ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Data Science Applications Project - Group 5

CHURN ANALYSIS FOR ABC TelCo

CHALLENGE BCG X @ UNIBO

2025

Made by:

Daniela Lopes	1900136918
Daniele Orio	0001194137
Giorgia Parisi	0001194901
Adriana Perrino	0001191441
Ranpatipura D.M.T. Jayaweera	0001176065
Alvin F. Terry	0001174918

Table of Contents

1. Introduction

Customer churn refers to the phenomenon where customers discontinue their use of a company's products or services for various reasons (Ahn et al., 2006; Filip, 2013). In the telecommunications sector, the annual churn rate reached 21% in 2022, resulting in losses of up to 65 million dollars (OvationCXM, 2025). Moreover, acquiring a new customer is estimated to be 5–10 times more costly than retaining an existing one (Lu, 2002). As a result, customer retention has become a strategic priority, especially in a highly competitive and rapidly evolving industry like telecommunications (Kumar & Reinartz, 2018).

Understanding the reasons behind customer attrition and being able to predict when it is likely to occur is essential for sustaining long-term business success. By identifying the factors that differentiate loyal customers from those likely to leave, telecom providers can design more targeted and effective interventions to reduce churn.

This project explores customer churn at **ABC TelCo**, a fictional telecommunications provider, using both **structured and unstructured datasets**. The main dataset contains 7043 customer records, while the complaints one has 1605 rows. The structured dataset has **21 variables**, including a unique customer identifier, demographic information, service subscriptions, usage patterns, and billing history, of which 16 are categorical and 3 are numerical. The key outcome variable is **"Churn"**, a binary variable indicating whether a customer has discontinued their service.

The unstructured dataset includes three additional variables consisting of **free-text customer complaints**, providing valuable qualitative insights into potential drivers of dissatisfaction. By combining both data types, this study aims to develop a deep understanding of customer behavior and the factors influencing churn.

All analyses were conducted using **R** and **Python programming**, with a focus on Gradient Boosting and LASSO. The goals of this study are to:

- Identify key drivers of customer churn
- Develop a predictive model to flag customers at risk of leaving.
- Provide actionable recommendations to improve customer retention.

The following sections will describe the dataset in detail, outline the methodology, present the main findings, and offer strategic suggestions based on the insights gained. A complete list of variables used in the analysis is included at the beginning of the following section.

2. Exploratory Data Analysis

2.1 Variable Description

Table 1: Structured Dataset

Variable	Description	Mode	Type
customerID	A unique identifier assigned to each customer		Statistical unit
gender	The gender of the customer	Male, Female	Categorical
SeniorCitizen	Whether the customer is a senior citizen	0, 1	Categorical
Partner	Whether the customer has a partner	Yes, No	Categorical
Dependents	Whether the customer has dependents	Yes, No	Categorical
PhoneService	Whether the customer has a phone service	Yes, No	Categorical
MultipleLines	Whether the customer has multiple phone lines	Yes, No phone service	Categorical
InternetService	Type of internet service the customer has	DSL, Fiber optic, No	Categorical
OnlineSecurity	Whether the customer has an online security add-on	Yes, No, No internet service	Categorical
OnlineBackup	Whether the customer has an online backup service	Yes, No, No internet service	Categorical
DeviceProtection	Whether the customer has device protection	Yes, No, No internet service	Categorical
TechSupport	Whether the customer has tech support	Yes, No, No internet service	Categorical
StreamingTV	Whether the customer has a streaming TV service	Yes, No, No internet service	Categorical
StreamingMovies	Whether the customer has a streaming movie service	Yes, No, No internet service	Categorical
Contract	The type of contract the customer has	Month-to-month, One year, Two years	Categorical

PaperlessBilling	Whether the customer has opted for paperless billing	Yes, No	Categorical
PaymentMethod	The method used by the customer to pay	Electronic check, Mailed check, Bank transfer (automatic), Credit card (automatic)	Categorical
tenure	The number of months the customer has been with the company	[0.00;72.00]	Continuous
MonthlyCharges	The current monthly charge billed to the customer	[18.25;118.75]	Continuous
TotalCharges	The total amount charged to the customer	[18.8;8684.8]	Continuous
Churn	Whether the customer has left the company or not	Yes, No	Binary

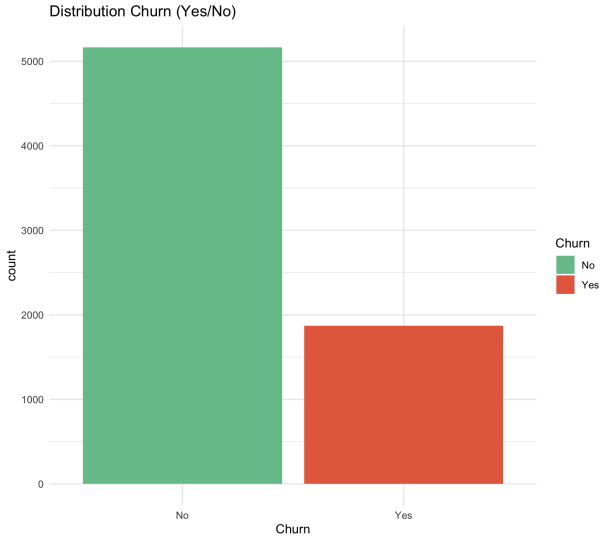
Table 2: Unstructured Dataset

Variable	**Description**	**Type**
customerID	A unique identifier assigned to each customer	Statistical unit
complaint	A free-text field containing the content of customer complaints	Text data
complaint_number	A numerical identifier unique to each complaint made by a particular customer. A single customer may have multiple complaints	Numerical

2.2 Data explanation

Before starting a detailed analysis, we first aimed to gain a basic understanding of the variables in the dataset. This preliminary exploration provides an overview of the data structure and highlights any initial patterns or irregularities worth further investigation.

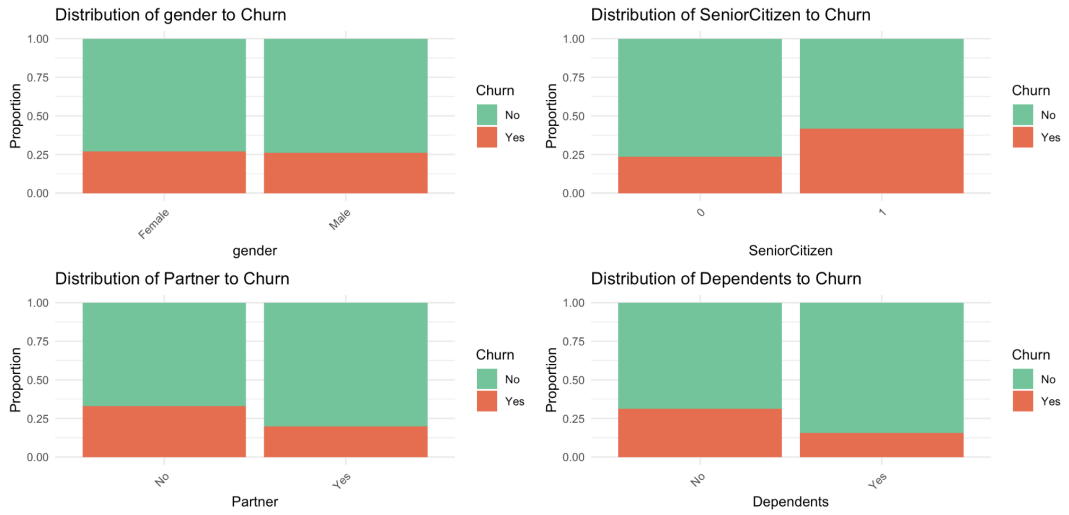
Figure 1:Distribution of Churn



To understand the overall distribution of the target variable, a bar chart was created to display the number of customers who have churned (Yes) versus those who have not (No). Out of the total 7,043 customer records, approximately 5,000 are labeled as "No", indicating that the majority of customers remained with the company. This suggests that the dataset is **imbalanced**, with a significantly higher proportion of non-churners compared to churners.

This imbalance is important to recognize early on, as it may affect the performance of classification models. Additionally, the relatively smaller number of churned customers emphasizes the need to identify patterns that distinguish them.

Figure 2: Churn Distribution by Customer Demographics



Distribution of Gender to Churn:

The analysis of gender distribution reveals that both male and female customers exhibit similar proportions of churn and non-churn. This observation suggests that gender does not significantly influence customer churn rates.

Distribution of Senior Citizen Status to Churn:

The data indicates that non-senior citizens have a higher proportion of non-churn compared to senior citizens. This finding suggests that senior citizens are more likely to churn than non-senior citizens.

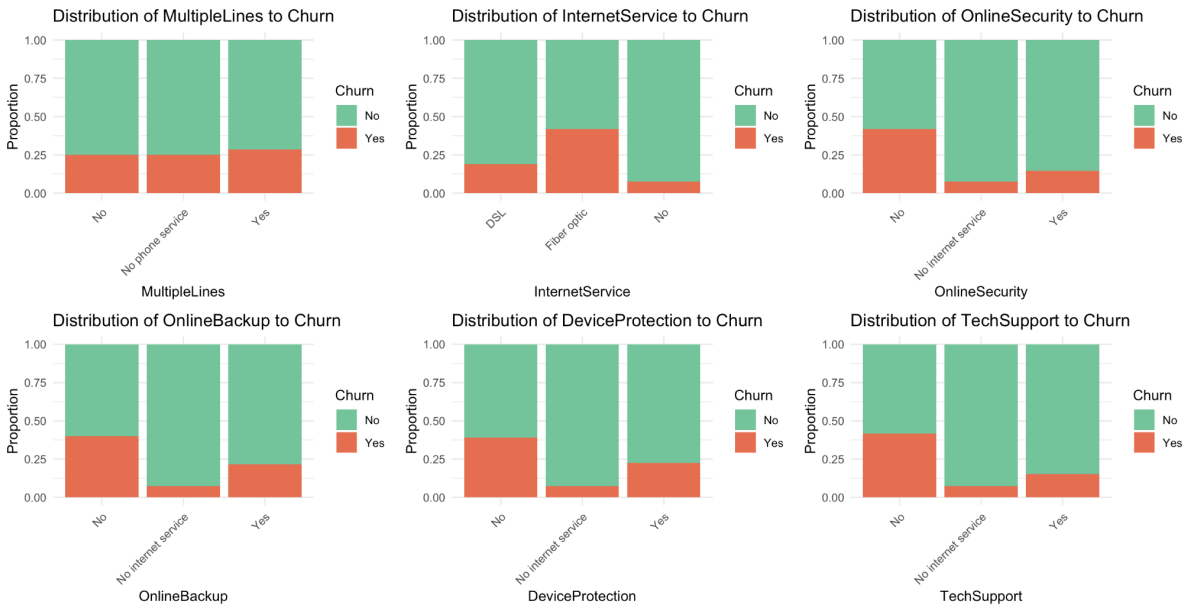
Distribution of Partner Status to Churn:

The distribution of partner status indicates that customers with partners exhibit a lower proportion of churn compared to those without partners. This suggests that having a partner may be associated with a reduced likelihood of customer churn.

Distribution of Dependents to Churn:

The analysis of dependent status reveals that customers with dependents have a lower proportion of churn compared to those without. This indicates that having dependents may be associated with increased customer retention.

Figure 3: Churn Distribution by Type of ABC TelCo Services Subscribed



Distribution of Multiple Lines to Churn:

The analysis of multiple lines of service shows only a slight difference in churn rates between customers with multiple lines and those without or with no phone service. This indicates that the presence of multiple lines does not appear to have a strong association with customer churn.

Distribution of Internet Service to Churn:

The data indicates that customers with fiber optic internet service have a higher proportion of churn compared to those with DSL or no internet service. This finding suggests that fiber optic internet service may be linked to higher churn rates.

Distribution of Online Security to Churn:

The distribution of online security services shows that customers without online security exhibit a higher proportion of churn compared to those with online security or no internet service. This implies that the absence of online security may be associated with higher churn rates.

Distribution of Online Backup to Churn:

The analysis of online backup service reveals that customers without online backup exhibit a higher proportion of churn compared to those with online backup or no internet service. This indicates that the absence of online backup may be linked to higher churn rates.

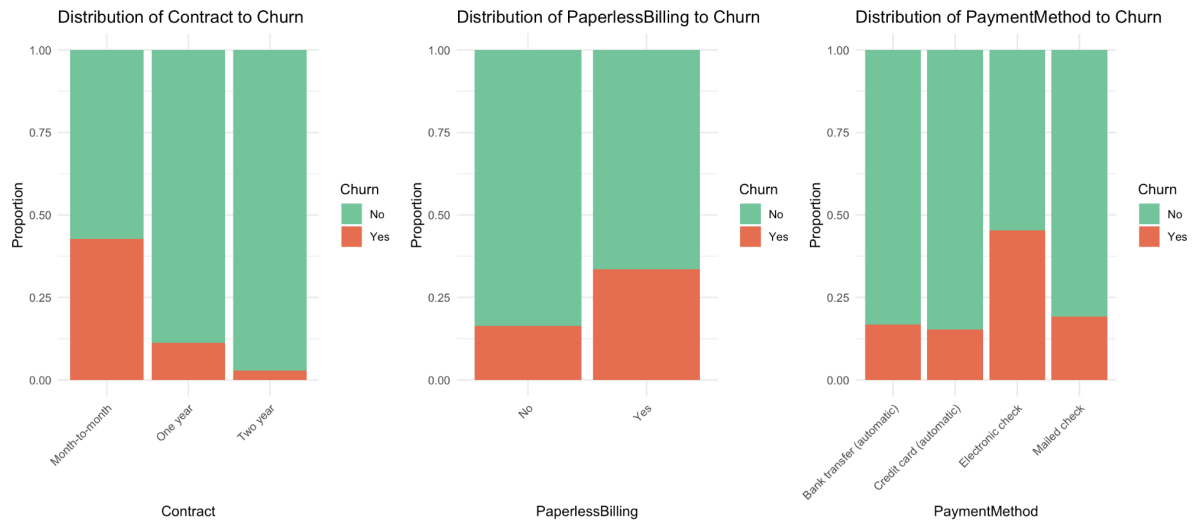
Distribution of Device Protection to Churn:

The data indicates that customers without device protection exhibit a higher proportion of churn compared to those with device protection or no internet service. This finding suggests that the absence of device protection may be associated with higher churn rates.

Distribution of Tech Support to Churn:

The distribution of tech support service shows that customers without tech support exhibit a higher proportion of churn compared to those with tech support or no internet service. This implies that the absence of tech support may be linked to higher churn rates.

Figure 4: Churn Distribution by Contract Characteristics and Payment Preferences



Distribution of Contract Type to Churn:

Customers on month-to-month contracts show a much higher churn rate than those on one-year or two-year contracts.

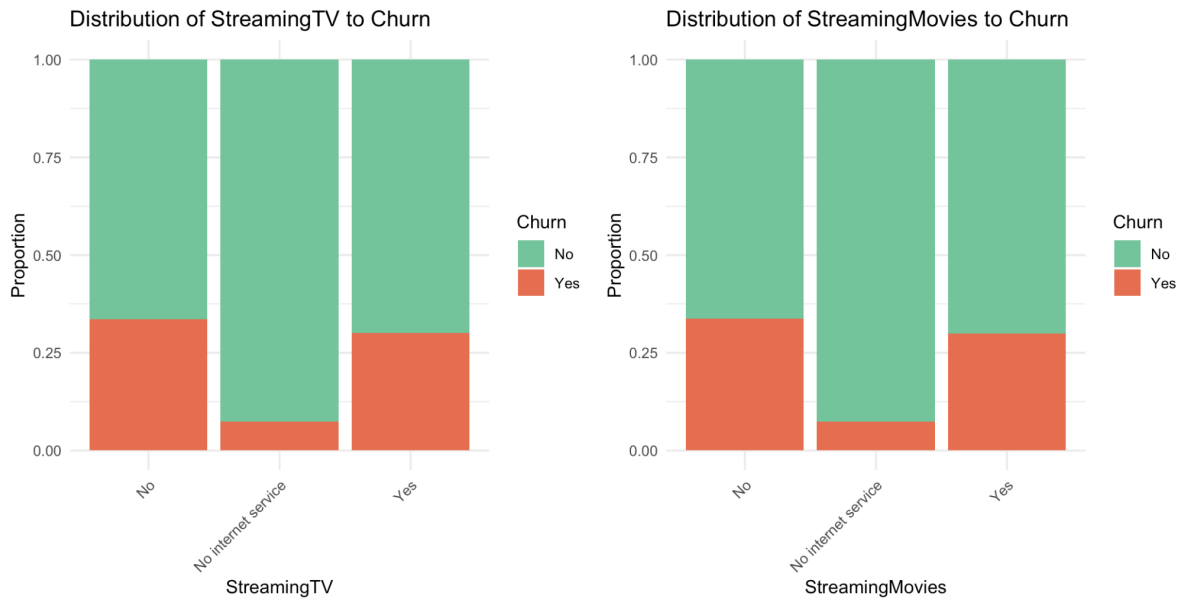
Distribution of Paperless Billing to Churn:

A higher proportion of churn is observed among customers who opted for paperless billing.

Distribution of Payment Method to Churn:

Customers who use electronic checks exhibit noticeably higher churn compared to those using bank transfers, credit cards, or mailed checks.

Figure 5: Churn Distribution by Streaming TV and Streaming Movies Subscription



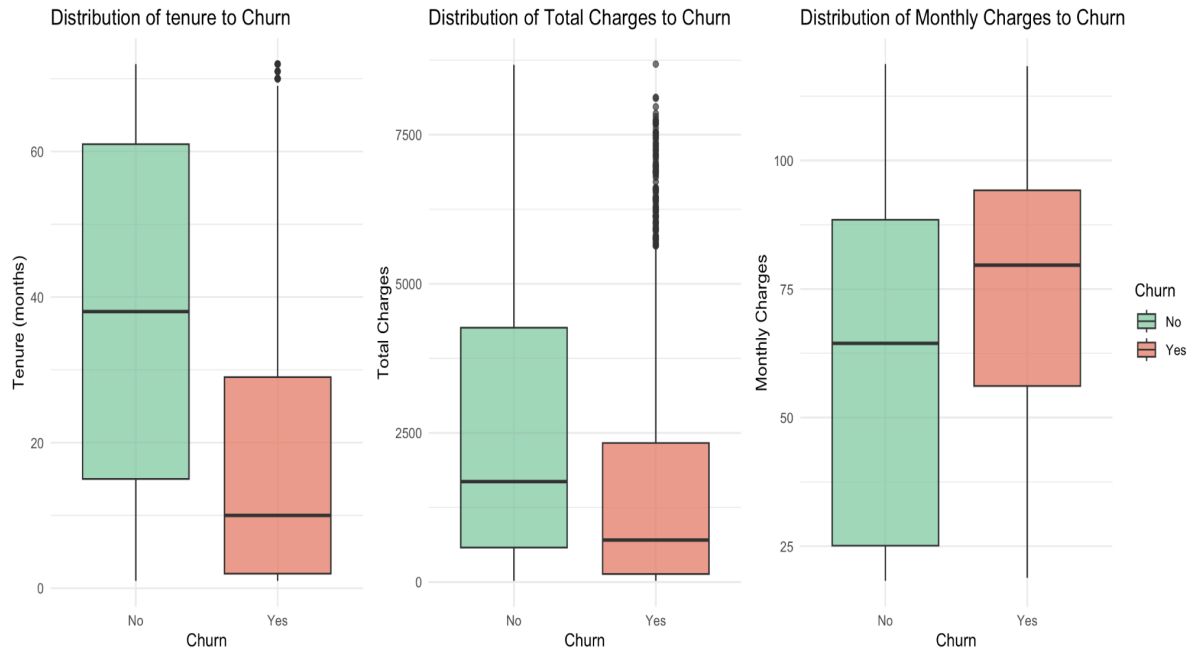
Distribution of Streaming TV to Churn:

Customers who do not subscribe to Streaming TV show a moderate churn rate. Those with no internet service exhibit very low churn. Interestingly, customers with streaming TV subscriptions show churn proportions similar to those of non-subscribers.

Distribution of Streaming Movies to Churn:

A nearly identical pattern is observed for Streaming Movies: churn rates are highest among those with the service and lowest for those without internet.

Figure 6: Churn Distribution by tenure, total charges, and monthly charges



Distribution of Tenure to Churn:

Customers who churned (Yes) had significantly lower tenure, typically under 25 months, compared to non-churned customers (No), who often had tenure above 50 months.

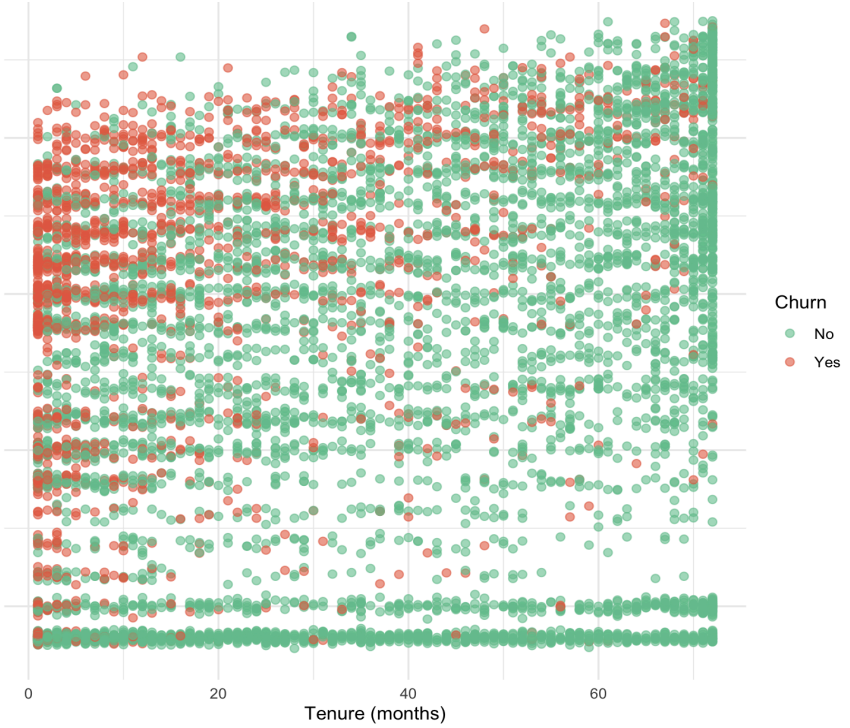
Distribution of Total Charges to Churn:

Customers who did not churn tend to have higher total charges. In contrast, most churned customers paid less, likely indicating early contract termination. Interestingly, there are several outliers among churned customers who paid very high total charges. These may represent long-term or high-value customers who left the company despite their significant spending.

Distribution of Monthly Charges to Churn:

Churned customers have slightly higher monthly charges than those who stayed.

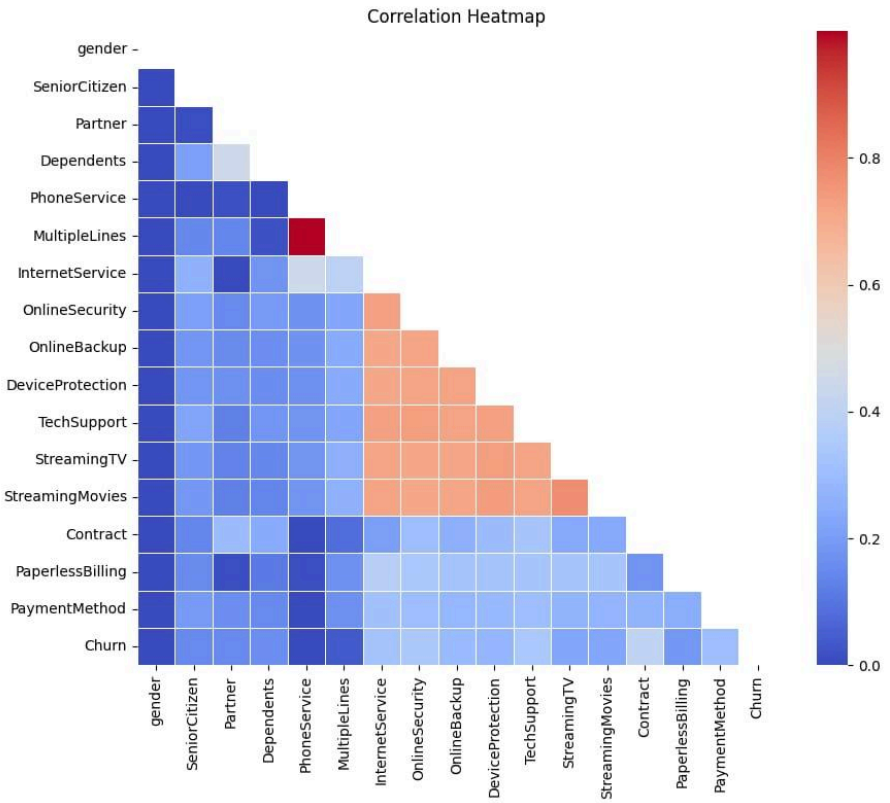
Figure 7: Scatter Plot of Tenure vs. Monthly Charges by Churn Status



Customers who churn (red) are more concentrated in the lower tenure range (especially below 20 months), regardless of monthly charges, but are more clustered around moderate to high charges. This suggests that newer customers who are paying more may be less satisfied or more likely to leave. Non-churned customers (green) are more prevalent across the entire tenure range and become dominant among customers with longer tenure, indicating that longer customer relationships are associated with greater retention.

Following the individual examination of each variable, we explored the relationships among the variables. Given the large number of categorical variables in our structured data, we employed **Cramér's V coefficient** as a statistical measure to assess the strength of association between pairs of categorical features. Cramér's V values range from 0 (no association) to 1 (perfect association), providing a useful metric for identifying potential dependencies. The figure below presents a heatmap of Cramér's V coefficients, illustrating the degree of association between all categorical variables in the dataset.

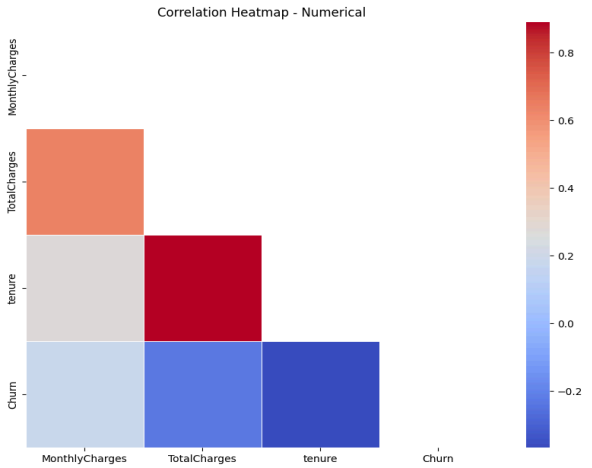
Figure 8: Heat Map for Categorical Variables



Key Observations:

- Service Feature Dependencies: A relatively high association is observed between several service-related features: OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, and StreamingMovies are moderately to strongly associated with InternetService.
- A strong association exists between PhoneService and MultipleLines. This results in a Cramér's V value close to 1, indicating near-perfect dependency, so considering both features is considered redundant information.
- The Contract type shows the strongest association with Churn (only 0.40).
- Other features, such as TechSupport, OnlineSecurity, and PaperlessBilling, also show lower moderate associations with Churn.
- Features like Gender, Partner, Dependents, and PaymentMethod exhibit low Cramér's V values in relation to most other variables, suggesting that they have limited interaction or overlapping information in the context of this dataset.

Figure 9: Heat Map for Continuous Variables



The continuous variables were accessed using Spearman Rank correlation, due to their not being normally distributed. A very strong positive correlation exists between tenure and Total Charges. A lower moderate negative correlation is observed between tenure and churn. Monthly charges show a weak positive correlation with churn. A moderate positive correlation exists between Monthly Charges and Total Charges.

3. Data Manipulation and Preprocessing

Before starting the analysis, several preprocessing steps were carried out to prepare the dataset: handling missing values, variable factorization, and feature engineering.

3.1 Missing values

As a first step, we checked the dataset for missing values. The only missing entries were found in the TotalCharges variable, with eleven cases. A cross-check with the tenure variable revealed that these eleven individuals were all newly acquired clients.
As a first approach, we tried to delete them, but, losing information, the error of the prediction increased; for this reason, we decided to set them equal to zero. Then we created a new binary variable indicating “new client” status.

3.2 Factorization of Categorical Variables

All categorical variables were converted into factors to ensure they were properly interpreted during the analysis and modeling phases. Specifically, the following variables were encoded as factors:
Gender, SeniorCitizen, Partner, Dependents, PhoneService, MultipleLines, InternetService, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies, Contract, PaperlessBilling, PaymentMethod, and Churn.

3.3 Feature Engineering

Several new variables were created to enrich the information content of the dataset and to capture patterns not directly visible in the original variables. This feature engineering phase required creativity and a deep study of the variables' meaning, aimed to enhance the predictive power of the model by transforming existing variables or combining them in more meaningful ways.

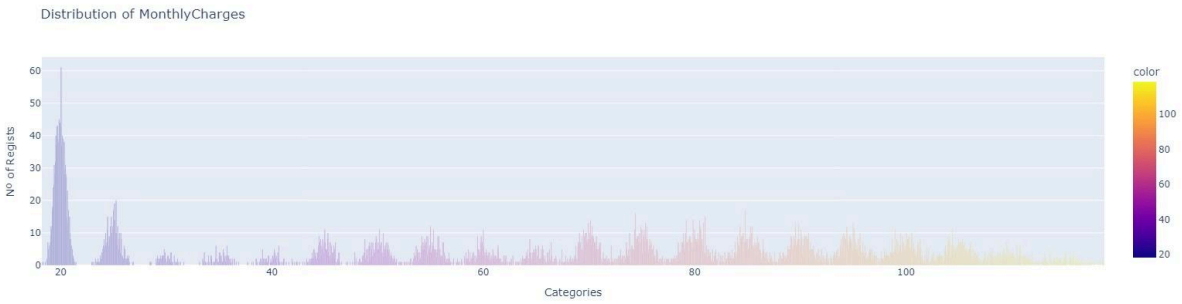
Tenure_cat. We found out that the variable holding information about how much time the subject has been a client is more informative when divided into categories.

First, we tried to make it a binary variable indicating whether the customer is a loyal client or a new acquisition. It takes the value 1 if the customer's tenure is 6 months or less, and 0 otherwise.

Later, we tried to divide the clients into years: the first category represents three months clients (the experiment period of time for new clients to test the service), the second contains clients of four - twelve months (1 year), the third represents one to two years clients, and so on until the sixth year.

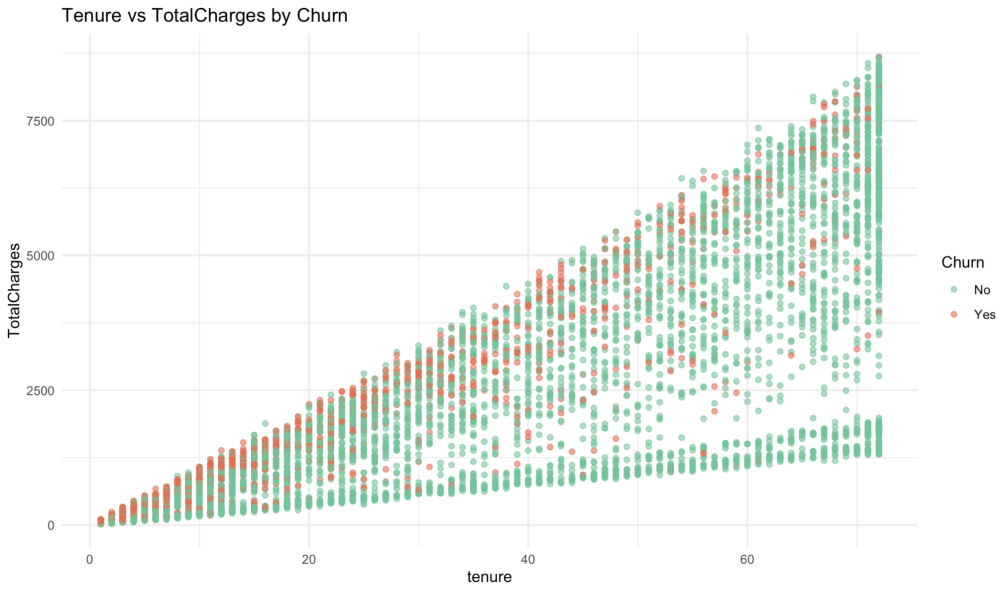
MonthlyCharges_cat. We grouped the MonthlyCharges variable into brackets to enhance interpretability and model performance, choosing the cut-off values based on the distribution of the variable

Figure 10: Distribution of Monthly Charges



TotalCharges_cat. As for MonthlyCharges, we grouped the variable into brackets, but with a different criterion for choosing the right value cut. As a matter of fact, TotalCharges is influenced by the tenure variable, as it represents the total amount charged to the customer, which naturally increases with time. For this case, we studied the distribution of TotalCharges for every year, and we chose the cut values based on the minimum value for every year.

Figure 11: Distribution of Tenure vs Total Charges by Churn



In the first 2–2.5 years, there appears to be a stronger concentration of churned customers (red dots), especially among those with higher total charges. However, as tenure increases beyond 30 months, the difference between churned and non-churned customers becomes less pronounced. In fact, both churned and retained customers are spread across similar total charge levels. This indicates that for longer-tenure customers, total charges isn't a strong distinguishing factor for churn.

ARPU. We calculated the Average Revenue Per User (ARPU) as the ratio between TotalCharges and tenure, and compared it with the MonthlyCharges variable. Although the two measures are conceptually similar, the correlation analysis and the distribution of their absolute differences revealed that they are not perfectly aligned. To better capture these differences, we created two binary indicators: change_charges (difference > 1) and change_charges5 (difference > 5).
According to our interpretation, the difference between ARPU and MonthlyCharges suggests that while MonthlyCharges reflects the current cost of the service, TotalCharges is a cumulative history of payments. Therefore, when the two differ, it may indicate that the customer has not always paid the same amount, possibly due to late payments or changes in the subscription plan. We found this signal potentially informative regarding the customer's decision to stay or leave the company. However, in practice, the inclusion of this variable did not substantially improve model performance.
We also interacted change_charges with the Contract type to explore whether inconsistencies between ARPU and MonthlyCharges are associated with specific subscription types, but this did not improve model performance nor add useful information.

senior_high_payment. Finally, we created a combined feature to indicate senior customers with above-average monthly payments. This did not improve model performance or add useful information, like many other combined features that we tried and decided to not report them in this document.

comments_number. During the sentimental analysis, we noticed that some clients did not have any written comments, while others had written many. We interpreted this as a potential

indicator of customer engagement or emotional investment, which could reflect either positive or negative experiences.

sentiment_variation. We noticed that it would be informative to know if someone who commented more than one time, had changed opinions with time, either positively or negatively. So we produced this new variable that is equal to zero if someone had never commented, commented one time, or commented more than one time having the same positive, neutral or negative score; and is equal to one if someone made two or more complaints not concordant in positive, negative and/or neutral.

Problems of multicollinearity for the Services variables

The variables concerning the offered services could create problems of multicollinearity.
Treating these variables was a necessary transformation to handle multicollinearity issues, because the category "No internet service" was perfectly predicted by the corresponding level in each variable.
So first we tried to create several binary variables to better represent the use and availability of specific services. For each of the following original categorical variable OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV and StreamingMovies, two new binary indicators were generated: one to capture whether the customer actively uses the service (*_yes), and another to indicate if the customer does not have internet access (*_no_internet), making the service inapplicable.
We chose not to group "No" and "No internet service" into a single category, as they reflect fundamentally different customer situations: in one case, the client has internet but opted out of the service, while in the other, they do not have internet access at all.

Then we experimented with summarizing the number of active services in a single variable, counting the number of "Yes" responses across the selected variables, and assigning the label of 99 to zero services, due to "no internet service".

We also tried to summarize only the streaming activity through a *streaming_number* variable (counting whether StreamingTV and StreamingMovies were active)

Unfortunately, all the three methods we tried to handle multicollinearity problems did not improve model performance nor add useful information. To address this issue more effectively, we will adopt modeling strategies that are better suited to handle this case.

All newly created variables were then converted into factors when needed, to be used in the modeling phase, but not all of them were included in the final model.

4. Model Development

In our work, we implemented different models, such as Decision Trees, CatBoost, Logistic Regression, and others, using two different programming languages: Python and R. We are trying to find the best model with the most optimal parameters possible. The following model was the model that gave us the best results.

4.1 Light Gradient Boosting Machine (LightGBM)

Gradient Boosting is a machine learning technique used primarily for regression and classification tasks. It belongs to the family of ensemble learning methods, which builds models by combining the strengths of multiple weak learners to create a robust predictive model. Ensemble learning involves combining multiple models to improve overall performance.
Boosting is a specific ensemble technique that focuses on converting weak learners— models that perform slightly better than random guessing—into strong learners with high predictive accuracy. The core idea is to sequentially train models, each compensating for the shortcomings of its predecessors.
Gradient Boosting leverages the principles of gradient descent, a first-order optimization algorithm used to minimize a loss function by iteratively moving toward the steepest descent as defined by the negative of the gradient. In the context of Gradient Boosting, gradient descent is employed to minimize a specified loss function by adding new models that predict the residuals (errors) of the combined ensemble.
Light Gradient Boosting Machine (LightGBM) is an advanced implementation of the Gradient Boosting framework. It is designed to be highly efficient, scalable, and capable of handling large-scale datasets with high dimensionality. LightGBM introduces several enhancements over traditional Gradient Boosting algorithms to improve speed and performance.

Theoretical Enhancements in LightGBM

LightGBM introduces several innovative techniques to improve efficiency and performance:

1. Gradient-based One-Side Sampling: Retains instances with large gradients (informative samples) while randomly sampling from those with small gradients, reducing computational demands without significant accuracy loss.
2. Exclusive Feature Bundling: Combines mutually exclusive features into a single feature, effectively reducing dimensionality in sparse data to minimize memory usage and speed up training.
3. Leaf-wise Tree Growth: Builds trees by expanding the leaf with the maximum loss reduction, capturing complex patterns. While it risks overfitting, it generally improves accuracy.
4. Categorical Feature Handling: Natively supports categorical features by optimizing splits for these variables without manual encoding, enhancing efficiency and model performance.
5. Parallel and Distributed Learning: Enables scalable training on large datasets by partitioning data across multiple machines or cores, maintaining high performance

4.2 Lasso

Sometimes it could be useful to fit a model that contains all the chosen predictors, but using some techniques that constrain the coefficient estimates. For this purpose, a Lasso regression has been used.
Lasso is an evolution of the Ridge Regression method, which performs better if some predictors are more important than the others, since it "forces" the coefficients associated with these less impactful predictors to 0.
Both these methods aim to reduce the RSS value by introducing a penalization term λ, which shrinks the coefficient estimates. This tuning parameter is estimated using cross-validation, and then the model is fit.
In our code, the cross-validation function performs a cross-validation in order to compute the misclassification error for the model. This function includes a threshold "choice" to improve the classification accuracy. A set of thresholds between 0.3 and 0.7 has been tried, since modifying the basic threshold of 0.5 to convert probabilities into predictions is useful when the classes of the response variable are not balanced. The threshold with the lowest error on each fold is selected (0.53), and then it is used to compute the predictions for that fold. With the function, we are able to obtain the list of predicted labels for the entire dataset.

4.3 Hyperparameters and Optimization

To enhance the LightGBM models and LASSO performance, we conducted hyperparameter optimization using Optuna, a powerful and efficient framework for automated hyperparameter tuning. We began the process by defining a broad search space for each model's key hyperparameters, allowing Optuna to explore a wide range of possible configurations. Through iterative sampling and evaluation, Optuna leveraged techniques such as Bayesian optimization to intelligently navigate the parameter space. Over multiple trials, the optimization process gradually narrowed down to regions with better performance, ultimately converging on a set of hyperparameters that yielded locally optimal results.

5. Sentiment Analysis

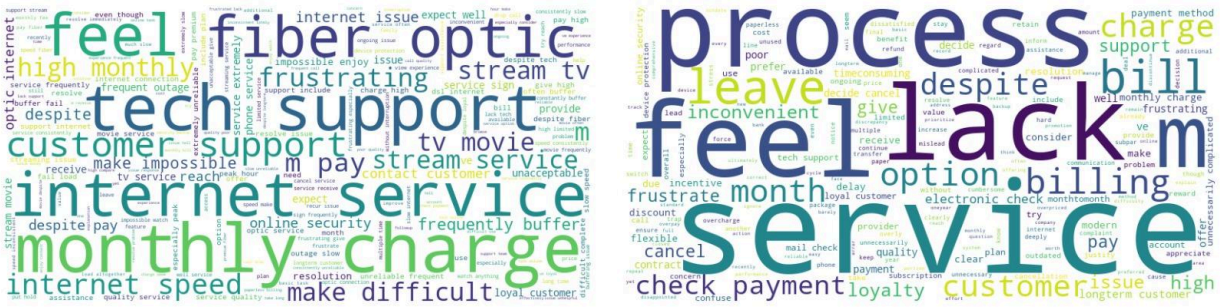
In this section, we will explore the dataset ‘ABCTelco_complaints’, which is a dataset that includes the complaints of 704 clients, and the number of complaints that each client made, so in total, we have 1605 different complaints. This dataset doesn’t have any missing values.

First of all, we start doing the preprocessing phase, which includes removing the stopwords, lemmatization of each complaint, and removing the punctuation.

Instead of using all the stop words that are predefined in ‘english’ vocabulary, we remove some of them, like ‘isn’t’, because if a client says that, for example: ‘My internet isn’t very good’, the model will consider this complaint as associated with a positive sentiment instead of a negative one.

The following WordClouds show the most commonly used words in all complaints and in the complaints that belong to the 73 clients who made more than one complaint.

Figure 12: Word Cloud of Common Terms in Customer Complaints. a) all complaints b) only referring to clients that made more than one complaint.



The word cloud highlights the most frequently mentioned words in customer feedback or complaints associated with churn. Larger words represent a higher frequency of occurrence. "Service," "process," "lack," and "feel" are among the most prominent terms. "Billing, "payment," "charge," and "check" might refer to financial and transactional matters. "Leave," "frustrate," and "inconvenient" suggest Emotional reaction. "Option" and "month" may imply dissatisfaction with limited billing cycle or service plan flexibility. “Internet service”, “fiber optic”, “tech support”, and “internet speed” indicate technical issues. Words like “monthly charge”, “frustrating”, and “make difficult” suggest cost concerns and usability problems. The repeated appearance of “support” emphasizes the role of customer sentiment and perceived lack of assistance.

Our objective is to determine the overall sentiment of each client complaint. We classify sentiment into three categories:

- Positive (1)
- Neutral (0)
- Negative (-1)

To achieve this, we use the SentimentIntensityAnalyzer model, which evaluates the proportion of negative, neutral, and positive words within each complaint. The model outputs a compound score that reflects the overall sentiment of the text

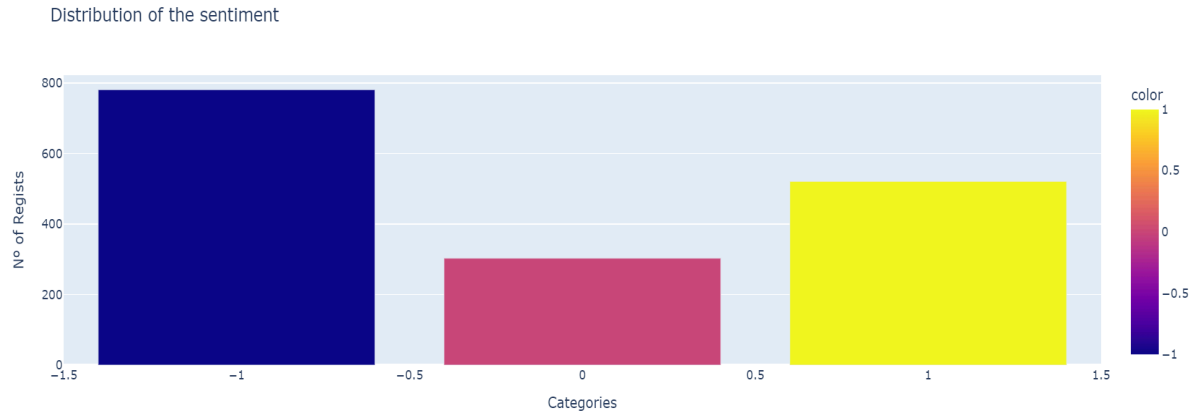
.
We then classify the sentiment based on the following thresholds:

- If the compound score is less than -0.15, the complaint is classified as Negative (-1).
- If the compound score is greater than 0.15, the complaint is classified as Positive (1).
- If the score falls between -0.15 and 0.15, it is classified as Neutral (0).

To assign a single sentiment per client, we created a function that processes cases where a client has submitted multiple complaints. This function counts the occurrences of each sentiment and returns the one that appears most frequently. In the event of a tie, we resolve it using the following priority: Negative less than -0.15, then Neutral between -0.15 and 0.15, and finally Positive greater than 0.15. This hierarchy is based on the assumption that users are generally more likely to take the time to write complaints when they have had a negative experience, rather than to compliment good service.

The following figure shows the distribution of the sentiments regarding the complaints, the new variable created to interpret the sentiment of each complaint (*sentiment*).

Figure 13: Distribution of the main sentiments



It is evident that the negative category has the highest number of complaints, with approximately 800 cases. The positive one reaches approximately 500, whereas the neutral category is the least frequent.

After computing the overall sentiment for each complaint, our next goal is to extract and summarize the main topics discussed across the complaints. To implement topic detection, we apply a method similar to Probabilistic Latent Semantic Analysis (PLSA) by using Truncated Singular Value Decomposition (SVD) on TF-IDF features. First, we convert the preprocessed complaint texts into a numerical representation using TF-IDF Vectorizer, limiting the vocabulary to the 1000 most relevant terms, excluding English stopwords, and filtering out overly frequent terms (max_df=0.5). Then, we apply TruncatedSVD with 20 components to reduce the dimensionality of the TF-IDF matrix, capturing the most salient patterns in the data. The resulting transformed matrix (*X_topics*) represents each complaint as a combination of topic scores, and we assign each complaint a dominant topic based on the highest-scoring component.
The following result consists of having 20 topics and 5 main categories:

1. Streaming & Buffering Issues

Topics: 0, 3, 15, 17, 19

Customers frequently express dissatisfaction with the quality of streaming services. Complaints focus on constant buffering, interruptions during video playback, and overall poor streaming quality. Keywords such as *"stream"*, *"buffering"*, *"tv"*, *"movie"*, and *"quality"* appear consistently, highlighting frustration, especially when trying to watch movies or live content. This suggests a significant gap in service delivery for one of the most used and valued features by customers.

2. Customer Support & Contact Issues

Topics: 1, 5, 7, 10, 12

Many complaints revolve around the inefficiency of customer support. Customers describe support as *"unhelpful"* or note repeated failed attempts to contact the service team. Words such as *"support"*, *"contact"*, *"reach"*, and *"frustrating"* indicate a perception of being ignored or not taken seriously. These concerns highlight that, beyond technical issues, customer service interactions significantly affect the overall experience and can strongly contribute to dissatisfaction and churn.

3. Payment & Billing Problems

Topics: 2, 8, 9, 16

A large number of complaints concern financial issues, including unclear billing statements, problems with payment processing, and dissatisfaction with pricing. Customers often mention feeling overcharged or not receiving value for the money paid. Keywords like *"charge"*, *"billing"*, *"monthly"*, *"payment"*, and *"money"* dominate these topics. This category indicates that smooth payment processes are essential to maintain customer trust.

4. Reliability & Outages

Topics: 6, 13, 18

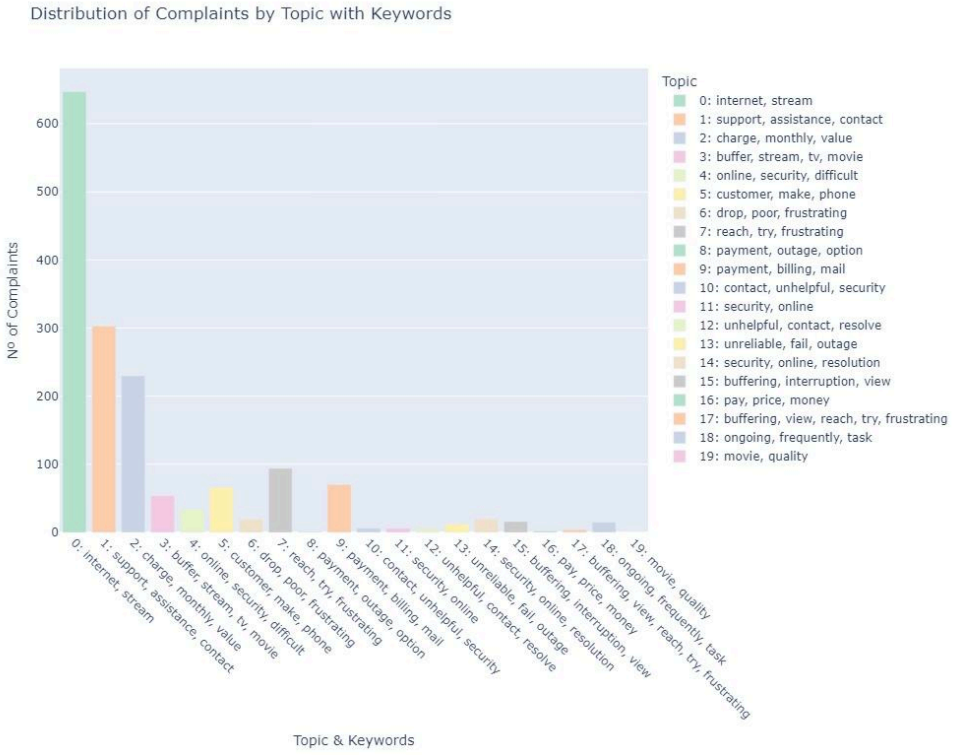
Another recurring issue relates to service reliability. Customers mention frequent *"outages"*, dropped connections, and general instability of the service, particularly during important tasks. The presence of terms like *"fail"*, *"unreliable"*, and *"frustrating"* reflects the critical impact that inconsistent connectivity has on customer satisfaction, especially as internet services become more integrated into daily life and work.

5. Security & Online Access

Topics: 4, 11, 14

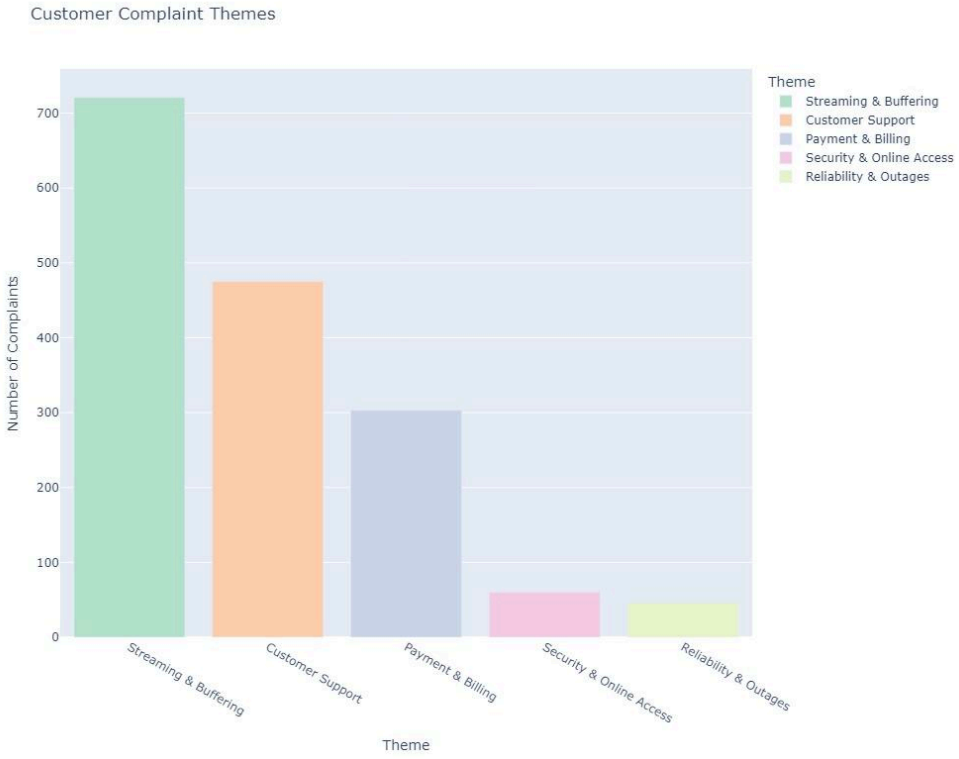
This group of complaints centers on issues of *"security"* and access to online services. Customers are concerned about account safety and have a lack of trust in the digital platforms used. Words such as *"online"*, *"security"*, *"difficult"*, and *"resolution"* point to technical and usability barriers that, if unresolved, can lead to loss of confidence and increased churn.

Figure 14: Distribution of Customer Complaints by Topic with Keywords.



This figure shows 20 common customer complaint themes. Internet and streaming are topic 0. This significantly leads to over 600 complaints, demonstrating that users care most about connection and streaming quality. Topic 1 (“support, assistance, contact”) and Topic 2 (“charge, monthly, value”) follow. These include complaints about billing transparency, customer service responsiveness, and value for money. Also significant are Topic 3 (“buffer, stream, tv, movie”) and Topic 17 (“buffering, view, reach, try, frustrating”), which emphasize media performance displeasure. Security, call quality, and problem resolution difficulties are highlighted in topics 4, 5, 6, and 10, underlying systemic flaws with technological infrastructure and support services. Topics 8 and 9 discuss payment and billing difficulties, while Topics 11 and 14 discuss security and internet access, where consumers demand smooth and secure experiences.

Figure 15: Distribution of Customer Complaints by Themes.



The plot shows that streaming and buffering are the most frequent problems, with over 700 complaints. Customer Support, Payment, and Billing rank as the second and third most prevalent subjects of complaint. Grievances about Security, Online Access, reliability, and outages are infrequent but significant.

As before, we encountered the issue of clients submitting multiple complaints. To address this, we applied a similar approach by assigning a single dominant topic per client, based on the most frequent topic among their complaints. In cases where there is a tie, we prioritize topics that are less represented overall. The prioritization order is as follows: 'Reliability & Outages', 'Security & Online Access', 'Streaming & Buffering', 'Customer Support', and 'Payment & Billing'. The final topic distribution, considering only one complaint per client, is as follows:

- 'Streaming & Buffering': 504 complaints;
- 'Customer Support': 113 complaints;
- 'Payment & Billing': 41 complaints;
- 'Reliability & Outages': 23 complaints;
- 'Security & Online Access': 23 complaints.

6. Results and Conclusion

This section aims to identify which model best fits the data and delivers the highest performance. The approaches that we adopt in this analysis use a consistent feature set consisting of 23 selected variables. These features were chosen through Recursive Feature Elimination (RFE), a method that iteratively removes the least significant features based on the performance of a given estimator. This process refines the feature set by retaining only those variables that contribute most to the model's predictive accuracy.
The final selected features are: *'SeniorCitizen', 'Partner', 'Dependents', 'tenure', 'PhoneService', 'InternetService', 'OnlineSecurity', 'OnlineBackup', 'DeviceProtection', 'TechSupport', 'StreamingTV', 'Contract', 'PaperlessBilling', 'PaymentMethod', 'MonthlyCharges', 'TotalCharges', 'New_client', 'Num_services', 'MonthlyCharges_cat', 'tenure_cat', 'TotalCharges_cat', 'sentiment', 'theme'*. The first 16 features are directly derived from the original dataset, while the remaining 7 were generated through feature engineering, enhancing the dataset with additional insights and structure. We use 4 final approaches:

- LGBM
- LGBM with Class Weights, given less importance to the majority class, due to the dataset that we are using is a bit unbalanced ('No' = 73.46%; 'Yes' = 26.54%)
- LASSO
- Voting Systems, a combination of the predictions of the three previous models using a majority vote (mode) strategy. For each client, the final prediction is the most frequently predicted class across the individual models.

The table below presents the optimal hyperparameters that yielded the best results for each model within its respective experiments.

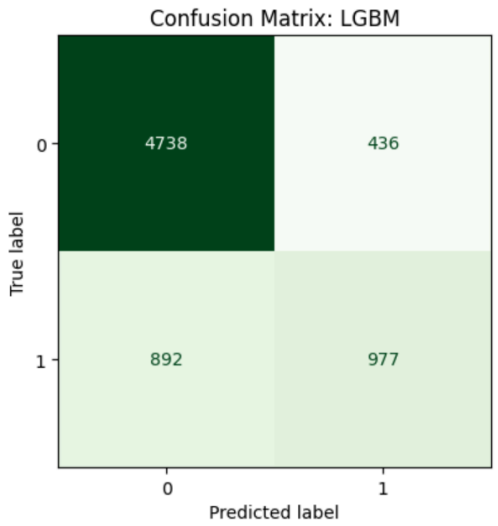
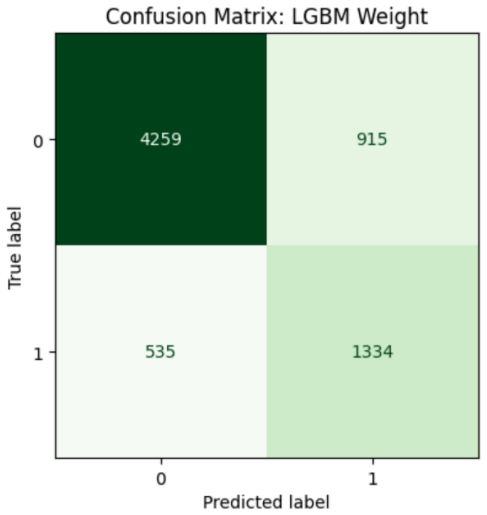
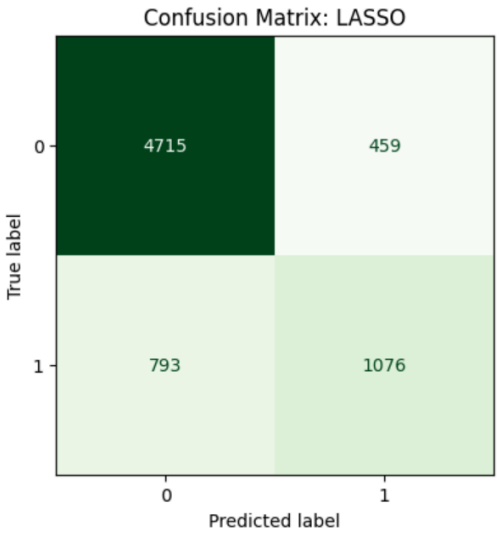
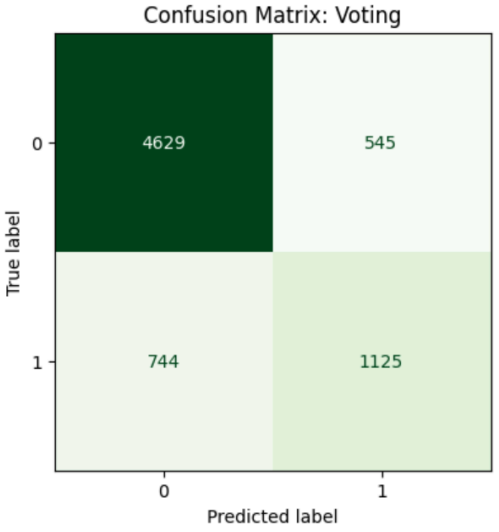
Table 3: Tables with the 3 best different models that proved the best score. a) Hyperparameters for the LGBM models. b) Hyperparameters for the LASSO model.

Hyperparameters	LGBM	LGBM with Class Weights
num_leaves	32	32
min_data_in_leaf	1	1
n_estimators	270	270
max_depth	75	75
learning_rate	0.01	0.01
bagging_fraction	0.8804043970736347	0.8804043970736347
feature_fraction	0.8613388610163988	0.8613388610163988
bagging_freq	1	1
class_weight	-	{0: 0.3, 1: 0.6}

Hyperparameters	LASSO
Cs	10
cv	5
penalty	l1
solver	liblinear
scoring	accuracy
max_iter	1000

Finally, we plotted a confusion matrix to visualize the model's performance. The Confusion Matrix revealed a high TP and TN rate, with minimal misclassifications, reinforcing the model's reliability.

Table 4: Confusion Matrices.

LGBM	**LGBM with Class Weights**
	
LASSO	**Voting Strategy**
	

LGBM confusion matrix shows that the model has a strong performance in identifying non-churners, but has a relatively high number of FN, meaning many actual churners were incorrectly classified.

LGBM with Class Weights makes a good trade-off, due to the model misclassifying more non-churners (FP), but can capture quite well the actual churners (TP and FN).

LASSO has a behavior very similar to the LGBM, but improves slightly in identifying churners, so we can say that it is a bit more balanced, but at the same time conservative.

Finally, as expected, the Voting Strategy is the one that represents the most balanced approach, since it combines these methods; it reduces both FN and FP moderately, showing a good trade-off.

The following table has the classification reports for each model used and the final accuracy using cross-validation with k_folds equal to 10.

Precision is a metric that measures how often a model correctly predicts the target class. The calculation is based on dividing the number of correct positive predictions (TP) by the total number of instances the model predicted as positive (TP and FP).

Recall is a metric that measures how often a model correctly identifies the target class from all the actual values that belong to the target class. The calculation is based on dividing the number of TP by the number of positive instances (TP and FN). For example, tell the percentage of clients who effectively left the company if the predictions classify others as having left the company.

The F1 score, which combines precision and recall into a single value, provides a balanced assessment of the model's performance. This ensures that both precision and recall are given equal importance in the evaluation.

The final metric is the Misclassification error, which is computed as 1 - accuracy.

Table 5: Classification reports, based on the best different models.

	Precision		**Recall**		**F1-score**		**Misclassification error of the cross-validation**
	0 (No)	**1 (Yes)**	**0 (No)**	**1 (Yes)**	**0 (No)**	**1 (Yes)**	
LGBM	0.84	0.69	0.92	0.52	0.88	0.60	**0.1885**
LGBM Weight	0.89	0.59	0.82	0.71	0.85	0.65	**0.2058**
LASSO	0.86	0.70	0.91	0.58	0.88	0.63	**0.1777**
Voting	0.86	0.67	0.89	0.60	0.88	0.64	**0.1777**

Analyzing Table 4, we observe that all models perform consistently well when classifying clients who did not churn. Precision, recall, and F1-scores for this class are consistently above 0.84, indicating that the models are highly effective in identifying the majority class. This strong performance is largely due to the class imbalance, where 'No' constitutes over 73% of the dataset. However, performance drops notably for the minority class (Churn = Yes), with recall values ranging from 0.52 to 0.71 and F1-scores from 0.60 to 0.65. This indicates that the models are less capable of correctly identifying clients who are likely to churn, which is critical from a business perspective. Among all models, LGBM with class weighting achieved the highest recall for the minority class (0.71), demonstrating the benefit of incorporating class imbalance into the learning process. By analyzing the obtained results about

misclassification error, we can see that the LGBM approaches seem to perform worse than Lasso, which achieves the best performance. Also, in this case, the Voting strategy summarizes this information, but does not improve the overall result provided by Lasso.

Although the Voting System does not outperform all individual models in every metric, it provides a balanced performance, particularly in F1-score for the Churn = Yes class (0.64), which suggests improved stability.

This analysis highlights the importance of not only optimizing the overall accuracy but also ensuring recall for the minority class, especially in 'Churn' prediction tasks, where missing a true churner can have a direct financial impact. So, if the business company's priority is to detect as many churners as possible, the best solution is LGBM with Class Weight. Otherwise, if the company requires a balanced model that reduces both types of errors reasonably, the Voting System, the consideration of the 3 models offers the most stable overall performance.

7. Suggestions and possible solutions

In this section, we aim to identify and quantify the key factors that influence customer churn in order to provide concrete, data-driven recommendations to the company. While earlier analyses helped highlight several variables significantly associated with churn, we now move beyond simple associations to determine which features have the strongest independent impact on the probability of a customer leaving. To investigate which variables most strongly influence the variation in customer *Churn*, we performed a stepwise multivariate logistic regression. This method allows for the selection of the most relevant predictors by sequentially including or excluding variables based on their statistical significance, while also accounting for potential multicollinearity, using as a decisive criteria the AIC value.
The estimated coefficients from the final model will help quantify the strength and direction of each factor's influence, offering valuable insights into where the company should focus its efforts to reduce customer attrition.

Table 6: Outcome of the multinomial logistic regression.

Variable	**OR**	**2.5% CI**	**97.5% CI**	**p-value**	**Significance**
(Intercept)	0.604	0.466	0.782	0.00014	***
ContractOne year	0.513	0.419	0.627	4.85e-10	***
ContractTwo year	0.256	0.179	0.364	1.05e-14	***
InternetServiceFiber optic	2.104	1.730	2.561	3.72e-14	***
InternetServiceNo	0.464	0.359	0.599	2.42e-08	***
tenure	0.942	0.931	0.954	<2e-16	***
PaymentMethodElectronic check	1.357	1.126	1.636	0.00121	**
MultipleLinesNo phone service	1.880	1.457	2.427	1.98e-06	***
MultipleLinesYes	1.282	1.091	1.507	0.00181	**

PaperlessBillingYes	1.412	1.213	1.643	3.50e-06	***
TotalCharges	1.000	1.000	1.000	5.30e-06	***
OnlineSecurityYes	0.668	0.561	0.795	2.00e-06	***
TechSupportYes	0.648	0.543	0.774	7.23e-06	***
StreamingMoviesYes	1.213	1.034	1.423	0.01711	*
SeniorCitizen1	1.240	1.053	1.461	0.01035	*
OnlineBackupYes	0.841	0.719	0.983	0.02550	*
StreamingTVYes	1.203	1.026	1.411	0.02272	*
DependentsYes	0.862	0.737	1.007	0.06817	.

p≤0.05 p≤0.01** p≤0.001****

The table reports only variables that are statistically significant and not excluded due to multicollinearity.

Key predictors negatively associated with churn include having a long-term contract (*one-year* or *two-year*), longer *tenure*, and services such as *Online Security* and *Tech Support*. On the other hand, factors positively associated with churn include using *electronic check* as a payment method, subscribing to *fiber optic* internet, and enabling *paperless billing*.

As we can read from the table, long-term contracts, such as one-year and two-year agreements, are significantly associated with a reduced probability of churn. Specifically, customers with a one-year contract are about 49% less likely to churn compared to those with a month-to-month contract, while a two-year contract reduces the odds of churn by about 74%. These findings suggest that long-term contracts strongly promote customer retention. Additionally, longer customer tenure is associated with a 6% decrease in churn risk for every additional month, indicating that loyalty builds over time.

Factors that decrease churn also include having online security and technical support services. Customers with online security are 33% less likely to churn, while those with technical support are 32% less likely to leave, highlighting the value of these additional services in fostering customer loyalty and satisfaction. Furthermore, customers who use online backup are 16% less likely to churn, further supporting the importance of supplementary services in reducing churn risk. Having dependents is also marginally associated with lower churn, though this result is borderline statistically significant.

On the other hand, several services are positively associated with an increased likelihood of churn. Customers who pay via electronic check are 36% more likely to churn, possibly due to lower engagement or trust in the service. The use of paperless billing is associated with a 41% higher likelihood of churn, suggesting that this segment may include more digitally autonomous users. Fiber optic internet subscribers are twice as likely to churn, potentially due to higher costs or technical issues. Additionally, customers with no phone service are 87% more likely to churn, possibly reflecting a lower level of investment in the full service package, or, again, technical issues. Customers with multiple lines are 28% more likely to churn, which could be due to higher billing complexity or dissatisfaction with the service.

Similarly, the presence of streaming services and streaming TV is slightly associated with increased churn, maybe due to competition with other platforms.

In conclusion, from this analysis, we can derive several insights to improve customer retention and thus enhance the company's overall performance.

First, the data suggests that customers who opt for longer commitments are significantly less likely to leave the service compared to those on month-to-month contracts. To leverage this insight, the company could focus on promoting long-term contracts by offering attractive incentives, such as discounts or exclusive benefits for customers who commit to a one- or two-year agreement. This would encourage customer retention and foster long-term loyalty.

Furthermore, the analysis highlights the importance of value-added services, such as online security, technical support, and online backup, in reducing churn. To capitalize on this, the company should consider bundling these services into standard packages or offering them as part of premium options. Additionally, targeted marketing campaigns that emphasize the value and benefits of these services could further reinforce their role in boosting customer retention.

Another key issue identified is the higher churn rate among customers who pay via electronic check. To address this, the company could explore offering alternative payment options that are more engaging or build greater trust with customers. For example, providing more flexible payment plans or offering additional incentives for those using electronic checks could improve retention in this segment.

Finally, the analysis suggests that customers using paperless billing may be at a higher risk of churn. This behavior could indicate one of two tendencies: either a preference for digital autonomy or simply a desire to reduce physical mail and paper clutter. In the first case, these customers may be more accustomed to managing services independently and could be more open to switching providers if they perceive better value elsewhere. To address this, the company could adopt a more personalized and proactive approach, offering tailored digital communications that emphasize the advantages of the current service and highlight new features, promotions, or bundled offers.

In the second case, where paperless billing is more about convenience than digital engagement, the lack of physical touchpoints might lead to a weaker emotional connection with the brand. For these customers, the company could consider reintroducing occasional physical communications—such as personalized letters, postcards, or promotional materials—as a way to maintain visibility and reinforce the relationship. Alternatively, timely notifications through other non-intrusive channels like SMS or push notifications can serve a similar purpose, keeping the brand present without overwhelming the customer. Addressing both profiles thoughtfully can help strengthen engagement and reduce churn in the paperless billing segment.

In conclusion, based on the analysis and supporting economic and marketing theories, we recommend that the company focus on strengthening customer commitment as a key strategy to reduce churn. According to the **Commitment-Trust Theory (Morgan & Hunt, 1994)**, fostering a strong emotional and psychological attachment between the customer and the service provider significantly increases loyalty and reduces the likelihood of defection. This can be achieved through long-term contracts, personalized support, and the inclusion of high-value services that enhance the customer experience.

Moreover, the **Customer Retention Theory (Reichheld & Sasser, 1990)** emphasizes that retaining existing customers is far more cost-effective than acquiring new ones. This further supports the need to invest in initiatives that deepen the relationship with current clients, such as loyalty programs or proactive engagement strategies. Additionally, drawing on **Disconfirmation Theory (Oliver, 1980)**, companies should aim to consistently exceed customer expectations by delivering services that add tangible value—this can play a pivotal role in increasing satisfaction and minimizing churn.

Finally, principles from Behavioral Economics, particularly **Loss Aversion (Kahneman & Tversky, 1979)**, suggest that customers are more motivated to avoid losing a benefit than to gain a new one. Communicating the potential loss associated with discontinuing services—such as the loss of security features, support, or bundled discounts—can be a powerful tool in encouraging customers to stay.

In addition to these retention strategies, the company should also focus on effectively marketing these offers and service bundles across multiple communication channels. Using targeted campaigns through email, SMS, in-app notifications, and social media can help ensure that customers are aware of the available benefits and feel continuously engaged with the brand. A strong multichannel marketing approach can amplify the impact of loyalty-building efforts and further support long-term customer retention.

Together, these theories provide a solid foundation for a customer-centric retention strategy focused on commitment, satisfaction, and perceived value.

8. References

Ahn, J. H., Han, S. P., & Lee, Y. S. (2006). Customer churn analysis: Churn determinants and mediation effects of partial defection in the Korean mobile telecommunications service industry. *Telecommunications policy*, *30*(10-11), 552-568.

Filip, A. (2013). Complaint management: A customer satisfaction learning process. *Procedia-Social and Behavioral Sciences*, *93*, 271-275.

Kumar, V., & Reinartz, W. (2018). *Customer relationship management*. Springer-Verlag GmbH Germany, part of Springer Nature 2006, 2012, 2018.

Lu, J. (2002). Predicting customer churn in the telecommunications industry—An application of survival analysis modeling using SAS. *SAS User Group International (SUGI27) Online Proceedings*, *114*, 27.

Reichheld, F. F., & Sasser, W. E. (1990). Zero defections: Quality comes to services. *Harvard Business Review*.

Morgan, R. M., & Hunt, S. D. (1994). The commitment-trust theory of relationship marketing. *Journal of Marketing*, 58(3), 20–38.

Oliver, R. L. (1980). A cognitive model of the antecedents and consequences of satisfaction decisions. *Journal of Marketing Research*, 17(4), 460–469.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2), 263–291.

Ovation (2025). What's Happening in 2023 Telecom Customer Experience Management. Accessed from https://www.ovationcxm.com/blog/whats-happening-in-2023-telecom-customer-experience-management?utm_source=chatgpt.com

R Core Team. (2024). R: A language and environment for statistical computing (Version X.X) [Computer software]. R Foundation for Statistical Computing. https://www.r-project.org/

Python Software Foundation. (2024). Python (Version X.X) [Computer software]. https://www.python.org/