SCHOOL OF SCIENCE & TECHNOLOGY

Corporate Capstone Project

Boosting Customer Retention for a Telco Company

The authors Rafaella Jereissati, Mohamed Atwa, Yupeng Chen, Alessio Mastropietro, Bernardo Antunes dos Santos and Max Uebele declare:

This capstone project is guided by strict ethical principles, ensuring confidentiality, compliance, integrity, and respect for all stakeholders. We safeguard proprietary information, adhere to data protection laws, conduct research with honesty and fairness, and uphold professionalism in all collaborations. By embedding these values into our work, we affirm our commitment to transparency and responsible execution, ensuring a project built on trust and integrity.

Master of Science in Business Analytics and Data Science
IE School of Science and Technology

Corporate Sponsor: BCG X
Assigned Mentor: Gustavo de Carvalho Homem

Boosting Customer Retention for a Telco Company

Corporate Capstone Project

by

Rafaella Jereissati

Mohamed Atwa

Yupeng Chen

Alessio Mastropietro

Bernardo Antunes dos Santos

Max Uebele

Corporate Sponsor:	BCG X
Assigned Mentor:	Gustavo de Carvalho Homem

SCHOOL OF SCIENCE & TECHNOLOGY

Executive Summary

Predicting customer churn is crucial for telecom providers aiming to retain high-value customers and drive revenue growth. This study develops a machine learning framework that leverages distinct types of customer data as user activity, demographic attributes, and text from complaint forms, uncovering key signals of dissatisfaction that contribute to churn risk. By systematically linking textual insights with structured behavioral data, we aim to refine customer retention strategies while maintaining model interpretability and scalability.

Our model employs a three-step pipeline. The first step categorizes complaints using feature engineering and state-of-the-art NLP models, enhanced by LLMs for deeper semantic recognition. The second step predicts churn probability through ensemble machine learning techniques, incorporating structured and unstructured features. The final step maps extracted text patterns to actionable interventions, ensuring that at-risk customers are not merely retained but repositioned within the company's service offerings based on enhanced, data-driven segmentation.

Further, we propose a revenue-maximizing framework to uncover trends in price sensitivity and lifecycle stages and to pair churn insights with effective retention measures. Unlike traditional churn mitigation strategies that focus solely on risk reduction, our approach seeks to leverage churn prediction as a means of customer re-engagement and product optimization. Hence we aim to identify opportunities for tailored upselling and dynamic discount allocation, utilizing the full range of available telecommunication services. Given the absence of direct pricing data, we construct a comparative business case using real-world benchmarks, establishing cost-effectiveness thresholds to evaluate potential interventions and long-term revenue impact.

A central methodological aspect is ensuring the model's resilience against overreliance on specific variables, such as phone usage behavior, which may implicitly correlate with plan characteristics rather than true churn risk. To enhance generalizability, we incorporate multiple layers of robustness, enabling dynamic recalibration upon deployment with new data. The solution follows a platform-agnostic design, employing a virtual environment to encapsulate dependencies and ensure reproducibility.

Future work may explore reinforcement learning-based intervention strategies and expanded model explainability techniques to further enhance decision-making precision. With the chosen approach, we shed light on the path towards increasing long-term customer value by aligning intervention strategies with business growth objectives.

Contents

Nomenclature

Abbreviations

Abbreviation	Definition
LLM	Large Language Model
ML	Machine Learning
NLP	Natural Language Processing

1
Project Description

1.1. Introduction

The modern business landscape in service and technology sectors is increasingly characterized by subscription-based models rather than singular sales and finite contracts. Industry leaders such as Microsoft, Apple, and Google have successfully embedded themselves into everyday life by recognizing a fundamental shift in consumer behavior: customers are not merely one-time buyers but long-term assets. The traditional model of selling software as a one-time purchase, such as CD-ROM packages, has been replaced by integrated ecosystems of hardware and software, ensuring continuous customer engagement.

However, sustaining customer loyalty is not merely about creating dependency. Companies must continuously evolve their offerings to align with user needs while also navigating regulatory constraints and market competition. In this paper, we analyze a specific application of this paradigm: customer retention within the telecommunications industry — a sector characterized by regulatory pressure (European Union, 2022), shrinking profit margins, and increasing consumer demand for high-speed connectivity at minimal cost.

Telecommunications services are foundational to modern society, extending beyond the provision of SIM cards to the development of comprehensive connectivity solutions. A customer-centric approach is therefore essential, not only for mitigating churn but also for fostering long-term trust. Embracing a customer-centric business model has been shown to enhance customer satisfaction, drive revenue growth, and improve operational efficiency (Sankur et al., 2024).

Trust is a key factor in service expansion: will a customer subscribe to additional products from a provider that struggles to deliver stable mobile internet at a reasonable price? The answer is evident. Yet, customer service should not merely act as a reactive mechanism for addressing complaints; it should serve as the foundation for customer engagement and loyalty. Many providers fail to meet modern user expectations, missing valuable opportunities for sustained customer relationships.

This study explores (1) the potential revenue losses resulting from inadequate churn prevention strategies and (2) the effectiveness of automated approaches for analyzing, reacting to, evaluating, and adjusting service offerings based on customer data. Leveraging state-of-the-art machine learning models, we analyze real-world customer data and unstructured text, demonstrating how artificial intelligence can enhance scalability and optimize customer retention.

1.2. Methodology

Our approach follows four main steps: exploratory data analysis, text analysis, predictive modeling, and retention strategy optimization. Each phase builds upon the previous one, ensuring that our analysis is grounded in both the data and the business realities of the telecommunications industry.

We begin by thoroughly assessing the dataset within the framework of industry-specific characteristics. Instead of viewing customer data in isolation, we examine it through the lens of service structures, contract dynamics, and known behavioral patterns in telecom churn. This provides a foundation for understanding usage trends, forming initial hypotheses on segmentation, and identifying potential churn drivers before moving into more detailed analysis.

Building on these insights, we apply natural language processing (NLP) to systematically analyze unstructured complaint data. Standard preprocessing techniques refine the text, while large language models (LLMs) enable deeper semantic analysis. This step allows us to extract dissatisfaction patterns that are not immediately apparent in structured data, translating qualitative insights into measurable and iterable features.

With both structured and text-derived features in place, we develop a machine learning pipeline to predict churn probability. The model is trained using a combination of behavioral attributes and extracted complaint insights, ensuring that it captures both explicit and implicit risk factors. Feature selection is performed rigorously to avoid overfitting and improve generalizability. The resulting predictions quantify churn risk at the individual level, providing a structured output that informs decision-making.

Finally, we design a dynamic segmentation and recommender approach that adapts retention strategies to different risk profiles. Instead of applying generic churn mitigation measures, we define intervention rules based on customer engagement levels and expected value. This allows for targeted actions that optimize both retention and profitability, ensuring that efforts are aligned with the company's business objectives.

2

Exploratory Data Analysis

2.1. Scope and Limits of Datasets

A comprehensive churn analysis requires both behavioral patterns and qualitative insights. The dataset provides a detailed look at customer history, service usage, and complaints, allowing us to examine potential churn drivers. However, key limitations—such as the absence of pricing details, past interventions, and long-term usage trends—mean that certain assumptions are necessary to interpret the findings.

For this analysis, we were provided with three major datasets: customer_cata containing demographic details and contract information; (monthly(activity data, capturing service usage and engagement metrics; and complaint records, offering unstructured insights into customer dissatisfaction. Together, these datasets allow us to assess behavioral patterns and potential churn drivers, though certain limitations must be considered when interpreting the results.

Customer records outline the basic profile of 10,000 individuals, including their contract duration, plan type, and whether they churned in early 2022. While these records establish a foundation for segmentation, they offer little insight into past company actions, making it difficult to distinguish between voluntary and externally influenced churn.

Monthly activity logs track data and phone usage over the course of 2021, as well as whether customers engaged with the provider's app. This allows for an assessment of usage trends leading up to churn but does not capture real-time reactions to service disruptions, price changes, or targeted retention efforts. Additionally, since the dataset spans only a single year, it assumes that 2021 behavior is indicative of broader customer habits.

Complaints add an unstructured layer to the analysis, with 3,500 documented issues categorized by type. While they highlight sources of dissatisfaction, the data does not indicate resolution times, company responses, or the impact of these issues on customer decisions.

To enable a more holistic analysis, all datasets were merged through customer IDs, linking long-term account characteristics with monthly behavior and complaints. This structure makes it possible to trace patterns between engagement, dissatisfaction, and churn, but also introduces gaps. Without explicit timestamps for company interventions, we cannot determine whether churners left due to unresolved issues, lack of incentives, or external market factors.

These constraints underscore the need for a careful balance between data-driven insights and reasonable assumptions.

2.2. Characteristics of Customer Base

The dataset reflects a broad and diverse customer base, spanning different age groups, service plans, and usage patterns. While individual behavior varies significantly, some characteristics remain consistent across the board, forming an essential backdrop for our later analysis.

A key aspect of the dataset is its strong focus on prepaid customers, who make up 70.4% of the sample. Postpaid contracts account for 19.7%, while the remaining 9.9% of customers follow a pay-as-you-go model. This distribution already shapes the kind of insights we can expect: prepaid users tend to have more flexibility in their engagement, whereas postpaid customers may represent a more stable revenue stream but are also bound by contract terms that could influence their churn behavior. Given this dominance of prepaid plans, it will be important to assess whether their churn drivers differ fundamentally from those of contract-based customers.

On average, customers have been with the company for about 5.2 years, though individual tenure varies widely, from newcomers with just one year of history to long-term customers who have been loyal for a decade. Age distribution is similarly broad, with a median age of 36 years, spanning from as young as 12 to over 60. These long-standing relationships suggest that many customers have had repeated interactions with the company, but the dataset does not contain explicit records of past retention efforts or contract changes, leaving a gap in understanding how previous experiences influence churn.

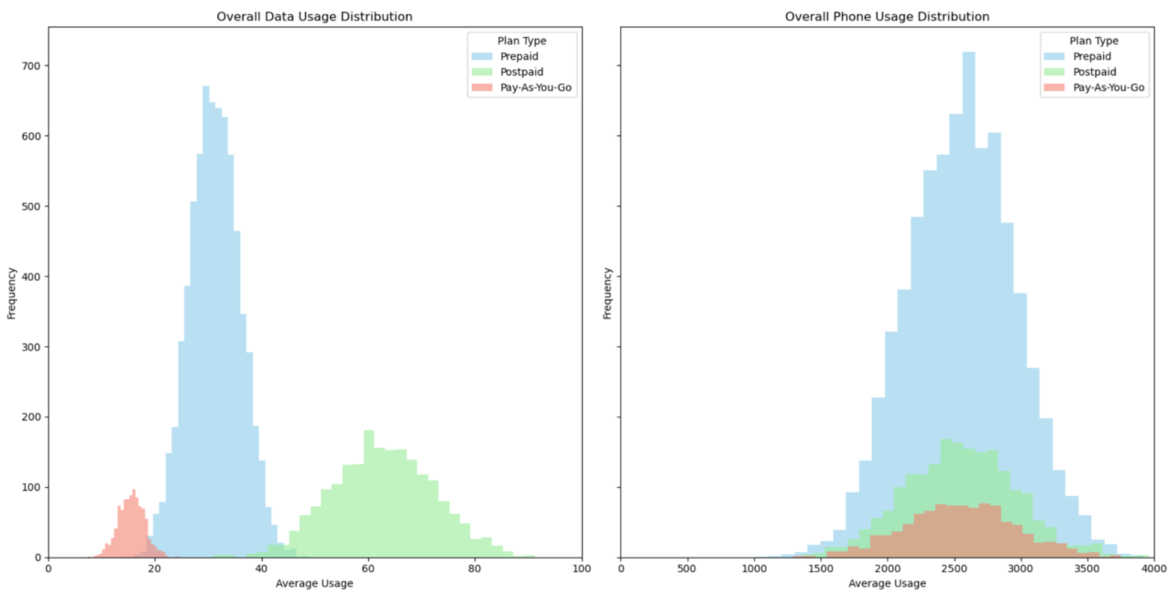

Figure 2.1: Average Data and Phone Usage across Plans

Looking at service usage, customers consume an average of 35.9 GB of data per month, with substantial variation across the user base. Some customers barely use mobile data, while others exceed 40 GB per month, suggesting very different levels of reliance on the service. Phone usage averages 2,554 minutes per month, though again, there are users who barely make calls and others who rely heavily on voice services. 74.9% of customers actively use the provider's app, showing a high degree of digital interaction across all plan types.

It is remarkable that data usage differs clearly by plan type, whereas phone usage appears to be similar across all customer groups. This suggests that data consumption may be the primary cost driver, while phone services are likely covered by a flat-rate structure, making them less relevant to spending behavior. If true, this could have implications for churn, particularly if high-data users are more price-sensitive than those with lower usage levels.

To assess whether these patterns are meaningful in the context of customer retention, the next step is to compare them between churners and non-churners. If significant differences emerge, they could provide early indications of what separates retained customers from those who leave. This will be examined in the following section.

2.3. Churn Characteristics and Behavioral Patterns

While the customer base as a whole provides insights into sociodemographic structures and plan preferences, behavioral patterns are far more telling when it comes to assessing churn risk. A closer examination reveals that customers do not leave at random—certain traits and engagement levels set churners apart from those who stay.

A total of 30.6% of customers canceled their service within the first trimester of 2022. Notably, churn is not strongly tied to plan type, as prepaid, postpaid, and pay-as-you-go customers all exhibit similar churn rates. Likewise, age and tenure distributions remain consistent between churners and non-churners, implying that these factors alone are not reliable indicators of customer stability.

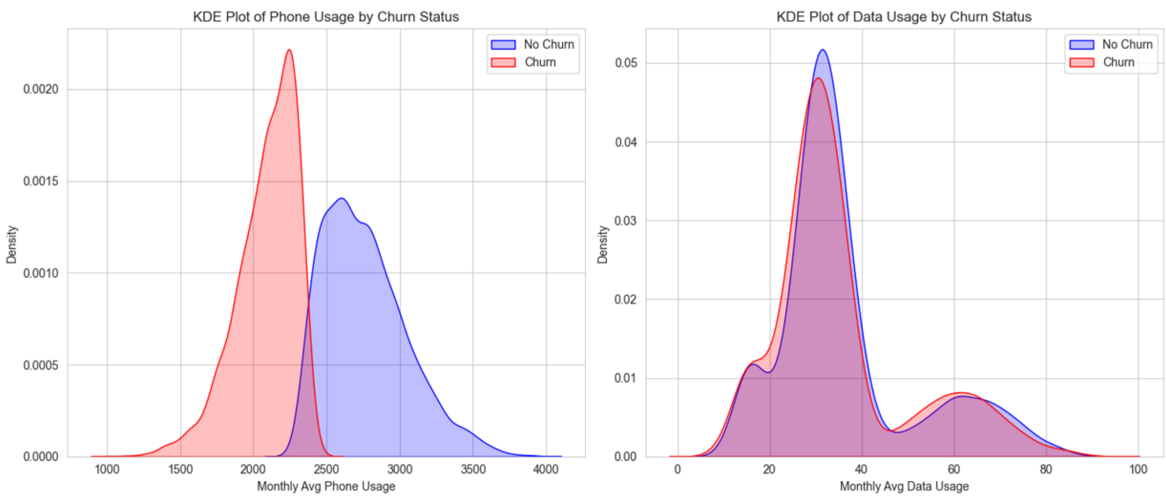

Figure 2.2: Density of Churners and No-Churners per Average Phone and Data Usage

The most striking behavioral difference between churners and retained customers appears in phone usage. Churners average 2,087 minutes per month, while non-churners reach 2,761 minutes. This suggests that low and predictable phone usage might indicate lower attachment to the service, making it a valuable early warning signal. Data consumption, in contrast, does not show a clear distinction—churners and retained customers use similar amounts of mobile data, and usage levels remain relatively stable leading up to cancellation. This aligns with expectations, as mobile data usage often continues even in the presence of service dissatisfaction, whether through essential functions or background processes. The density plot of phone usage further illustrates this divergence, showing a concentration of churners in the lower consumption range.

Customer complaints emerge as one of the most decisive factors in predicting churn. Clients who have filed a complaint have a 70% probability of churning, compared to just 9% among those who have not. While dissatisfaction does not always result in immediate departure, the sheer proportion of churning customers within this group underscores the need for effective resolution strategies. Breaking complaints down by category reveals an even more concerning insight—every single customer who lodged a promotional issue complaint ended up churning. Unlike other complaint categories such as billing issues, connection problems, unauthorized charges, and service disruptions, which all show a similar churn rate of around 67%, dissatisfaction with promotions appears to be a definitive dealbreaker. This suggests that grievances related to misleading or disappointing promotional offers may leave customers with no perceived path to resolution, making them highly predictive of churn. A bar chart of churn rate by complaint category highlights this stark contrast.

Beyond complaints, engagement trends provide further nuance. App usage, while commonly cited as a marker of digital interaction, does not appear to correlate with churn. Both churners and non-churners engage with the provider's app at similar rates, suggesting that app adoption alone does not equate to customer satisfaction. However, within certain complaint categories, notable distinctions appear—customers filing connection problem complaints display a wider range of phone usage patterns than those lodging other complaints, indicating that connectivity issues affect a more diverse user base.

One particularly important observation is that churners do not exhibit a noticeable decline in phone or data usage before canceling their service. Monthly activity levels remain steady throughout 2021, meaning that there is no slow disengagement phase—customers continue their usual behavior right up until the moment they decide to leave. This reinforces the idea that telecom churn is often a decisive event rather than a gradual drift and suggests that retention efforts must take place before dissatisfaction escalates to a cancellation decision. A line chart showing monthly data and phone usage trends confirms this stability.

While these patterns highlight key behavioral signals of churn, they do not fully explain the drivers behind customer dissatisfaction. The structured data provides clear numerical differences, but the most direct indicators of frustration and unmet expectations lie within the unstructured text of customer complaints. To deepen the analysis, the next step involves extracting meaningful patterns from complaint texts, identifying recurring themes, and linking them to churn behavior. This will allow for a more comprehensive understanding of why customers leave and how their concerns could be addressed more effectively.

3

Automation of Complaint Analysis

3.1. Model Selection

A core component of our dataset consists of customer complaint forms, which convey crucial information about the nature of reported issues, their severity, and whether further action is necessary. However, these complaints are highly unstructured, often lengthy, and include both redundant and unique details. Extracting standardized, reliable insights from them is challenging due to the variability in human expression. Additionally, computational constraints impose a trade-off: more sophisticated models that achieve high accuracy may be too slow for practical deployment at scale.

Given these constraints, we assessed various model families and iteratively tested their effectiveness. Initially, we experimented with traditional text-processing techniques such as TF-IDF vectorization and sentiment scoring. These methods provided speed and interpretability but were insufficient for tasks requiring contextual understanding, such as detecting negation, implicit frustration, or indirect requests for assistance.

To improve extraction quality, we explored classical Natural Language Processing (NLP) models, including Named Entity Recognition (NER) and dependency parsing. These approaches enhanced our ability to structure complaint data but struggled with context-dependent phrasing and the dynamic nature of customer grievances. Literature has shown that such models can be useful for structured text extraction but often require extensive rule-based fine-tuning (Manning et al., 2008).

To overcome these limitations, we turned to transformer-based Large Language Models (LLMs), known for their ability to handle complex textual structures and capture contextual meaning. Specifically, we tested BERT, RoBERTa, and T5-based architectures, as these models have been widely used for various text classification and language understanding tasks (Devlin et al., 2018; Liu et al., 2019). While these models significantly improved classification accuracy, their inference times made them impractical for real-time complaint analysis. Summarization models such as BART and PEGASUS were also considered for condensing long complaints into key highlights, but they often failed to preserve crucial details necessary for resolution tracking.

3.2. Text Processing and Key Phrase Extraction

After multiple iterations, we developed a hybrid solution: lightweight classification models for rapid filtering, combined with transformer-based LLMs for deeper linguistic analysis. This approach balances efficiency and accuracy, ensuring that critical complaint elements are captured without excessive computational overhead.

A major challenge in automating complaint analysis was extracting meaningful and distinct phrases from lengthy, unstructured texts. Customer complaints frequently contained redundant information, such as standard service acknowledgments, that added little value for assessing issue resolution. To address this, we first preprocessed the text by replacing user-specific identifiers with generic placeholders to prevent biases in similarity detection. Similar strategies have been employed in text anonymization efforts (Neudecker et al., 2017).

To structure the complaint texts, we segmented them into smaller sub-phrases using punctuation marks as delimiters. We then classified these phrases into two categories: general phrases, which appeared across multiple complaint types, and category-specific phrases, which were unique to a particular issue. To avoid excessive noise, general phrases were retained only if they appeared in a moderate proportion of complaints. Initially, we attempted lemmatization to standardize variations in wording, but this introduced distortions in meaning, particularly for negations (e.g., "not able to resolve" being transformed into "able to resolve"). Given these risks, we abandoned lemmatization in favor of direct phrase classification.

3.3. LLM-Based Text Analysis

The core of our complaint analysis pipeline is a transformer-based LLM optimized for structured insight extraction. Given the ambiguity and variability in customer complaints, traditional statistical models struggled to capture sentiment, negation, and implicit dissatisfaction.

Initially, we explored sentiment scoring as a means of inferring issue resolution, under the assumption that resolved complaints might exhibit more positive sentiment. However, as previous research has shown (Mohammad, 2016), sentiment-based approaches can be misleading, as strong emotional wording does not necessarily correlate with resolution or dissatisfaction. Similarly, direct resolution detection using simple classifiers led to frequent misinterpretations, particularly when complaints contained negations (Morante and Blanco, 2012).

Given these challenges, we selected a prompt-based LLM for structured classification, fine-tuned to assess issue resolution, ongoing concerns, and customer demands. This approach follows recent advancements in zero-shot and few-shot classification methods, where text-to-text models outperform traditional classifiers in tasks requiring nuanced language interpretation (Chung et al., 2022).

For issue resolution detection, we designed a structured prompt:

> "Was the issue resolved? Only answer 'Yes' or 'No'. Context: {complaint excerpt}"

This formulation significantly reduced errors in interpreting negations and improved classification reliability.

3.4. Feature Creation

To ensure that extracted complaint insights were structured in a meaningful and actionable way, we fine-tuned a language model using labeled data from our complaint records. This allowed us to move beyond general sentiment analysis and instead extract specific indicators of customer dissatisfaction and issue resolution. The feature creation process was divided into three key stages.

First, we assessed whether an issue described in a complaint had been resolved. This step aimed to differentiate between customers whose problems had been addressed and those who remained dissatisfied.

Second, for complaints classified as unresolved, we determined whether the issue was still ongoing. Many grievances refer to past events, such as temporary service disruptions, that have since been restored. Distinguishing between these cases and active, unresolved issues allowed us to refine churn risk predictions.

Third, we identified cases where customers explicitly demanded corrective action, particularly refunds or compensation. Customers who go beyond merely reporting an issue and actively seek financial redress may be at a heightened risk of dissatisfaction. By incorporating this information into our dataset, we were able to create a structured representation of customer grievances, making it possible to analyze their influence on churn more effectively.

The extracted features included:

- **issue_resolved**: Whether the complaint was classified as resolved.
- **ongoing_issue**: Whether the issue described in the complaint was still ongoing.
- **no_customer_demand**: Whether the customer explicitly demanded a refund or compensation.

These features provided a structured foundation for integrating complaint analysis into our churn prediction models. By systematically distinguishing between resolved and unresolved complaints, identifying whether issues persisted, and detecting direct customer demands, we were able to extract valuable insights from unstructured text without relying on general sentiment scores or traditional classification techniques.

4

Churn Prediction Model

4.1. Initial Analysis and Data Cleansing

The first process of data analysis was to examine the data and data cleaning for the datasets used in the telecom customer churn project. Key actions undertaken included:

Data Type Assessment and Correction: First checks showed that some fields contained wrong data for example dates. Columns such as month, birth_date, and join_date were cast as datetime type so as to enable proper analysis of time series. Null Value Inspection: An exhaustive check for null values in all the datasets (activity, customer details, and complaints) revealed that there are none of them, hence the data is complete. Feature Engineering: New features were systematically engineered to enhance predictive power. That is, features like data usage, phone usage, the age of the customer at the time of joining the plan, and customer tenure (in years on the plan) were created to gain a better understanding of the customer's behavior.

Categorical Data Handling: One-Hot Encoding (OHE) was used on categories like plan_type and month to change them into 0 and 1 for use in the machine learning models to avoid having two columns with the same name in the combined dataset.

Dataset Integrity and Merging: Checks were made to ensure that all customer IDs from the activity dataset were present in the customer dataset before merging on the customer_id index. Next, the data was combined and prepared for analysis, and there was no way that there was duplication of data or misalignment of the data.

Complaint Data Processing: The textual complaints were parsed systematically using regular expressions to identify the key components of the complaint, the issue, the action required, and the general language used. The components were then scored to determine their consistency and then normalized for further NLP analysis.

These structured preprocessing steps show how it is important to clean data and engineer features so that the analysis is correct and reliable. Data types were corrected, all categorical variables were handled carefully, all complaint text data was systematically extracted and structured, and all merging operations were done thoroughly. Thus, the problem of ensuring the correctness of the primary data has been solved to the maximum, which is why all the above-mentioned factors have been taken into account right from the beginning. Hence, the dataset was ready for further analysis, which included advanced machine learning techniques to achieve precise telecom customer churn modeling and practical recommendations for customer retention management.

4.2. Preprocessing and Feature Selection

We first developed our models with little to know preprocessing, and the truth is that both our pycaret SVM model and our Neural Networks model were performing well on test and training data (close to 0.99 accuracy for both), and in a normal situation we'd leave it like this. The problem would arise if we were eventually given real world data, which is more imperfect and would almost definitely lead to overfitting, as our model would struggle to adapt to new data.

This prompted us to perform a very deep analysis of the data through intuition and also by analyzing the correlation between the features, using a correlation matrix. We started by exporting the dataset as an excel; first analyzing through intuition and our expertise, which resulted in the removal of 11 features that were deemed unnecessary or redundant, such as customer_id which offered very little explanatory value, therefore in this stage we dropped all categorical and purely informational features.

The next step was to generate a correlation matrix and analyze the correlation between variables. What high correlation between two features means, let's call them feature A and feature B, is that a big proportion of the explanatory value of feature A is explained by feature B, and therefore we could drop or take out one or the other, without worsening the model, while reducing overfitting. We defined our correlation threshold as +/- 0.7, so anything above 0.7 and below -0.7 (range -1 to 1), again, this means that any given variables in this range would have very strong correlation and share most of the explanatory value for the model.

In this step we ended up performing an extra step of feature engineering, as we noticed that the average monthly data usage was highly correlated to every individual month's data usage, but still below our 0.7 threshold (between 0.6 and 0.69). Therefore, we started by creating trimester averages ,which still did not make us confident enough to drop every month's data usage with some variables sharing a correlation above 0.7, but some not. Ultimately, we decided to go with quarterly averages, which were correlated to every month's data usage at values above 0.7.

With this feature engineering step, and further correlation analysis, we ended up dropping 12-month data usage variables, as well as other highly correlated variables such as joint_date and birth_date (preferring to keep engineered features such as current_age and years_on_plan, which are not datetime and can be read by the Pycaret model).

Ultimately, we ended up reducing the number of features from 59 to 35, even with this reduction the model appeared to perform remarkably well on test data, with an Accuracy of 0.989 and a AUC metric of 0.9996. Therefore, even after dropping 24 features, our model is performing remarkably well, further validating that this was the correct decision that strengthens our model and will help prevent future overfitting, especially when applied to more imperfect, real-world data.

4.3. PyCaret Setup and Modelling

The modeling process commenced with the splitting of the dataset into training and testing subsets in an 80/20 ratio. PyCaret was initialized using `setup()` with the following critical parameters:

- `normalize=True`: This was set to standardize feature scales across all features, hence, improving convergence and model performance.
- `target='churn_in_3mos'`: The binary target variable indicating customer churn.
- `session_id = 123`: For reproducibility.
- Using the default (mode) for categorical imputation and (mean) for numerical imputation.

Standardization of numeric features was performed using z-score normalization.

Models were trained and automatically compared using PyCaret's `compare_models()`. The Logistic Regression obtained higher performance for the first time, with:

- AUC = 0.9995
- Accuracy = 0.9884
- Recall = 0.9783
- F1 Score = 0.9809

SVM with a linear kernel was a close second.

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
lr	Logistic Regression	0.9884	0.9995	0.9783	0.9836	0.9809	0.9726	0.9726	1.1180
svm	SVM - Linear Kernel	0.9875	0.9993	0.9789	0.9802	0.9795	0.9705	0.9706	0.0510
ada	Ada Boost Classifier	0.9520	0.9898	0.9064	0.9345	0.9200	0.8857	0.8861	0.7790
xgboost	Extreme Gradient Boosting	0.9432	0.9872	0.8800	0.9306	0.9043	0.8640	0.8649	0.5370
qda	Quadratic Discriminant Analysis	0.9427	0.9873	0.8900	0.9203	0.9044	0.8635	0.8642	0.0440
lightgbm	Light Gradient Boosting Machine	0.9354	0.9834	0.8742	0.9106	0.8917	0.8457	0.8463	1.4750
ridge	Ridge Classifier	0.9316	0.9895	0.8871	0.8887	0.8878	0.8386	0.8387	0.0430
lda	Linear Discriminant Analysis	0.9316	0.9895	0.8912	0.8855	0.8883	0.8390	0.8391	0.0450
gbc	Gradient Boosting Classifier	0.9282	0.9806	0.8561	0.9038	0.8790	0.8280	0.8289	2.9500
rf	Random Forest Classifier	0.9134	0.9692	0.8204	0.8873	0.8523	0.7912	0.7926	1.2900
et	Extra Trees Classifier	0.8989	0.9635	0.8168	0.8464	0.8312	0.7591	0.7595	0.5330
nb	Naive Bayes	0.8779	0.9725	0.8268	0.7855	0.8052	0.7164	0.7173	0.0410
knn	K Neighbors Classifier	0.8754	0.9180	0.7156	0.8522	0.7775	0.6919	0.6974	0.0920
dt	Decision Tree Classifier	0.8239	0.7953	0.7220	0.7075	0.7142	0.5870	0.5876	0.1220
dummy	Dummy Classifier	0.6948	0.5000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0370

Figure 4.1: Scores of different models

4.4. SHAP Model Analysis Report and Feature Evaluation

This section presents an evaluation of feature importance using SHAP (Shapley Additive exPlanations) analysis, which gives a clear picture of the role that different features play in the model's predictions. The analysis includes three primary visualizations:

- A heat plot for global feature impact assessment (Plot 1).
- A waterfall plot for local feature contribution analysis (Plot 2).
- A PCA explained variance plot to check on the quality of the dimensionality reduction (Plot 3).

Analysis and Observations

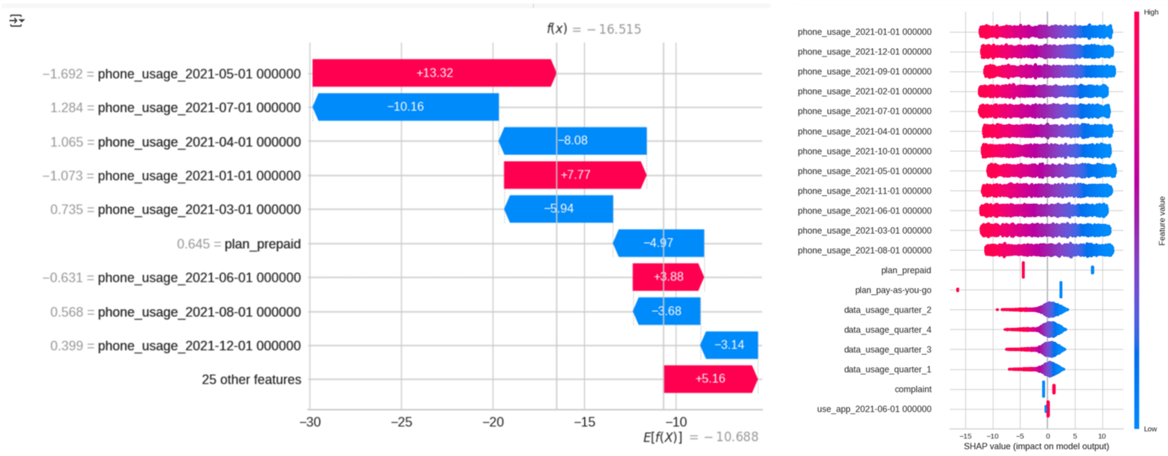

Figure 4.2: SHAP 1 (Left) and SHAP 2 (Right)

The heat plot shows the overall effect of each feature on the model's predictions. Key observations include:

- Features related to phone usage across different time periods are most important in the ranking – this shows that past phone usage is a key determinant of model results.
- The color gradient, which changes from blue to red, represents the SHAP value, where red represents positive values and blue represents negative or close to zero values.
- Quarters of data usage (Q1, Q2, Q3, Q4) are also equally important in determining the forecasts, revealing patterns in data consumption.
- The complaint feature has a mixed distribution of SHAP values, suggesting that its importance varies for different instances and that non-complainers are also less likely to churn.
- The plan type (prepaid/pay-as-you-go) has a somewhat limited but still significant impact on the predictions of user payment strategies.

The waterfall plot provides a local view of interpretability, showing how feature values affect an individual prediction. Key insights:

- The most positive influence on this prediction is `phone_usage_2021-05-01`, which increases the prediction score by +13.32.
- The largest negative influence is `phone_usage_2021-07-01`, which subtracts -10.16 from the score.
- Historical phone usage values from March and April are also significant and can contribute either positively or negatively, indicating that each month represents distinct behavioral aspects deemed important by the model.

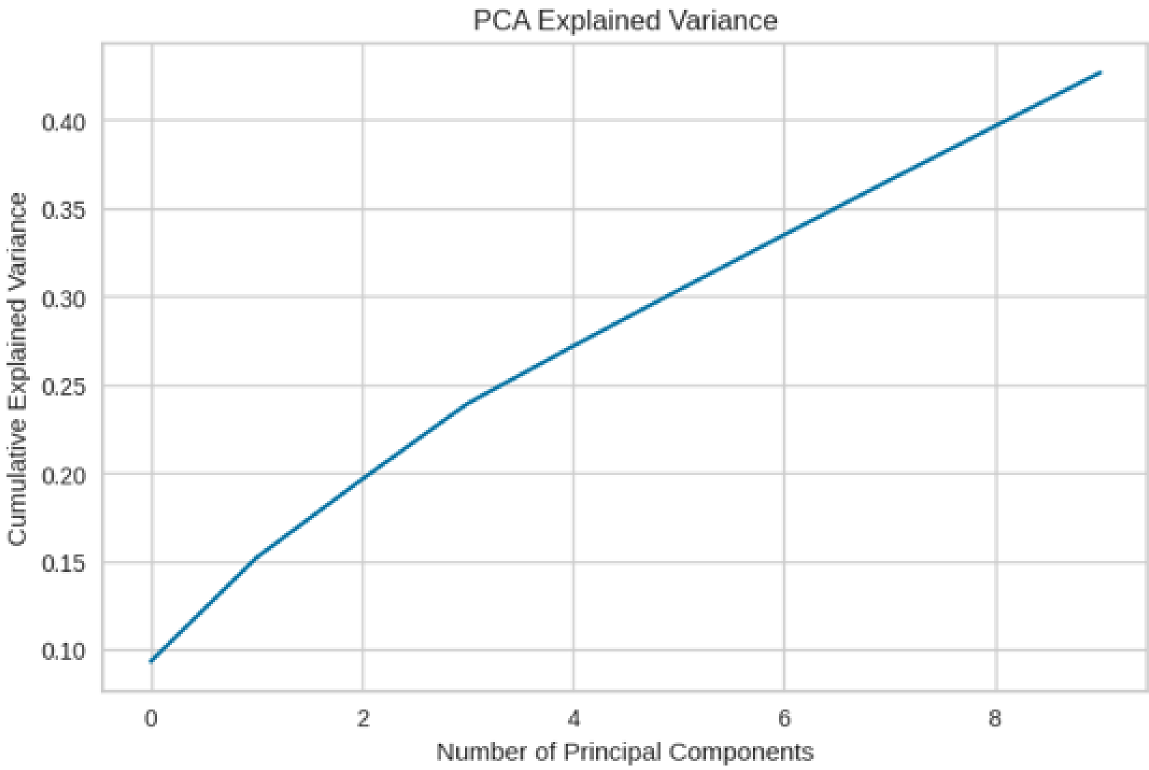

Figure 4.3: PCA: Explained Variance

The PCA plot demonstrates how principal component analysis (PCA) can be applied for dimensionality reduction of the dataset. Observations include:

- The cumulative explained variance increases steadily, with the first 8 principal components explaining about 40% of the variance.
- This relatively slow growth indicates that a large number of features contribute to the variance, suggesting that feature selection or other dimensionality reduction techniques might improve the model.

- However, since the dataset has already undergone feature selection, reducing the number of features from 59 to 35, the current set appears sufficient for maintaining predictive accuracy without overcomplicating the model.

The SHAP analysis confirms that phone usage history is the most important predictor in the model, both from a global and local perspective. Payment plans and complaint history also play a role but are not as influential as phone usage features. The analysis of monthly variables in Plot 2 shows that months are not merely repetitions of each other but provide unique information that influences predictions differently. The PCA analysis indicates that the dataset is likely high-dimensional and that prior feature selection has already enhanced efficiency. Given that the number of features has been reduced from 59 to 35, further reduction may not be necessary unless it leads to improved performance. Future enhancements should focus on leveraging SHAP values for targeted feature manipulation rather than outright removal.

4.5. Secondary PyCaret Run (Post-SHAP Feature Selection)

After feature pruning based on SHAP values, a new PyCaret session was initialized, reducing the feature set from 34 to 19. Preprocessing ensured no need for explicit categorical encoding, and to prevent data leakage, date-related features were excluded.

The Linear SVM model achieved superior performance:

- Accuracy: 0.9886
- AUC: 0.9995
- Recall: 0.9772
- Precision: 0.9853
- F1 Score: 0.9811

The slight decrease in AUC from 0.9995 to 0.9993 indicates a trade-off in feature reduction.

Stacked Model Approach

A stacked classifier was developed by combining the top three models:

- Tuned Logistic Regression (LR)
- Linear SVM (SVM)
- XGBoost (XGB)

Hyperparameter tuning was applied for 50 iterations to enhance generalization and avoid overfitting. Logistic Regression served as the meta-model for aggregation.

Hyperparameter Justification

- **Logistic Regression:** Regularization optimized to prevent overfitting.
- **Linear SVM:** SHAP analysis justified the use of the linear kernel, outperforming other kernels.
- **XGBoost:** Tuned estimators, learning rate, and depth maximized non-linear predictive ability.
- **Stacking:** Leveraged diverse model strengths for accuracy improvement.

Despite its robustness, the stacking model underperformed compared to the Linear SVM, leading to the SVM's selection as the final model.

4.6. Model Selection and Performance Evaluation

After feature pruning using SHAP values, the feature set was reduced from 34 to 19 before initializing a new PyCaret session. Preprocessing confirmed no need for categorical encoding, and date-related features were excluded to prevent data leakage. Among tested models, the Linear SVM demonstrated superior performance with an accuracy of 98.86%, AUC of 0.9995, recall of 97.72%, precision of 98.53%, and an F1-score of 98.11%. The slight AUC reduction (from 0.9995 to 0.9993) indicated a minor trade-off due to feature reduction.

A stacked classifier was developed by combining Tuned Logistic Regression, Linear SVM, and XGBoost, using Logistic Regression as the meta-model. Hyperparameter tuning (50 iterations) improved generalization while avoiding overfitting. Justification for the selected hyperparameters included optimized regularization for Logistic Regression, a SHAP-validated linear kernel for SVM, and fine-tuned estimators, learning rate, and depth for XGBoost. Despite leveraging the strengths of multiple models, the stacked model underperformed the standalone Linear SVM, leading to the latter's selection.

Performance comparison among individual models showed that the Linear SVM (AUC: 0.9865, Accuracy: 98.90%) outperformed Logistic Regression (AUC: 0.9826, Accuracy: 98.75%) and XGBoost (AUC: 0.9878, Accuracy: 94.55%). The stacked model achieved a mean accuracy of 98.21% and an AUC of 0.9983, offering improved stability but still falling short of the Linear SVM. Confusion matrices reinforced the SVM's reliability, with test accuracy reaching 98.9%, precision at 98.55%, and recall at 97.91%. Final validation metrics confirmed SVM as the optimal model for deployment, with train accuracy at 99.01% and test accuracy at 98.90%.

A neural network (NN) model was also explored for customer churn prediction, using standardized numerical features and one-hot encoding for categorical variables. The model was trained for up to 50 epochs with a batch size of 32, incorporating early stopping based on validation loss to prevent overfitting. The NN achieved a slightly higher test accuracy (99.55%) than the SVM (99.01%), but the significantly higher computational costs and resource demands made it less practical. Given its efficiency and reliability, the Linear SVM was ultimately chosen as the best model for real-world deployment in telecommunications, balancing predictive power with operational feasibility.

5
Business Case

5.1. Technical Implementation

The delivery strategy for the customer churn prediction solution is designed to ensure efficiency, simplicity, flexibility, security, and compliance with industry best practices. The solution is packaged in a ZIP file containing all necessary components for deployment, including the primary script `classify_customer_churn.py`, which executes the churn prediction, a setup script `setup.py` for automating dependency installation, a pre-trained model `model.pkl` to enable rapid inference without requiring retraining, and a user guide `classify_customer_churn_guide.pdf` that provides instructions for execution.

The deployment process is streamlined for seamless execution across multiple environments, including Windows, MacOS, and Linux. The `setup.py` script automates the installation of dependencies, ensuring compatibility and reducing setup complexity. The primary execution script utilizes `argparse` for flexible command-line input, allowing users to dynamically specify file paths and output locations, thus enhancing adaptability across various operational contexts.

A key advantage of this solution is local model execution, which significantly enhances data security and compliance. Running the predictive model locally prevents the exposure of sensitive customer data to third parties, addressing regulatory requirements such as GDPR, CCPA, and HIPAA. Local execution ensures enhanced data privacy, regulatory compliance, and performance stability by avoiding latency issues commonly associated with cloud-based systems. This guarantees uninterrupted service availability, which is crucial in decision-making scenarios that require high-speed access to insights.

Additionally, the use of a pre-trained model minimizes computational overhead, enabling organizations to generate insights quickly without the need for extensive model retraining. This allows businesses to respond in a timely manner to customer behavior trends and optimize retention strategies. The overall solution is designed to reduce technical barriers, improve the adoption of predictive analytics, and streamline customer churn prediction while maintaining high standards of security and efficiency.

This deployment strategy provides organizations with a scalable and secure approach to predictive analytics. By balancing technical efficiency with regulatory compliance, the solution offers an integrated framework that enhances decision-making while safeguarding sensitive data.

5.2. Retention Measures

Our recommendation system focuses on mitigating customer churn while optimizing cross-sell and upsell opportunities by leveraging a structured approach that integrates segmentation, strategy formulation, predictive modeling, and business actions. The process begins with the identification of key customer segments based on historical interactions, plan details, and issue categories. These segments are stored in JSON rule-sets, allowing for scalable and adaptable decision-making. Once segmented, customers are assigned to different strategic pathways, including retention, upsell, and cross-sell, each designed to maximize their lifetime value.

Retention strategies focus on preventing customers from leaving by offering incentives tailored to their engagement patterns. For instance, high-value customers at risk of churn may receive a six-month free service extension or a substantial store coupon, followed by a smaller follow-up incentive such as a one-month half-price plan in exchange for feedback. Other customers with moderate churn risk might be offered additional data roaming or international call minutes to encourage continued usage. These offers are measured through an estimated business value metric, ensuring that the most valuable customers receive the most effective interventions.

Upsell strategies target existing customers with opportunities to transition to higher-tier plans. A customer currently on a standard plan might be presented with an offer such as a free upgrade to a higher plan for six months, coupled with a follow-up discount after the promotional period. The goal is to gradually transition users into premium plans by making the switch seamless and rewarding. If the customer is not fully convinced, a counter-offer strategy is employed, where an alternative incentive, such as extended premium support, is presented to maximize retention within the upgraded segment.

Cross-sell strategies, on the other hand, aim to enhance customer value by introducing complementary products and services. Customers identified as likely candidates for additional services might be offered six months of free device insurance or premium support, with a follow-up offer for a discounted continuation of the service. Similarly, customers engaging with digital services like streaming or eSIM functionalities could receive time-limited free trials, followed by strategically placed discount offers to encourage long-term adoption.

To ensure these strategies are optimally deployed, we apply machine learning techniques, particularly XGBoost, to assess the influence of various factors such as tenure, unresolved issues, plan type, and intensity of support interactions. By using SHAP (Shapley Additive Explanations), we gain deeper insights into how these factors drive retention, cross-sell, and upsell probabilities. This allows us to refine our intervention models dynamically, ensuring that high-value customers receive prioritized engagement while minimizing unnecessary expenditures on lower-value segments.

Example Business Actions

The structured nature of our approach allows us to propose concrete business actions that are aligned with the estimated business value of a customer. Below are three examples of actions, each assigned to a specific strategy and test group:


```
{
    "action": "6 months free of charge",
    "strategy": "Retention",
    "test_group": "A",
    "estimated_business_value": 5,
    "follow_up": "1 month half-price for Feedback"
}

{
    "action": "12 months free upgrade to higher plan",
    "strategy": "Upsell",
    "test_group": "B",
    "estimated_business_value": 5,
    "follow_up": "Switching offer to higher plan with discount in first
}

{
    "action": "6 months Partner voucher (Streaming, Music, etc.)",
    "strategy": "X-Sell",
    "test_group": "A",
    "estimated_business_value": 5,
    "follow_up": "Offer to keep partner content at discount in first 3 m
}
```


These actions illustrate how we tailor recommendations to maximize retention, encourage plan upgrades, and cross-sell relevant services based on estimated customer value.

Ultimately, this approach allows us to align business objectives with customer-centric actions, optimizing outcomes by balancing immediate retention needs with long-term growth opportunities. The structured nature of the model ensures that it remains adaptable, allowing business teams to modify strategies as needed while maintaining a data-driven foundation for decision-making.

5.3. Economic Viability

Our churn prediction model, developed using the Linear Support Vector Machine (SVM), demonstrated excellent performance on test data, achieving an accuracy of 98.90%, precision of 98.55%, and recall of 97.91%. The confusion matrix indicates that out of 2,000 test cases, the model correctly identified 1,368 non-churners and 610 churners, with only 9 false positives and 13 false negatives. Compared to traditional customer retention methods, this targeted approach enhances operational efficiency and financial returns by optimizing resource allocation.

The high precision and recall of the model allow the company to effectively differentiate between churners and stable customers, ensuring that retention efforts are focused on the right individuals. This minimizes financial losses due to missed churners while avoiding unnecessary spending on customers who would have remained loyal without intervention. Consequently, the model maximizes revenue retention with minimal costs. To assess the financial impact, we use the following profit/loss framework:

$$\text{Profit/Loss} = (\text{Revenue from retained True Positives}) - (\text{Retention costs for True Positives} + \text{Retention costs} \tag{5.1}$$

Without intervention, the estimated revenue loss due to churn is €49,029.48. A brute force retention strategy, which applies intervention to all 2,000 customers at a cost of €5 per customer (€10,000 total), achieves an 80% retention success rate among the 623 actual churners, leading to the retention of 498 customers and generating €92,726.84 in annual retained revenue. After deducting intervention costs, this approach results in a net benefit of €82,726.84.

In contrast, the SVM model identifies 610 churners, reducing intervention costs. Given an 80% success rate, it is expected to retain 488 customers, yielding €90,768 in retained revenue. Accounting for intervention costs for true positives (€610), false positive costs (€45), and lost revenue from missed churners (€2,418), the predictive model achieves a net annual benefit of €87,695. This represents a financial advantage of €4,968.16 per year over the brute force approach (€87,695 vs. €82,726.84), an increase of approximately 6.00%. When scaled to 10,000 customers, this translates to an additional annual benefit of €24,840.80. Furthermore, if churn rates increase, the model's efficiency advantage would become even more pronounced, as targeted interventions would be even more cost-effective.

A Neural Network (NN) model was also tested and achieved a slightly higher accuracy of 99.55%. However, the NN approach entails higher operational complexity, scalability challenges, and increased computational costs. Given that the accuracy improvement over SVM is marginal, the additional investment required for NN does not currently justify its adoption.

Based on these findings, the SVM predictive model should be implemented immediately, with continuous monitoring using real-world data to prevent model drift and maintain optimal financial performance. Future assessments should periodically evaluate the feasibility of transitioning to an NN model, carefully weighing potential accuracy gains against operational costs.

Deploying predictive models in an organization entails some risks, including model drift, which can lead to decreased accuracy over time, and potential misclassification of customers. While these risks are relatively minor, they require proactive management through periodic performance evaluations, frequent model updates, and well-defined feedback mechanisms to ensure sustained effectiveness.

The next step involves testing the model with actual company data to validate its accuracy and practical efficiency. Additionally, the company should periodically reassess the potential benefits of migrating to an NN model and explore ways to enhance overall operational performance in the future.

Acknowledgement

We would like to express our sincere gratitude to our mentor, Gustavo Alberto Vouga de Carvalho Homem, for his invaluable guidance and support throughout this project. His expertise, insightful feedback, and unwavering encouragement have been instrumental in shaping our work and deepening our understanding of the subject. We are truly grateful for his mentorship and for the time and dedication he has invested in helping us achieve our goals.

References

European Union (2022). *Roaming charges within the European Union*. Accessed: 2025-03-18. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=legissum%3A4593182.

Sankur, Ali et al. (2024). *True customer-centricity: An operating model for competitive advantage*. Tech. rep. Accessed: 2025-03-18. McKinsey & Company. URL: https://www.mckinsey.com/industries/industrials-and-electronics/our-insights/true-customer-centricity-an-operating-model-for-competitive-advantage.

Adherence to academic standards

We hereby declare in lieu of an oath that we have written this thesis in the Master of Business Analytics and Data Science at IE School of Science and Technology independently and have not used any aids other than those specified. All passages taken verbatim or in spirit from publications are marked as such. We further confirm that we have not previously submitted the thesis in another examination procedure.

Madrid, the 18th of March 2025