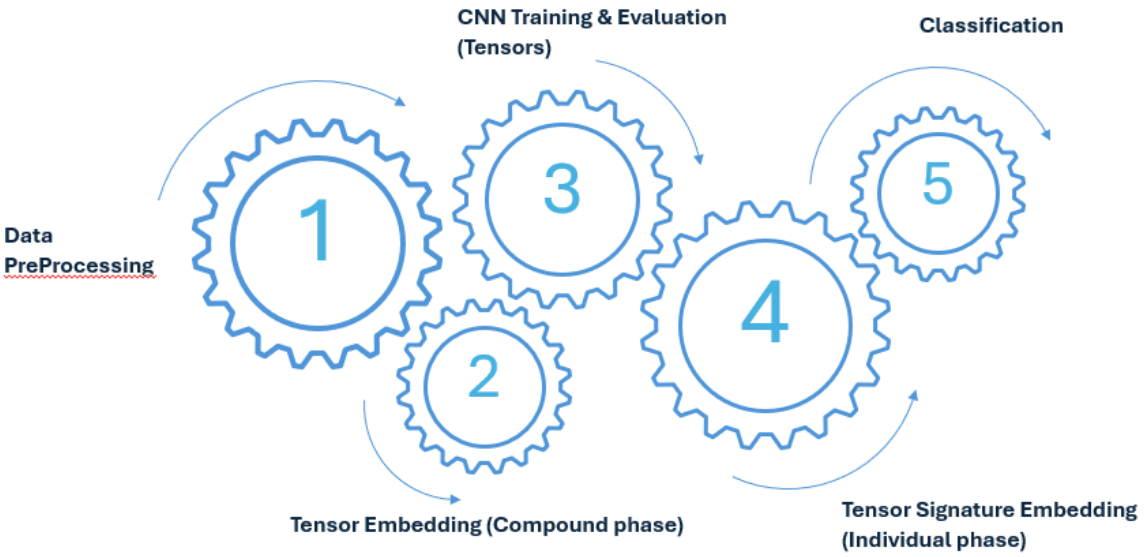

Figure 10 – Proposed method pipeline

First, the training set was concatenated to combine feature data and the target churn labels. The combined data was then separated based on the churn label into two distinct subsets, one containing non-churn cases and the other containing churn cases. Then these subsets were divided into groups of clients. Given the imbalance in the dataset, the number of groups into which each subset was split was set differently, 150 groups for non-churners and 50 groups for churners, ensuring manageable group sizes that still retained meaningful variance within each group. The splitting into groups was done using a custom function designed to divide the data into nearly equal-sized subsets. This function calculates the approximate size of each subset and divides the data accordingly. In possible cases where the data does not divide evenly, the leftover samples are distributed across the first few groups to maintain balance. This resulted in groups of roughly the same size, with the churn groups having 30 samples each and the non-churn groups having 28 samples each. This process results in groups that capture the local structure and variance within each class.

After creating the groups, the next step was to compute covariance matrices for each group. Covariance matrices capture the relationships between features within each group, providing a summary of the internal structure and variability of the data. To improve the representational power of these covariance matrices, fractional matrix powers were computed (3, 3.25, 3.5, 3.75, and 4), creating a set of powered covariance matrices for each group. These matrices reflect different degrees of correlation structure in data, improving the

feature representation. To guarantee numerical stability during these operations, a small value was added to the diagonal of each covariance matrix before raising it to fractional powers, preventing possible issues. Since fractional powers can form matrices with complex values, only the real part of these matrices was kept, as it contains the meaningful variance information relevant to classification.

The resulting powered covariance matrices were stacked to form a three-dimensional tensor for each group. This tensor-based approach is designed to allow the model to recognize differences in correlation structure between churn and non-churn groups, offering a more informative input to the classifier.

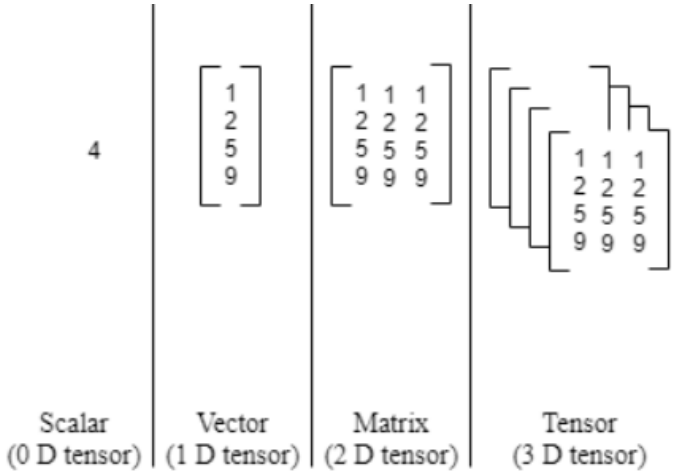

Figure 11 – Tensor representation example

Fractional powers of the covariance matrix were first deployed to estimate brain structural connectivity from fMRI time series data (Liégeois et al., 2020). As such, they are structure-enriched features that convey informative signatures relevant to the target variable, in this case, churn prediction.

Following tensor creation, the 150 non-churn tensors and 50 churn tensors were randomly split into training and testing sets. Specifically, 40 churn tensors and 120 non-churn tensors were used as a training set, while the remaining 10 churn tensors and 30 non-churn tensors were used for testing. This 80/20 split enables the model to be trained on a representative sample of the data and evaluated on unseen groups, getting an estimate of its performance in real-world scenarios.

To process these structured tensor inputs, a Convolutional Neural Network (CNN) model was implemented. The CNN architecture was chosen for its ability to identify spatial patterns and hierarchies within structured inputs. In this case, the tensors represent transformed feature relationships, and the CNN model can detect consistent patterns across the matrices that are

associated with churn or non-churn behaviour. To improve training and prevent the model from overfitting, early stopping was used. This technique monitors the validation loss during training and stops the process once the model's performance on unseen validation data stops improving. By doing so, it guarantees that the model remains general and performs well on new, unseen data rather than fitting to the training examples.

The CNN model trained on the tensor representations of groups of clients produced good performance in distinguishing between churn and non-churn groups. Over 10 training epochs, accuracy on the validation split reached up to 100%, with the validation loss steadily decreasing, indicating effective learning without signs of overfitting.

When evaluated on the test set, the model reached an overall accuracy of 90%, correctly classifying 28 out of 30 non-churn groups and 8 out of 10 churn groups. The confusion matrix also displayed a balanced performance in both churn classes, with only a few misclassifications.

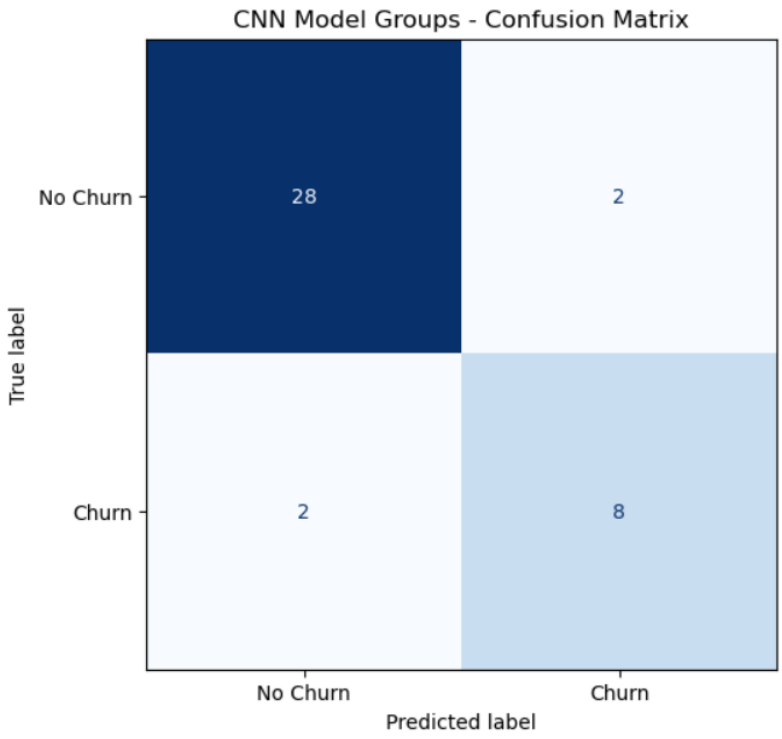

Figure 12 – Confusion matrix of the CNN model

Classification Report on Test Set:

	precision	recall	f1-score	support
0.0	0.93	0.93	0.93	30
1.0	0.80	0.80	0.80	10
accuracy			0.90	40
macro avg	0.87	0.87	0.87	40
weighted avg	0.90	0.90	0.90	40

Figure 13 – CNN model results

These results demonstrate that the structured covariance-based tensors can effectively retain information that differentiates between churn and non-churn groups. By identifying statistical relationships within each group, these representations present a meaningful basis for classification, enabling a model to separate the two classes.

Building upon the tensor representation, the goal is to use these learned signature tensors to refine and individualize churn prediction. Two distinct signature tensors were constructed, one using all training samples labelled with class 0, and the other using all training samples labelled with churn 1. These tensors contain structural signatures specific to each churn class, serving as a macro-level summary of their internal correlations and dependencies.

Previously, the tensors were computed from groups of clients within the training data, capturing local patterns with smaller subsets. In this step, however, the tensors were derived from the full set of training samples for each class to obtain a more global and robust structural representation. This approach reduces the influence of noise and sampling bias, which is important when churn behaviours are subtle and when class distributions are imbalanced.

To use these representations for individual prediction, each client was joined with each class signature, producing two datasets, one where each client's data is combined with the class 0 signature, and another where each individual vector is combined with the class 1 signature.

For training, each example was labelled as 1 if the class label of the sample matched the signature it was combined with, otherwise it was labelled as 0. This essentially trains the model to learn how well a given sample aligns with each class signature. The final training set was created by stacking the two datasets, doubling the size of the original data, and embedding each instance in the context of both class perspectives. This approach allowed the model to act as a kind of similarity function, judging which class signature better matches each sample.

A neural network was then trained on this structural data. Throughout training, both the training and the validation accuracies improved steadily, with validation accuracy stabilizing around 80%, showing the model was learning to identify significant correspondences between

the client sample and the class signature tensors. The training progression suggested the model generalized well without overfitting, even with this expanded representation.

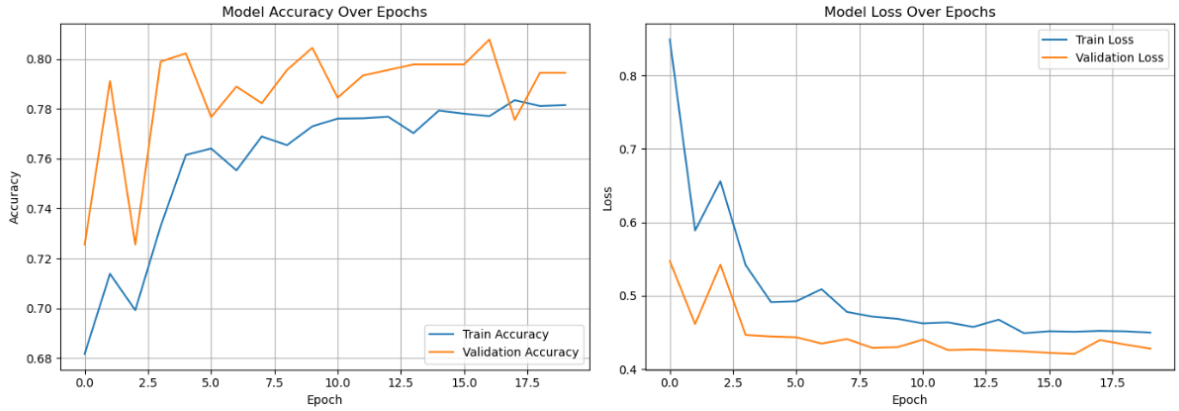

Figure 14 – Model accuracy and loss over epochs

After training, predictions were computed by scoring each test sample against both class signatures independently. These scores were then compared: for a given instance, if the score under the class 1 (churners) signature was higher than under the class 0 (non-churners) signature, the model predicted class 1; otherwise, it predicted class 0. To refine this decision rule, a threshold was optimized to tune the margin between the two scores. This threshold was achieved using a separate validation set specifically held out for this purpose. This approach guarantees that the test set stays untouched during threshold tuning, avoiding data leakage and securing unbiased evaluation of the model's performance. Instead of simply getting the threshold that maximized accuracy, a weighted score was used to better balance the trade-off between precision and recall. Specifically, a weighted combination of accuracy and recall was computed using a weighting factor alpha = 0.8, where accuracy was given a weight of 0.8 and recall a weight of 0.2. This approach was chosen to consider the class imbalance, recognizing that in churn prediction, it´s not only important to achieve high overall accuracy, but also crucial to maintain a good churn recall, assuring that as many true churn cases as possible are correctly identified.

The optimal threshold found was negative (-0.347), indicating that, on average, samples needed to favour class 1 just moderately to be considered churn. This aligns well with the observed score distributions, where scores from class 0 and class 1 signatures showed clear separation, but with enough overlap to justify a nuanced decision boundary.

Finally, when evaluated on the test set, this approach achieved solid performance. The method reached 80% overall accuracy with a strong true negative rate. It also captured 72% recall on the positive class (churners), which is often a critical metric in customer churn prediction contexts, especially in industries like telecommunications, where retaining customers is crucial. The confusion matrix also showed the effectiveness of this approach,

showing that the method predicted well across both classes, accurately identifying churners and minimizing false positives.

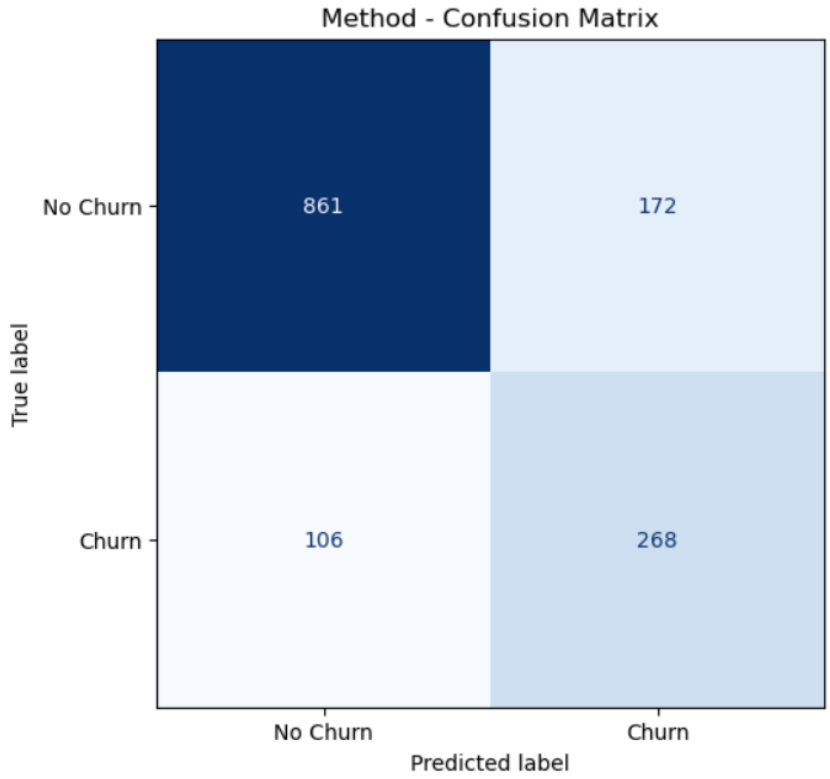

Figure 15 – Confusion matrix of the Proposed Method

Classification Report on Test Set:

	precision	recall	f1-score	support
0	0.89	0.83	0.86	1033
1	0.61	0.72	0.66	374
accuracy			0.80	1407
macro avg	0.75	0.78	0.76	1407
weighted avg	0.82	0.80	0.81	1407

Figure 16 – Proposed Method results

4. RESULTS AND DISCUSSION

This section shows and compares the performance of different models for predicting churn in the telecommunications sector, including standard machine learning techniques, ensemble models, and a new approach. The main objective is to assess the results of each model when identifying customers at risk of churning, avoiding techniques that may introduce bias or lead to misleading conclusions. It is also important to assess the model performance using a span of evaluation metrics, not just accuracy, since metrics as precision and recall provide a more comprehensive understanding of a model's true predictive capability, especially in imbalanced datasets and real-world scenarios. The results are summarized in Fig. 17, which presents the metrics accuracy, recall for class 0 (non-churn), and recall for class 1 (churn).

Methods	Accuracy	Recall (Class 0)	Recall (Class 1)
Logistic Regression	0.8124	0.9013	0.5668
Gradient Boosting	0.8053	0.9032	0.5348
AdaBoost	0.8138	0.9080	0.5535
Voting Classifier	0.8145	0.9080	0.5561
Proposed Method	0.8024	0.8335	0.7166

Figure 17 – Comparison of results

4.1. PERFORMANCE OF TRADITIONAL MODELS

Initial analysis was carried out using Logistic Regression, Gradient Boosting, and AdaBoost. These models are used in churn prediction due to their simplicity and reliability. All models had a good performance in identifying customers who did not churn, with recall values for class 0 above 90%. However, the recall for actual churners (class 1) remained low. Gradient Boosting reached a churn recall of 53.5%, with AdaBoost reaching 55.4% and Logistic Regression slightly higher at 56.7%. These results demonstrate that while the models can predict non-churners effectively, there is a struggle to detect customers who are likely to churn.

Similar observations were made in earlier research. For instance, Prabadevi et al. (2023) evaluated several models and reported that Stochastic Gradient Booster achieved the best performance with an accuracy of 79.1%. However, the recall for churners in many such models tends to be limited, with the best recall reported being only 46%.

4.2. Impact of Oversampling Techniques

One important difficulty in churn prediction is the imbalance between churners and non-churners. In most datasets, non-churners represent a larger portion, causing models to favor predicting the majority class. To handle this issue, the SMOTEENN technique was applied. This approach combines oversampling (SMOTE) with data cleaning (ENN) to remove noisy instances, resulting in more balanced training datasets.

When SMOTEENN was used only in the training set, an impressive improvement in detecting churners was observed. For instance, recall for the churn class increased significantly, with AdaBoost reaching up to 87%. However, this came at the cost of overall accuracy, which dropped to 72%. This trade-off highlights the challenge of improving minority class detection without sacrificing general model performance.

It is also critical to ensure that oversampling techniques like SMOTEENN are applied only to the training data. Applying this technique to both training and test sets can cause data leakage, which happens when the model is exposed to test information during training. As a result, the test data would no longer be truly unseen data, weakening the validity of model evaluation and leading to optimistic performance metrics.

4.3. Ensemble Learning Techniques

An ensemble method, the Voting Classifier, was also assessed. This approach combined the predictive capacities of different classifiers already tested (Logistic Regression, Gradient Boosting, and AdaBoost) to build a more robust model. This classifier combines the predictions from each of these models using a majority voting scheme, which helps to balance the bias and variance of the individual algorithms.

In evaluation, the ensemble model slightly outperformed the individual classifiers, achieving an accuracy of 81.4% and a churn recall of 55.6%. This reveals a small improvement in identifying customers who are likely to churn. However, despite getting a better performance, the improvement was not substantial, suggesting that the base models may already be capturing similar patterns in the data, or that further gains may require more sophisticated techniques.

4.4. Proposed Method

The evaluation of the proposed churn prediction method showed promising results, by identifying churners in a more effective way compared to already established techniques. The model was trained over 20 epochs, and its performance was monitored using accuracy and loss metrics for both training and validation sets.

The created plots of model overall accuracy and loss over epochs showed that both training and validation overall accuracy improved during early epochs, stabilizing at around 78% for training and 80% for validation. This means the model performed well without overfitting. It is also possible to see that the training loss dropped in the first few epochs and then stabilized, while the validation loss shows a similar tendency, maintaining a lower value. This can be seen as a sign that the model is not only fitting the training data but is also capturing meaningful patterns in the data.

The histograms of predicted probabilities from the additional validation set used for threshold selection provide a perception into the model's output distribution before applying the classification threshold. For the validation set, the predicted probabilities show a clear distinction between the two signature groups. Samples when trained with the signature from class 1 (churners) tend to have probabilities heavily concentrated between the values 0.0 and 0.2, meaning the model is confident in assigning low churn probabilities to many samples with this signature. However, there are probabilities that reach values close to 0.8, showing that the model is also confident in assigning high churn probabilities to certain samples. On the contrary, probabilities for samples trained with signature 0 (non-churners) are more distributed, typically from values between 0.5 to 0.9, revealing a higher level of uncertainty in the signature matching for these samples. Based on this analysis, it is possible to say that the samples associated with signature 1 are more informative. Although they show a higher concentration in lower values (which is expected given the imbalance in the data), these probabilities cover a broader range of values, reflecting the model's confidence in separating samples with different churn risks.

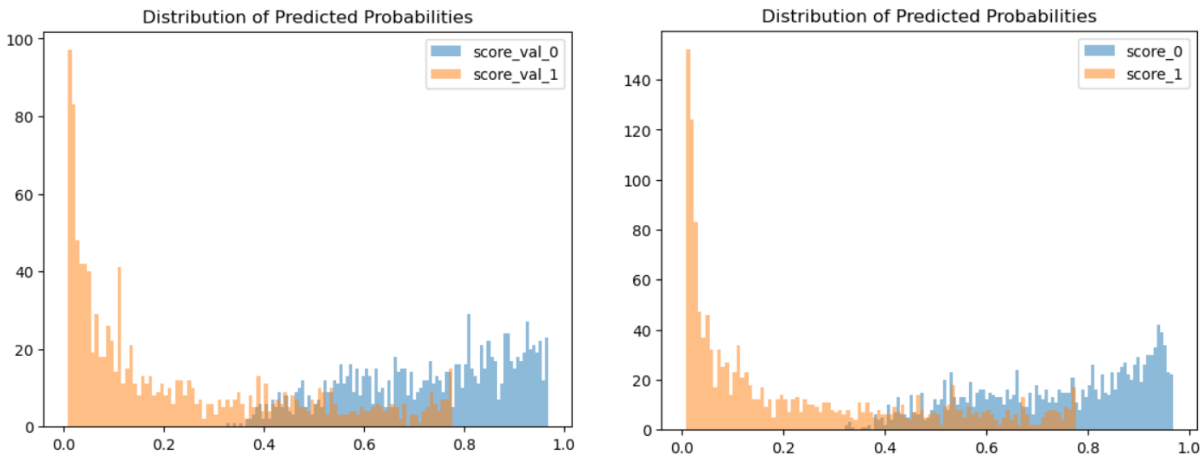

Figure 18 – Histograms of predicted probabilities (validation and test)

It is important to note that, while there is visible separation, there is also a region of overlap where both groups intersect. This overlap highlights the necessity of using a well-defined

classification threshold to separate churners and non-churners effectively. In this case, the selected threshold was -0.347, which favors the churn class in the model, allowing for a more aggressive detection of potential churners while controlling, at the same time, false positives.

Additionally, the predicted probability distributions for the test set are identical to those observed in the validation set. This consistency shows that the model is stable across different data splits and generalizes well to unseen data, supporting its practical applicability for churn prediction in real scenarios.

The evaluation on the test set (unseen data) demonstrates the effectiveness of the proposed method. As shown in the confusion matrix, the model identified 268 churners out of 374, reaching a recall of 72% for the churn class (class 1). There is a notable improvement over other approaches, which achieved churn recalls in the range of 53% to 56%. Additionally, this approach maintained a high recall of 83% and precision of 89% for the non-churn class (class 0) and achieved an overall accuracy of 80%, which is comparable to the overall accuracy obtained by other methods.

It is important to note that, unlike some of the earlier methods and other authors' methods, this proposed method did not rely on any oversampling techniques to assist classification. Instead, it had to handle the inherent class imbalance during training, making the achieved recall even more significant. This presents the method’s ability to learn meaningful patterns related to churn without needing artificial data balancing techniques, reducing the risk of overfitting to synthetically generated samples.

The proposed method demonstrates good improvements when identifying customers at risk of churning compared to conventional machine learning models, while keeping high overall performance and stability. This indicates its applicability for real-world scenarios in churn prediction within the telecommunications sector, supporting proactive retention efforts with a reliable data-driven foundation.

5. FUTURE WORK

Although telecommunications companies already have their churn prediction models established, there is an increasing interest in researching new and innovative methods to improve proactive retention efforts. One possible approach involves using sentiment analysis on call center interactions to detect early signs of churn risk and understand customer concerns. This represents an innovative method to proactively act on retention using audio data.

In collaboration with KPMG, a pilot study was conducted using artificially generated call recordings (not real customer calls, due to the restricted and private nature of such data) to evaluate the feasibility of analysing sentiment during customer support interactions. The objective was to evaluate the emotional tone of customers during calls, evaluate their potential risk of churn, and typify the reasons for contacting the call center.

It is important to note that, due to the use of artificial data rather than real customer calls, it was not possible to validate the results. However, this exploration provides a foundation for developing tools that can support the respective teams in identifying customers with a high risk of churn, supporting faster and more personalized retention actions based on the sentiment analysis and the call reasons of each client.

5.1. SENTIMENT ANALYSIS FOR CHURN DETECTION

In this context, audio sentiment analysis is defined by the analysis of customer emotions directly from their voices during calls instead of only using text from transcriptions. It focuses on how something is said, not just on what is said, analyzing the customer's speed of speaking, voice pitch, and tone to detect if the customer is frustrated, satisfied, or neutral during interactions with the call center.

In this work, sentiment analysis was applied to both customer and call center agent, making it possible not only to assess the customer's emotional state for churn prediction but also to support employee quality control by monitoring the emotional signals of agents during calls.

The methodology for using audio sentiment analysis to predict churn was implemented in several stages. This process can be performed in multiple languages, as the transcription system allows specifying the language during processing to adapt to different call center environments.

The first step consists of transcribing customer service calls using speech-to-text services with speaker diarization to create transcriptions with structure, labels, and timestamps. This guarantees that each speaker's segments are correctly identified, allowing alignment between the audio and specific speaker turns in the conversation.

In the second step, audio sentiment analysis is performed by analyzing the audio signals instead of relying only on textual sentiment. Some acoustic features, such as average pitch, volume, and speech rate, are extracted for each speaker. Based on these extracted features, the emotional state of each speaker is classified into categories such as 'Frustrated', 'Energetic', 'Calm', and 'Assertive'. This classification is determined by using thresholds related to these emotions, allowing the detection of emotional signals important for churn prediction.

The third step consists of analyzing the call reason and evaluating churn risk using the transcription combined with the extracted emotional information. A large language model (LLM) is prompted to summarize the reason for the call, categorize the type of request, evaluate the sentiment for both customer and call center agent, identify customer preferences, and estimate the likelihood of churn. The LLM can also identify the main action required after the call, promoting follow-up actions and providing insights into retention strategies.

By integrating analysis using LLMs, vocal emotion detection, and structured call reason categorization, this approach produces an improved dataset for predictive churn modeling. Combining sentiment analysis with the transcription allows a better understanding of customer behavior, supporting early churn risk detection and the development of proactive retention strategies.

With this methodology, the company can assess churn risk at an earlier stage, get more time to act on retention strategies, and maintain customer relationships in a proactive way. While detecting dissatisfaction signals and understanding customer needs during calls, the company can intervene before customers decide to leave, offering tailored solutions that directly address concerns. This not only improves customer retention but also supports operational efficiency by prioritizing cases requiring immediate action, enabling the company to retain valuable customers while improving overall service quality.

6. CONCLUSIONS

This thesis aimed to explore the increasingly relevant challenge of predicting customer churn in the telecommunications sector, accentuating the need for innovative methods and approaches to support proactive retention strategies. The main objectives of this work were: (i) to analyse the performance of conventional machine learning models in predicting churn in imbalanced datasets, and (ii) to create and evaluate a new tensor-based method to improve churn prediction.

Predicting customer churn represents an important challenge in the telecommunications industry, where intense competition and high operational costs keep bringing more awareness to companies about the need to retain customers.

This research confirmed that traditional machine learning models, including Logistic Regression, Gradient Boosting, and AdaBoost, while being effective in achieving high overall accuracy, demonstrate limitations in identifying customers likely to churn. The different models achieved a lower recall for the churn class, with values below 60%, revealing the problem of dealing with class imbalance in churn datasets.

To address this challenge, the SMOTEENN resampling technique was applied, increasing the recall for churners to above 80%, at the cost of reducing the overall accuracy of the model and having potential risks of data leakage if this technique is not applied correctly (e.g., if artificial samples are deployed on training and testing).

Considering these challenges, this research proposed a new tensor-based feature engineering approach to churn prediction that leverages covariance structures, capturing complex relationships and patterns often missed by traditional models. This approach achieved promising results, with a churn score of 72% for the churn class while maintaining an overall accuracy of 80%, demonstrating its potential to outperform traditional models in detecting customers at risk of churning, without relying on resampling methods.

Additionally, this thesis explored the use of sentiment analysis on customer interactions. By analysing customer emotional states and the content of the call, companies can gain insights into dissatisfaction signals that are early indicators of churn risk. This approach leverages another type of data, such as the content of calls to the call center, to identify churn at an earlier stage than usual, giving the company more time to act with the objective of retaining the customer.

Finally, the findings of this work demonstrated that while traditional models provide a baseline for churn prediction, advanced approaches such as the proposed tensor-based method can enhance the identification of at-risk customers in the telecommunications sector. As future work, the integration of sentiment analysis from customer interactions can improve churn prediction capabilities by capturing emotional signals. While supporting earlier

detection and enabling proactive and personalized retention actions, these methods can help telecommunications companies maintain customer relationships, reduce losses related to churn, and secure revenue stability in an increasingly competitive market.

7. REFERENCES

Ahmad, A. K., Jafar, A., & Aljoumaa, K. (2019). Customer churn prediction in telecom using machine learning in big data platform. *Journal of Big Data, 6*, Article 28. https://doi.org/10.1186/s40537-019-0191-6

Altalhan, M., Algarni, A., & Turki-Hadj Alouane, M. (2025). Imbalanced Data Problem in Machine Learning: A Review. *IEEE Access, 13*, 13686–13699. https://doi.org/10.1109/ACCESS.2025.3531662

Blastchar, D. (2017). Telco Customer Churn [Data set]. *Kaggle*. https://www.kaggle.com/datasets/blastchar/telco-customer-churn

Carriero, A., Luijken, K., de Hond, A., Moons, K. G. M., van Calster, B., & van Smeden, M. (2025). The Harms of Class Imbalance Corrections for Machine Learning Based Prediction Models: A Simulation Study. *Statistics in Medicine, 44*(3–4), e10320. https://doi.org/10.1002/sim.10320

Chen, W., Yang, K., Yu, Z., Shi, Y., & Chen, C. L. P. (2024). A survey on imbalanced learning: Latest research, applications and future directions. *Artificial Intelligence Review, 57*(137). https://doi.org/10.1007/s10462-024-10759-6

Cheng, X. (2024). A comprehensive study of feature selection techniques in machine learning models. *Artificial Intelligence and Digital Technology*, vol 1(1), pages 65-78. https://doi.org/10.70088/xpf2b276

Edwine, N., Wang, W., Song, W., & Ssebuggwawo, D. (2022). Detecting the Risk of Customer Churn in Telecom Sector: A Comparative Study. *Mathematical Problems in Engineering, 2022*, Article 8534739. https://doi.org/10.1155/2022/8534739

Ginson, M., Sousa, D., Arias, J. M. A., Watters, L., & Labio, D. D. (2024). From telco to techco: Towards tomorrow's telecom. KPMG International.

Lalwani, P., Mishra, M. K., Chadha, J. S., & Sethi, P. (2022). Customer churn prediction system: A machine learning approach. *Computing*, 104(2), 271–294. https://doi.org/10.1007/s00607-021-00908-y

Li, Y., Yang, Y., Song, P., Zhang, W., & Yang, J. (2025). An improved SMOTE algorithm for enhanced imbalanced data classification by expanding sample generation space. *Scientific Reports, 15,* Article 23521. https://doi.org/10.1038/s41598-025-09506-w

Liégeois, R., Santos, A., Matta, V., Van de Ville, D., Sayed, A.H. (2020). Revisiting correlation-based functional connectivity and its relationship with structural connectivity. *Network Neuroscience*, 4(4), 1235–1251. https://doi.org/10.1162/netn_a_00166

Liu, X., Xia, G., Zhang, X., Ma, W., & Yu, C. (2024). Customer churn prediction model based on hybrid neural networks. *Scientific Reports, 14*, Article 30707. https://doi.org/10.1038/s41598-024-79603-9

M, N., Daniel, E., Durga, S., & Seetha, S. (2025). Explainable multi-stage churn prediction using graph neural network in telecom sector. In *Proceedings of the 2025 8th International Conference on Trends in Electronics and Informatics (ICOEI)* (pp. 1100–1105). IEEE. https://doi.org/10.1109/ICOEI65986.2025.11013438

Mujahid, M., Kına, E., Rustam, F., Ashraf, I., Siddiqui, S. A., & Choi, G. S. (2024). Data oversampling and imbalanced datasets: An investigation of performance for machine learning and feature engineering. *Journal of Big Data, 11*, Article 87. https://doi.org/10.1186/s40537-024-00943-4

Poslavskaya, E., & Korolev, A. (2023). Encoding categorical data: Is there yet anything 'hotter' than one-hot encoding? (arXiv:2312.16930). *arXiv*. https://doi.org/10.48550/arXiv.2312.16930

Prabadevi, B., Shalini, R., & Kavitha, B. R. (2023). Customer churning analysis using machine learning algorithms. *International Journal of Intelligent Networks, 4*, 145–154. https://doi.org/10.1016/j.ijin.2023.05.005

Saha, S., Saha, C., Haque, M. M., Alam, M. G. R., & Talukder, A. (2024). ChurnNet: Deep learning Enhanced Customer Churn Prediction in Telecommunication Industry. *IEEE Access, 12*, 4471–4484. https://doi.org/10.1109/ACCESS.2024.3349950

Sikri, A., Jameel, R., Idrees, S. M., & Kaur, H. (2024). Enhancing customer retention in telecom industry with machine learning driven churn prediction. *Scientific Reports, 14*, Article 13097. https://doi.org/10.1038/s41598-024-63750-0

Wagh, S. K., Andhale, A. A., Wagh, K. S., Pansare, J. R., Ambadekar, S. P., & Gawande, S. H. (2024). Customer churn prediction in telecom sector using machine learning techniques. *Results in Control and Optimization, 14*, Article 100342. https://doi.org/10.1016/j.rico.2023.100342

Xia, X., Zeng, L., & Yu, R. (2018). HMM of Telecommunication Big Data for Consumer Churn Prediction. *Proceedings of the 2018 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation* (pp. 1903 1910). IEEE. https://doi.org/10.1109/SmartWorld.2018.00319

Zhu, W., Qiu, R., & Fu, Y. (2024). Comparative study on the performance of categorical variable encoders in classification and regression tasks (arXiv:2401.09682). *arXiv*. https://doi.org/10.48550/arXiv.2401.09682

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa