
A. Gary Anderson School of Management

School of Business

**STAT 206**: **Statistical Computing**

Dr. Alfonso Landeros

# Predicting Telecom Customer Churn: A Machine Learning Approach

**Authors:**

Saloni Bonde

Ramylakshmi Lakshmi Narasimhan

Red Kurti

**TABLE OF CONTENTS**

## 1 Abstract

Customer churn is a critical challenge in the telecom industry, impacting revenue and sustainability. This project aims to identify predictors of churn and develop a predictive framework. Utilizing the Telco Customer Churn dataset from IBM, with over 7,000 records and 33 features, we conducted comprehensive analysis, including EDA, data preprocessing, and deploying machine learning models like Logistic Regression, K-Nearest Neighbors, Random Forest, and XGBoost. Our methodology focused on understanding data patterns, addressing quality issues, and optimizing models for accuracy, precision, and recall. Factors like contract length, payment method, and service usage patterns emerged as significant predictors. Random Forest and XGBoost showed superior performance. This study emphasizes predictive analytics' value in enhancing customer retention strategies and suggests exploring advanced modeling techniques for better accuracy and operational applicability in the telecom sector.

## 2 Introduction

In the competitive telecom industry, customer retention is crucial. Understanding and mitigating customer churn are key to this. Churn directly impacts revenue and long-term sustainability. Predictive analytics helps anticipate customer behavior to reduce churn rates. This project aims to identify churn predictors and develop accurate predictive models. Using IBM's dataset of over 7,000 customer records, we employed Exploratory Data Analysis and preprocessing to analyze churn dynamics. Machine learning algorithms like Logistic Regression, K-Nearest Neighbors, Random Forest, and XGBoost were applied and evaluated based on accuracy, precision, and

recall. The goal is to provide telecom companies with predictive tools for anticipating churn and crafting targeted retention strategies. This project contributes to customer relationship management by highlighting churn predictors and evaluating predictive model performance. It emphasizes data-driven strategies for enhancing loyalty and fostering growth in the telecom industry.

## 3 Dataset Overview

The dataset utilized in this study, sourced from IBM data repositories, encompasses a comprehensive array of customer information from a telecom company. It consists of 7,043 individual records, each described by 33 features that capture a wide spectrum of customer demographics, account details, service usage, and billing information. These features include, but are not limited to, gender, age (senior citizen status), partnership and dependency status, tenure, phone and internet service options, contract type, payment method, and monthly and total charges.

### 3.1 Target Variable: Churn Value

The primary focus of this analysis is the binary target variable 'churn', indicating whether a customer has left the company's service within the last month. The dataset reveals a churn rate of

approximately 27%, signifying that out of the total customer base, 27% have churned, leaving a substantial 73% who have remained with the service.

### 3.2 Features

#### 3.2.1 Gender

The gender breakdown shows a near equal split, with 49% female and 51% male customers. This balanced distribution implies a diverse customer base. While initial analysis doesn't suggest a strong correlation between gender and churn, deeper exploration, considering other variables, might unveil nuanced insights.

#### 3.2.2 Senior Citizen

The distribution reveals that 16% of the customer base answered "Yes," while 84% answered "No." This indicates that senior citizens constitute a smaller portion of the customer base. Given this demographic's potentially distinct needs or service expectations, their churn rates might be influenced differently, warranting further investigation.

### 3.2.3 Partner

The distribution indicates that 48% of customers responded "Yes," while 52% responded "No." This suggests that a significant portion of the customer base has partners. Considering that customers with partners may have different usage patterns or service requirements, their churn decisions could be influenced in unique ways, underscoring the need for further examination.

### 3.2.4 Dependents

The distribution shows that 23% of customers responded "Yes," while 77% responded "No." This indicates that a smaller portion of the customer base has dependents. Customers with dependents may prioritize stability and affordability over service variety, potentially impacting their loyalty and churn behavior. This suggests the importance of considering their specific needs and preferences in retention strategies.

### 3.2.5 Tenure

The distribution spans from 1 to 72 months, with quartiles at 9, 32, and 55 months. The high frequency of customers with short tenures suggests a challenge in early retention efforts. Conversely, customers with longer tenures may exhibit lower churn rates, indicating a potential focus area for retention strategies to enhance long-term loyalty.

### 3.2.6 Phone Service

The distribution reveals that 90% of customers responded "Yes," while 10% responded "No." This high adoption rate of phone services underscores its significance to customers. The 10% without phone service represent a distinct segment that may have different needs or preferences, suggesting the necessity for tailored strategies to address their requirements effectively.

### 3.2.7 Multiple Lines

The distribution shows that 42% of customers responded "Yes," 48% responded "No," and 10% reported no phone service. Customers with multiple lines may indicate higher engagement or a greater need for connectivity, which could impact their churn risk. Understanding the behavior and preferences of this segment is essential for effective churn management strategies.

### 3.2.8 Internet Service

The distribution indicates that 34% of customers have DSL, 44% have Fiber Optic, and 22% have no internet service. The preference for Fiber Optic suggests a demand for high-speed internet among a significant portion of customers. However, churn rates may vary significantly by service type, with potential higher churn among Fiber Optic users due to expectations or

service-related issues. Understanding these dynamics is crucial for devising effective retention strategies tailored to different service types.

### 3.2.9 Device Protection, Tech Support

The distribution for both features shows similar patterns, with approximately 29-34% responding "Yes," 44-49% responding "No," and the remainder being not applicable due to no internet service. These add-on services likely play a role in influencing customer satisfaction and churn, particularly among users who heavily rely on technology. Understanding the implications of these services on churn is crucial for devising effective retention strategies tailored to different customer segments.

### 3.2.10 Contract

The distribution reveals that 55% of customers have month-to-month contracts, 21% have one-year contracts, and 24% have two-year contracts. The prevalence of month-to-month contracts suggests lower commitment barriers, which may contribute to higher churn rates. Understanding the impact of contract duration on churn behavior is essential for implementing retention strategies that encourage long-term customer loyalty.

### 3.2.11 Paperless Billing

The distribution indicates that 59% of customers prefer paperless billing, while 41% do not. This preference suggests a tech-savvy customer base, which may be associated with distinct churn behaviors. Understanding the relationship between paperless billing preference and churn is crucial for tailoring retention strategies to effectively address the needs and expectations of this customer segment.

### 3.2.12 Payment Method

The distribution reveals diverse payment preferences among customers: 34% prefer Electronic Check, 22% prefer Mailed Check, 22% prefer Bank Transfer, and 22% prefer Credit Card. Such diversity suggests the presence of varied customer segments. Electronic check users, for instance, may exhibit different churn patterns, potentially influenced by factors such as convenience or demographics. Understanding these differences can inform targeted retention strategies tailored to each payment preference group.

### 3.2.13 Monthly and Total Charges

Monthly Charges range from $19 to $119, with a median of $70, while Total Charges vary widely, ranging from $19 to $8,684, often correlating with tenure. Higher monthly and total charges may be associated with higher churn, particularly if customers perceive the value of the

services received does not justify the cost. Understanding the relationship between pricing and churn behavior is essential for optimizing pricing strategies and improving customer retention efforts.

### 3.2.14 CLTV (Customer Lifetime Value)

The distribution ranges from 2,003 to 6,500, with a median Customer Lifetime Value (CLTV) at 4,527. Higher CLTV values may correlate with lower churn rates, suggesting that nurturing high-value customers is essential for reducing churn and fostering long-term relationships. Understanding the relationship between CLTV and churn behavior can inform targeted retention strategies aimed at maximizing the value of high-LTV customers.

## 4 Exploratory Data Analysis (EDA)

The EDA stage of our study is essential for visualizing and understanding the data, identifying patterns and anomalies, and formulating hypotheses about the factors that influence customer churn. Various graphical representations provide insights into the relationships between customer characteristics and their churn status. The analysis includes violin plots for contract types and their association with churn, as well as other visualizations detailed in the presentation.

### 4.1 Churn Rate by Payment Method

The exploratory data analysis of the payment_method reveals insightful trends in customer behavior and retention. The initial graph displays a preference hierarchy among payment options, with bank transfers leading the way, followed by credit card usage, mailed checks, and electronic checks. This distribution may reflect customer preferences or could indicate the telecom company's more accessible or endorsed payment methods.

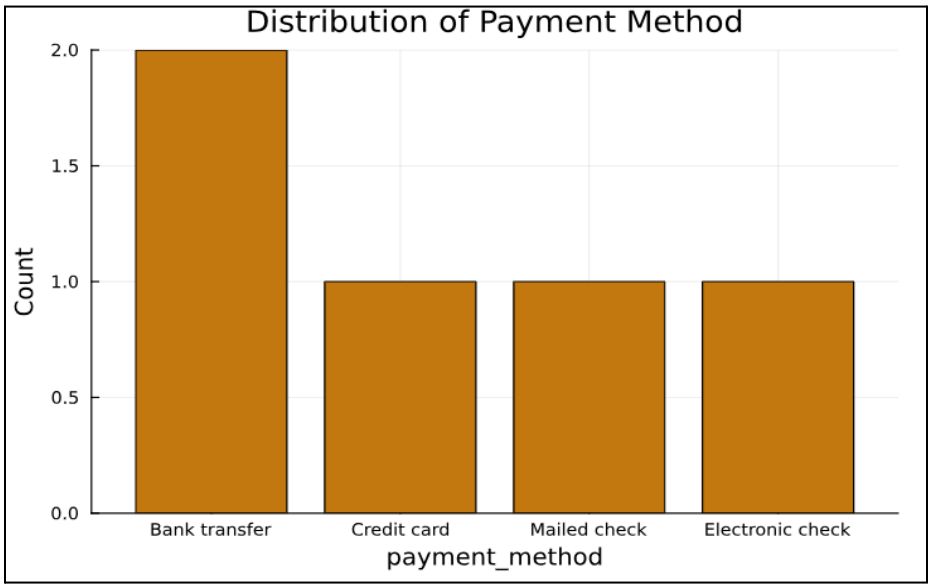


Figure 1: Distribution of Payment Method

When examining the relationship between payment methods and customer churn, the second graph offers a striking contrast. Customers utilizing electronic checks show a higher propensity to churn, which is indicated by the black segments. In stark contrast, those employing automated payment methods like bank transfers and credit cards exhibit a lower churn rate. This

discrepancy suggests that the ease and lack of friction provided by automated payments possibly contribute to higher customer satisfaction and retention.

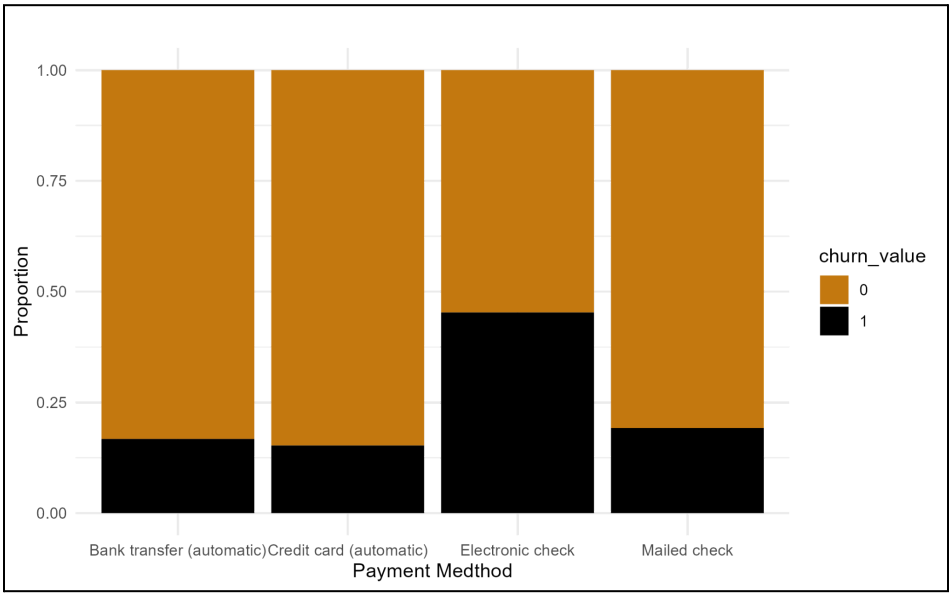


Figure 2: Distribution of Payment Method by Churn Value

Moreover, the choice of an automatic payment method might signal a deeper level of trust and commitment to the service provider. These customers could potentially exhibit less sensitivity to price changes or service disruptions, correlating with a lower churn rate. The preference for automated payments over manual ones like mailed checks might also reflect broader trends such as technological comfort and economic stability among different customer demographics.

In light of these findings, it is recommended to encourage the adoption of automatic payment methods as a strategic approach to mitigate customer churn.

### 4.2 Churn Rate by Tenure Months

The analysis of tenure months in the dataset showcases customer loyalty and retention trends over time. The first graph, illustrating the distribution of tenure months, reveals a high frequency of customers with a short tenure, tapering off as the tenure length increases. This could indicate that a considerable number of customers try the telecom services but a smaller percentage commit to longer-term use.

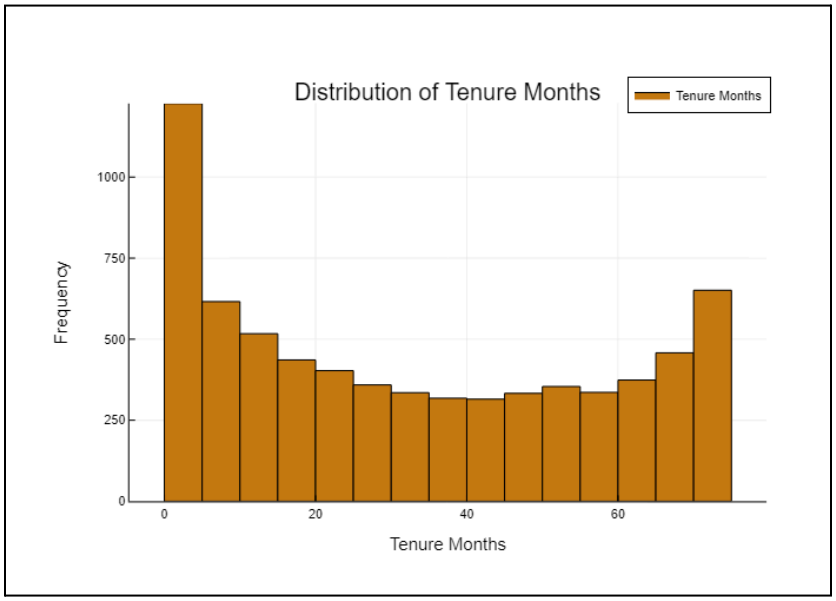


Figure 3: Distribution of Tenure Months

This trend becomes more nuanced when we overlay churn status, as shown in the second graph. Here, a large portion of customers with very low tenure months have churned (black bars), suggesting that the risk of churn is higher at the beginning of the customer life cycle. Conversely, customers with longer tenure are predominantly in the 'No Churn' category (orange bars),

indicating that the longer customers stay with the telecom provider, the less likely they are to churn.

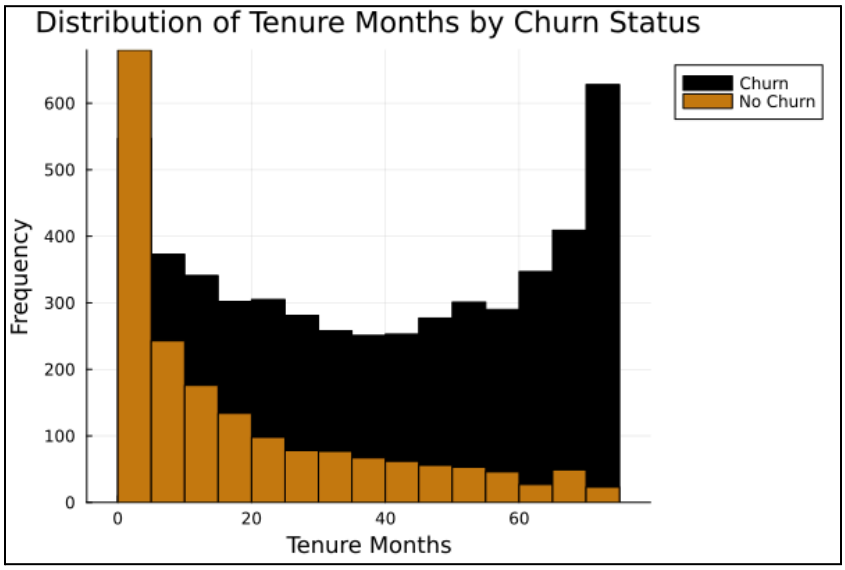


Figure 4: Distribution of Tenure Months by Churn Value

This pattern underscores the importance of the initial service period in customer retention. It implies that efforts to enhance customer experiences early on could be crucial in reducing churn rates. Furthermore, the data suggests that once customers cross a certain tenure threshold, they are more likely to remain loyal to the service.

Considering the findings, telecom companies might focus on improving customer onboarding processes, providing exceptional service in the initial months, and exploring loyalty programs aimed at customers within the critical early tenure window to bolster retention.

### 4.3 Churn Rate by Monthly Charges

The first visualization, Figure 5, outlines the distribution of monthly charges among the telecom company's customers. This violin plot highlights a concentration of customers at the lower end of the monthly charges spectrum, expanding to a broader distribution as charges increase. The widest part of the distribution suggests a substantial number of customers with mid-range monthly charges, with fewer customers at the highest end of the scale. This could point to a variety of service plans catering to different customer segments.

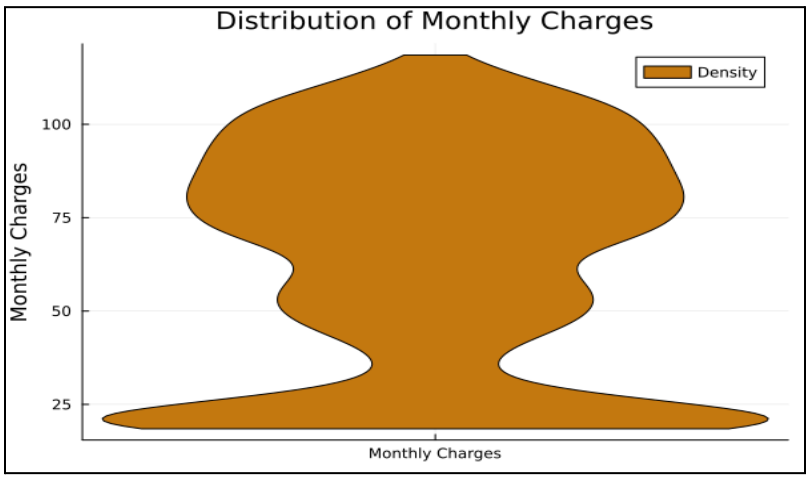


Figure 5: Distribution of Monthly Charges

Figure 6 delves into the correlation between monthly charges and customer churn. The distribution of monthly charges for those who have churned forms a stark contrast with those who haven't. Customers who have churned show a denser concentration at the higher end of the monthly charges spectrum, indicating that higher charges may correlate with a higher likelihood

of churn. In contrast, the distribution for customers who haven't churned is skewed towards the lower and middle ranges of monthly charges.

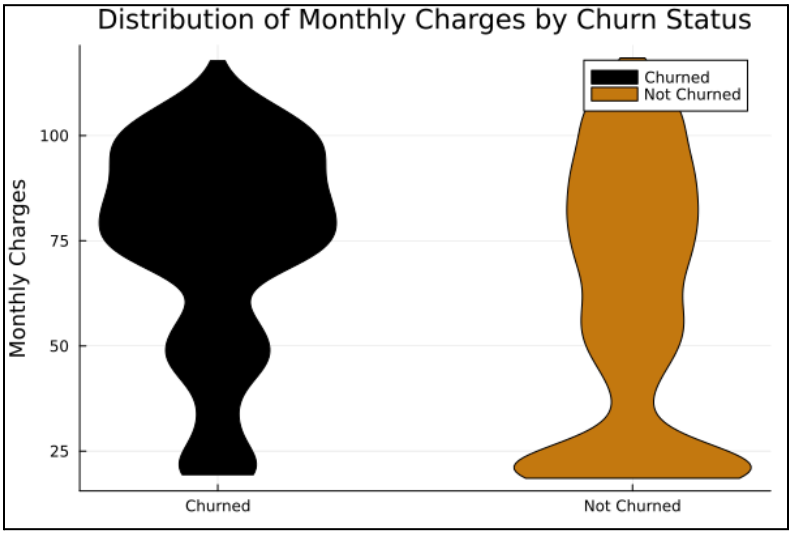


Figure 6: Distribution of Monthly Charges by Churn Value

This disparity suggests that price sensitivity could be a significant factor in customer retention, with higher monthly charges potentially prompting customers to consider other options in the competitive telecom market. This information could be particularly valuable in devising pricing strategies that balance revenue and customer retention. Telecom companies might leverage this insight by reviewing their pricing structures, especially for services at the higher end of the price range, or by introducing more competitive pricing tiers to improve customer loyalty.

### 4.4 Churn Rate by Service Features

The distribution of tech support availability among the telecom company's customers, as shown in Figure 7, suggests several aspects of customer service engagement. A larger number of customers do not subscribe to tech support, with a smaller yet significant portion opting in for this service. Interestingly, there is also a noticeable subset of customers who do not have internet service, hence the tech support service is not applicable to them.

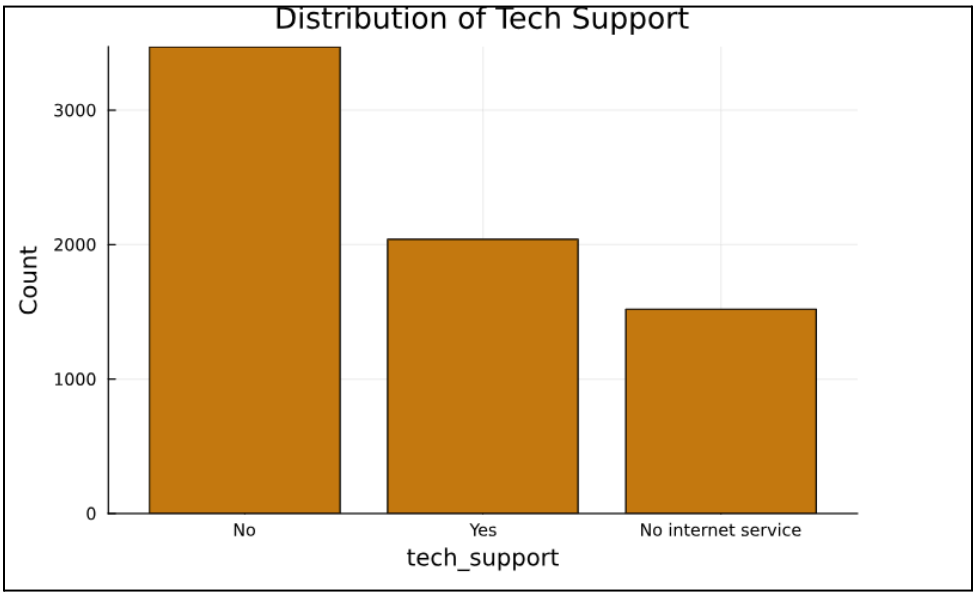


Figure 7: Distribution of Tech Support

When we cross-reference tech support with churn status in Figure 8, there's a discernible pattern in customer behavior. Among the churned customers, a large segment did not have tech support, which could imply a correlation between the lack of tech support and higher churn rates. In

contrast, the pie chart representing retained customers shows a more balanced distribution, with a considerable proportion of customers having tech support in place.

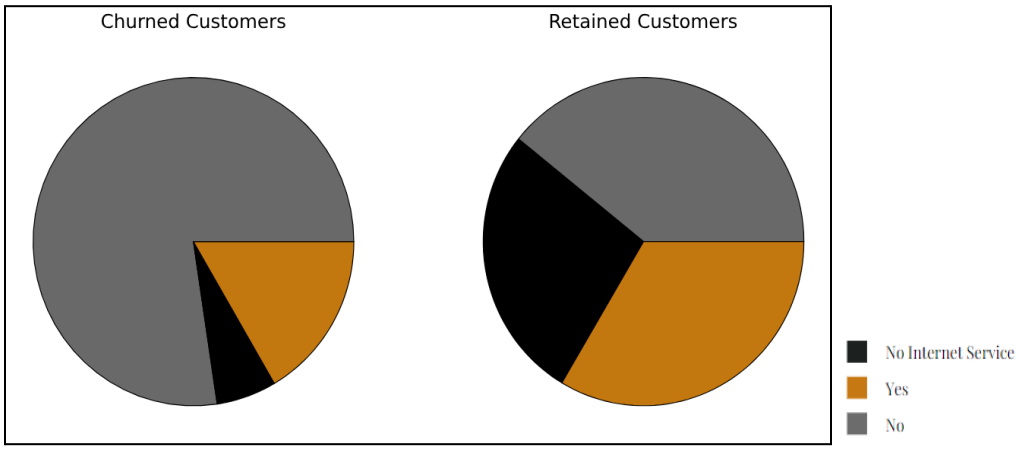


Figure 8: Distribution of Tech Support by Churn Value

This analysis suggests that the availability and possibly the quality of tech support could be influential in customer retention. It highlights a potential opportunity for the telecom company to focus on its tech support services as a means to enhance customer satisfaction and reduce churn. The company could explore strategies such as improving the accessibility and responsiveness of tech support, or offering it as part of a value-added package to encourage customer loyalty. Given the data, it would be prudent for the telecom provider to consider the role of tech support in its overall service offering, possibly as a pivotal point in the customer's decision-making process.

### 4.5 Contract Types and Churn

The bar chart in Figure 7 presents the distribution of customer contracts within the telecom company, displaying the prevalence of month-to-month contracts over one- and two-year

contracts. This indicates a customer base that favors flexibility, as month-to-month contracts typically offer the freedom to change or cancel services without long-term commitments or penalties.

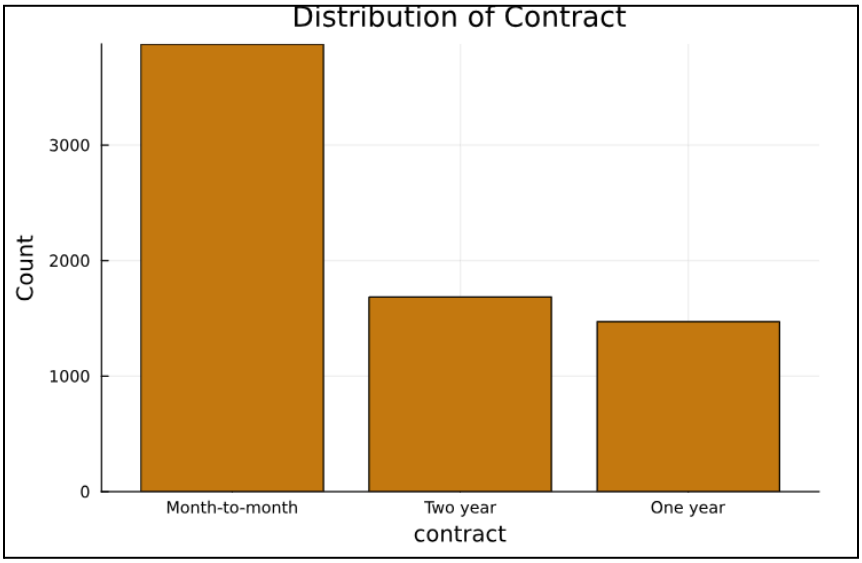


Figure 9: Distribution of Contract

The violin plots in the subsequent figures provide a deeper analysis, correlating contract types with customer churn. In the month-to-month contract category, there is a significant spread in churn, with a substantial number of customers both leaving (churned) and staying (not churned). This suggests that while month-to-month contracts are popular, they may also be volatile, with customers feeling less tied to the service and possibly more sensitive to price changes or competitive offers.

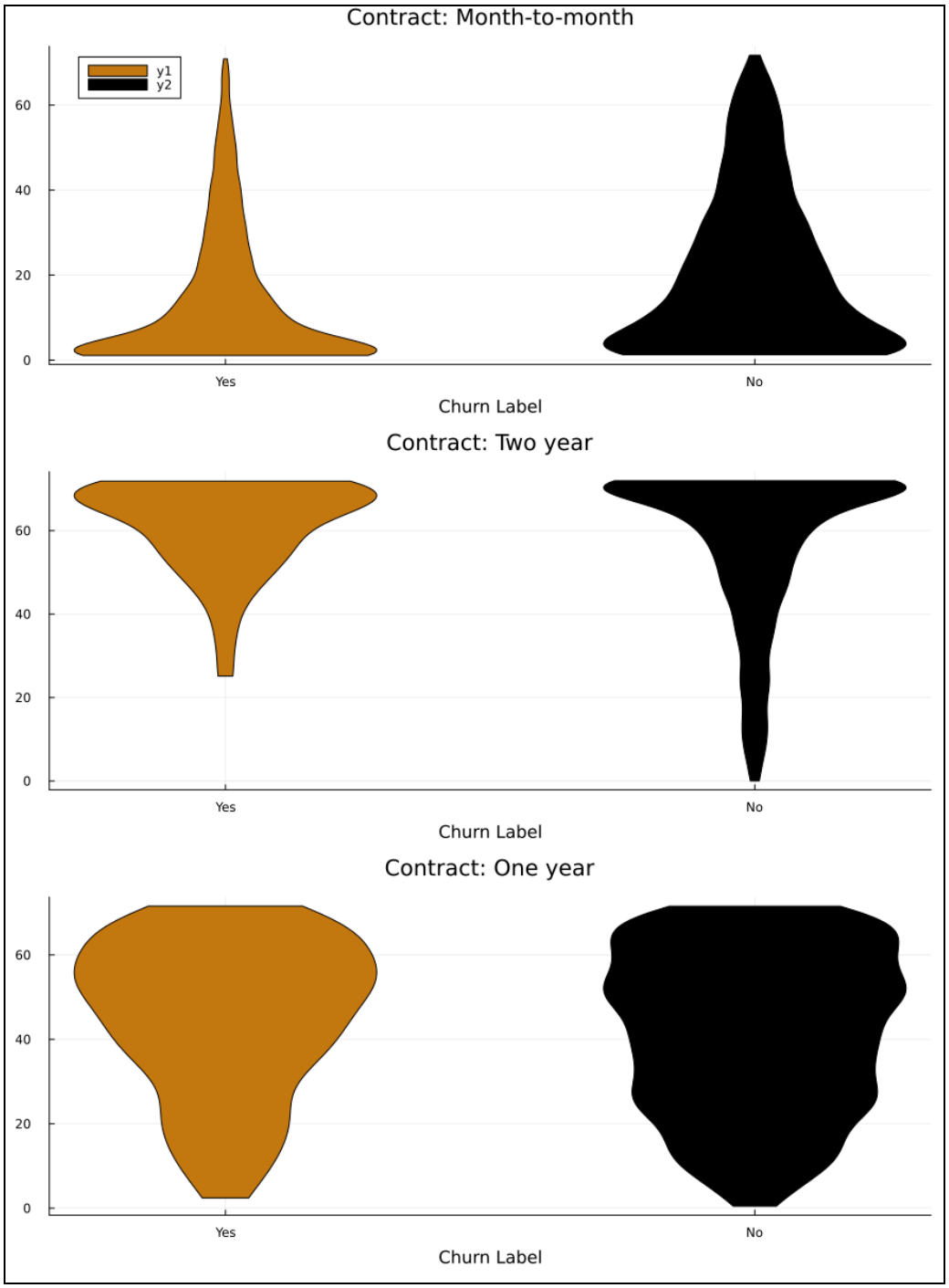


Figure 10: Distribution of Contract by Churn Value

For two-year contracts, the churn distribution is narrow, indicating that customers with longer contract terms are less likely to churn. This could reflect a more committed customer base or satisfaction with the service, considering the longer-term commitment. Similarly, one-year contracts show a narrower churn distribution than month-to-month contracts, though not as pronounced as two-year contracts, suggesting an intermediate level of commitment and churn risk.

The patterns across these figures point to the significance of contract length in customer retention strategies. The data suggests that while month-to-month contracts might attract customers due to their flexibility, they could also lead to higher churn rates. On the other hand, longer contracts seem to be associated with greater customer retention.

For telecom companies, these insights could indicate an opportunity to create a balance between offering flexibility to attract customers and providing incentives for longer-term contracts to maintain a stable customer base. Strategies such as loyalty benefits, gradual price discounts over time, or other value-added services could be effective in encouraging customers to commit to longer contract periods. Additionally, a closer look at the reasons behind the churn in different contract types could help tailor specific retention efforts.

### 4.6 Churn Rate by CLTV

Figure 12 depicts the distribution of Customer Lifetime Value (CLTV) across the telecom company's customer base through a Kernel Density Estimate (KDE) plot. This density plot suggests that CLTV has a multimodal distribution, with peaks around lower as well as higher CLTV values. This might indicate that there are distinct groups within the customer base—those who bring in varying levels of long-term value to the company. A concentration around lower values may represent a large segment of customers with lower projected revenue contribution, whereas higher values could represent more valuable, possibly more loyal customer segments.

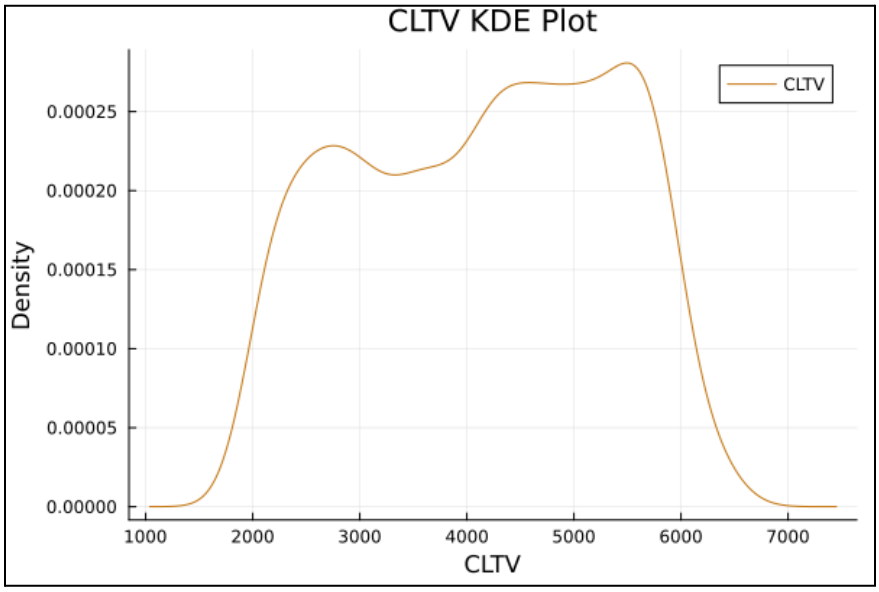


Figure 11: Distribution of CLTV

In Figure 13, the distribution of CLTV is segmented by churn status, which provides insights into the relationship between CLTV and customer retention. Customers who have not churned (No

Churn) appear predominantly in the higher CLTV segments, suggesting that customers with higher lifetime value are more likely to remain with the company. Conversely, customers who have churned (Churn) are more frequently found in the lower to middle CLTV segments.

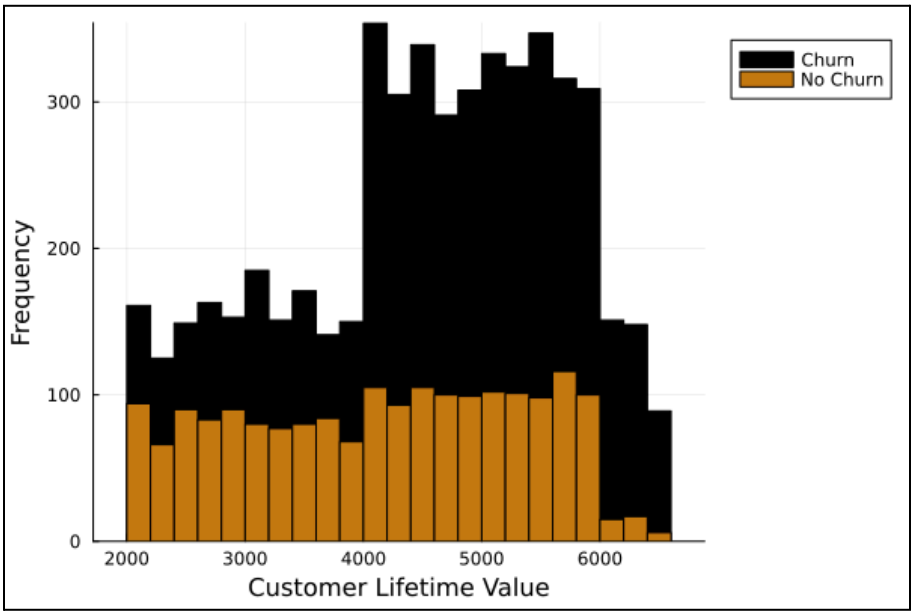


Figure 12: Distribution of CLTV by Churn Value

This information is particularly important for the telecom company's retention strategy. It implies that customers with higher CLTV are crucial as they are more inclined to stay, making them key targets for retention efforts. It could also suggest that initiatives to enhance customer value, such as upselling or cross-selling, may simultaneously increase CLTV and reduce churn.

The data supports the idea that strategies aimed at increasing the CLTV could be beneficial not only for immediate revenue but also for long-term customer retention. Understanding the factors that contribute to a higher CLTV could enable the company to foster loyalty and reduce turnover. This might include analyzing the service features, customer support quality, or pricing models that correlate with higher CLTV. Additionally, identifying why customers with lower CLTV are churning could offer insights into potential service improvements or changes needed to enhance overall customer satisfaction and value.

### 4.7 Churn Rate by Gender

Figure 14 illustrates the distribution of gender within the customer base, indicating a nearly equal division between male and female customers. This equitable distribution suggests that gender does not play a significant role in the uptake of telecom services within this dataset.

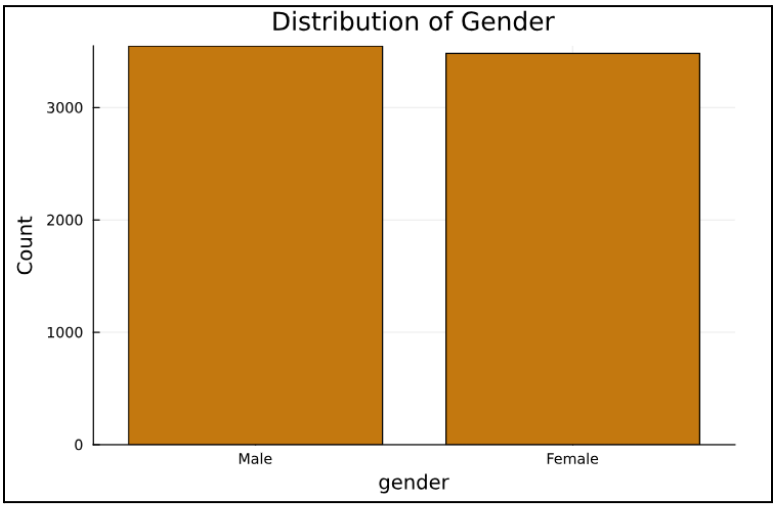


Figure 13: Distribution of Gender

Figure 15 shows the churn rate distributed by gender, with the proportion of churned and not churned customers represented for both genders. Interestingly, the churn rates between male and female customers are similar, as indicated by the proportionate segments within each gender category. This implies that gender is not a distinguishing factor in the likelihood of customers churning.

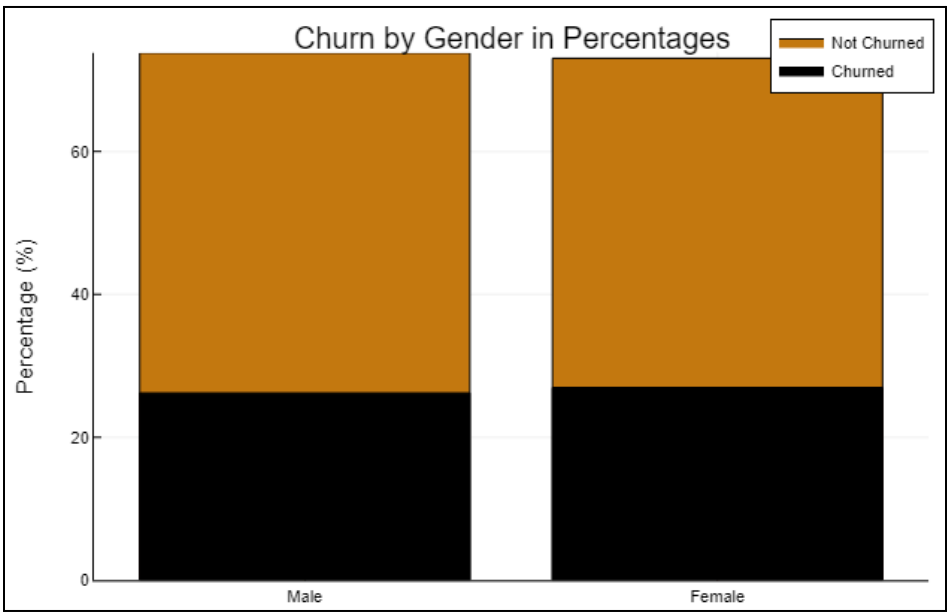


Figure 14: Distribution of Gender by Churn Value

From a strategic standpoint, these findings indicate that gender should not be a primary segment in targeting retention initiatives. Since the churn rate does not significantly differ by gender, efforts and resources might be better allocated towards factors that have a stronger correlation with churn.

### 4.8 Churn Rate by Total Charges

The first histogram, displaying the distribution of total charges, shows a right-skewed distribution, indicating that a large number of customers have accrued lower total charges, while fewer customers have very high total charges. This skewness might reflect a customer base with many new or short-term customers who haven't had the opportunity to accumulate high charges, or it might indicate that most customers are using lower-cost service options.

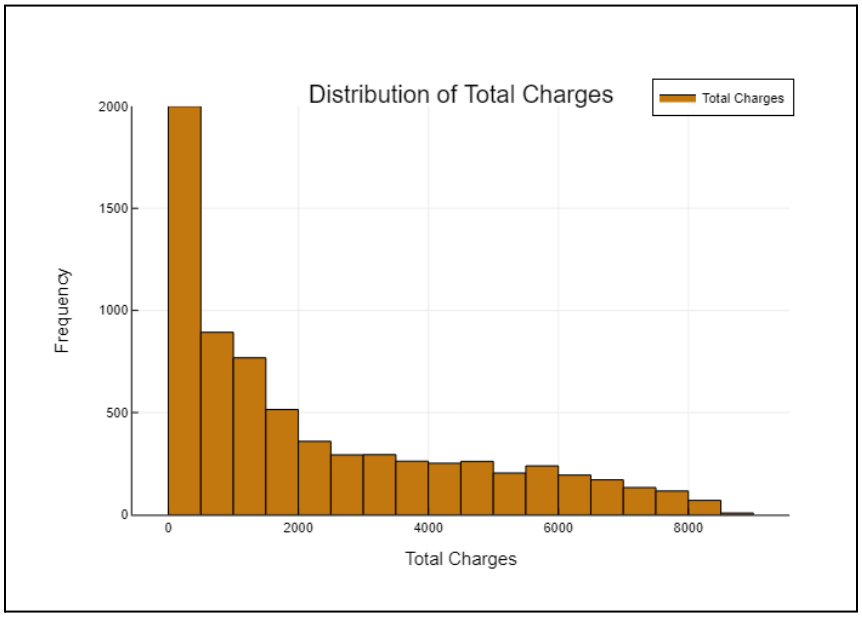


Figure 15: Distribution of Total Charges

In the second histogram, when we examine the distribution of total charges by churn status, we observe that a higher frequency of churn occurs at the lower end of total charges. This could suggest that customers who have not been with the company long enough to accrue higher

charges—or those who subscribe to lower-cost services—are more likely to churn. On the other hand, customers with higher total charges, potentially indicating longer tenure or higher service level usage, tend to have a lower churn rate.

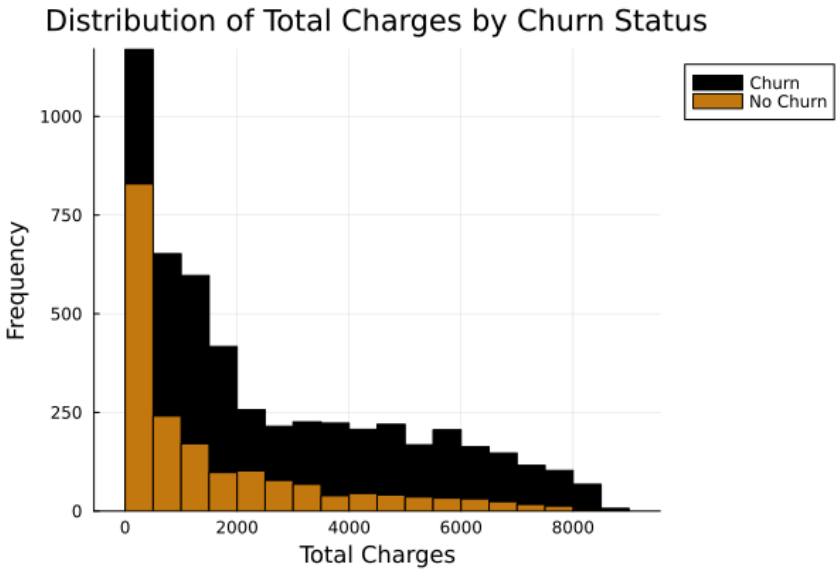


Figure 16: Distribution of Total Charges by Churn Value

These observations suggest that customer loyalty may increase with time or with the use of more premium services, as indicated by higher total charges. For the telecom company, this might imply that focusing on long-term customer engagement and value-addition could be key strategies for reducing churn

### 4.9 Correlation between variables and churn_value

This heatmap represents the correlation coefficients between different variables and the churn value, which indicates whether a customer has churned. The scale on the right shows the range of the correlation coefficient, where 1 indicates a perfect positive correlation, -1 indicates a perfect negative correlation, and 0 indicates no correlation.

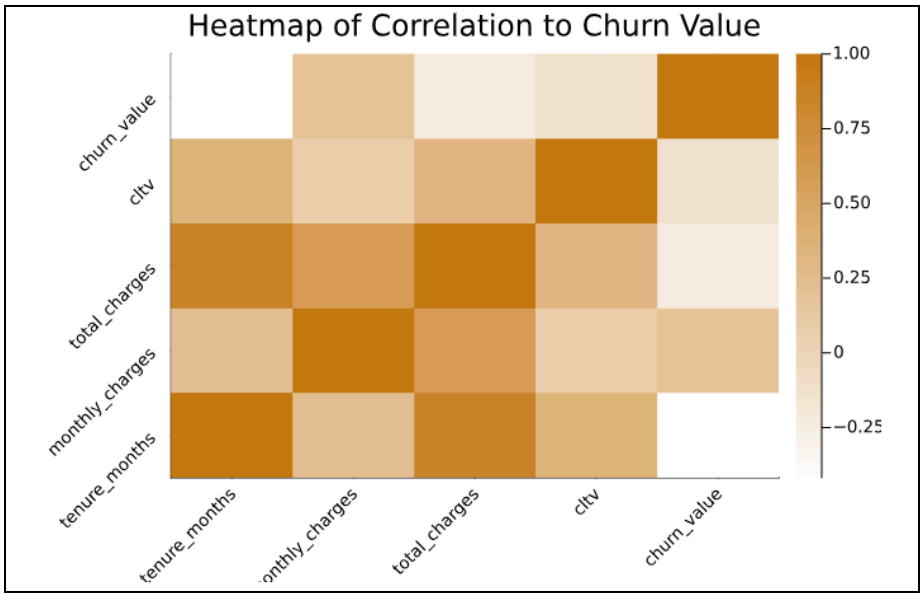


Figure 17: Correlation Heat Map between Numerical Variables and Churn Value

The strongest positive correlations with churn value appear to be with monthly charges and total charges, suggesting that as these charges increase, so does the likelihood of churn. This could indicate a sensitivity to price increases or higher charges among the customer base.

Conversely, tenure months shows a negative correlation with churn value, which means that the longer a customer stays with the company (higher tenure), the less likely they are to churn. This aligns with the common understanding that longer customer relationships often equate to increased loyalty.

The correlation with Customer Lifetime Value (CLTV) is not as strongly defined in either direction according to this heatmap, which could suggest that the calculation of CLTV might encompass factors both positively and negatively related to churn, making its direct relationship with churn less clear-cut.

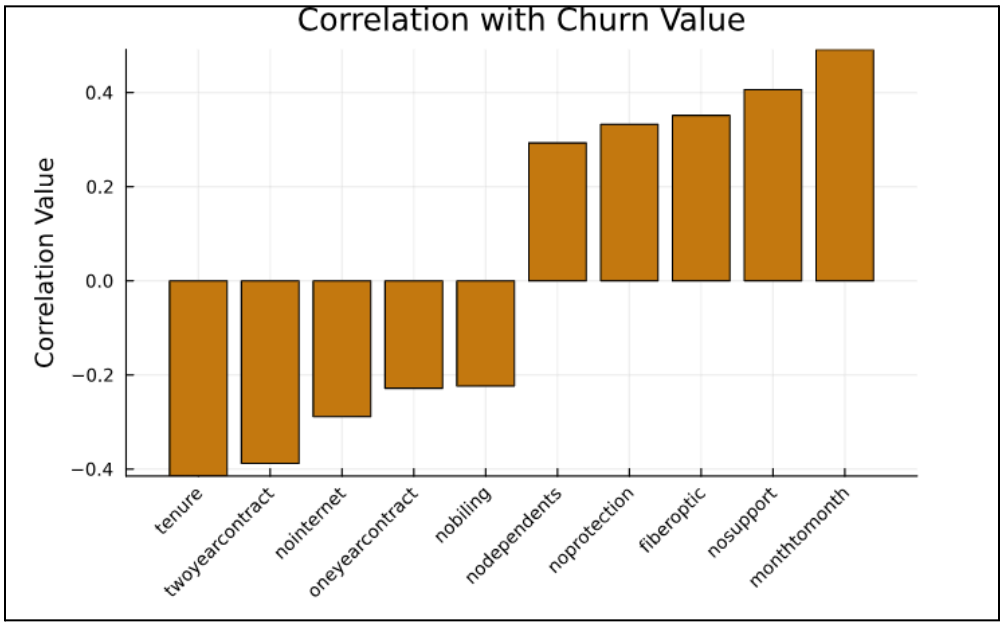


Figure 18: Variables with highest Correlation with Churn Value

This bar chart illustrates the correlation of various factors with the churn value of customers in the telecom dataset. The variables on the left side of the chart with negative correlation values — 'tenure', 'twoyearcontract', and 'oneyearcontract' — inversely relate to customer churn. This indicates that longer customer tenure and longer contract terms (one or two years) are associated with lower churn rates. It's clear from these correlations that customers who commit to a longer relationship with the telecom company are less likely to leave, suggesting that long-term contracts could be a key to customer retention.

On the other hand, the variables on the right with positive correlation values — 'monthtomonth', 'nosupport', 'fiberoptic', 'noprotection', 'nodependents', and 'nobilling' — have a direct relationship with churn. This suggests that customers with month-to-month contracts, without tech support, using fiber optic internet services, without a protection plan, with no dependents, or with no automatic billing are more likely to churn. These factors could be areas where the company may need to investigate further, potentially improving services or changing policies to address the increased churn risk. For example, customers on month-to-month contracts might value flexibility but could be swayed by incentives to switch to longer-term contracts, while those without tech support might benefit from improved service offerings in this area.

### 4.10 Reasons to Churn

The bar chart categorizes the reasons for customer churn, providing a visual summary of the factors that lead customers to leave the telecom company. The most prominent reason for churn is due to competitors, suggesting that customers are finding more attractive offers or services with other providers. This could reflect on the competitiveness of the telecom market and highlights the need for the company to stay competitive in terms of pricing, service features, and customer engagement.

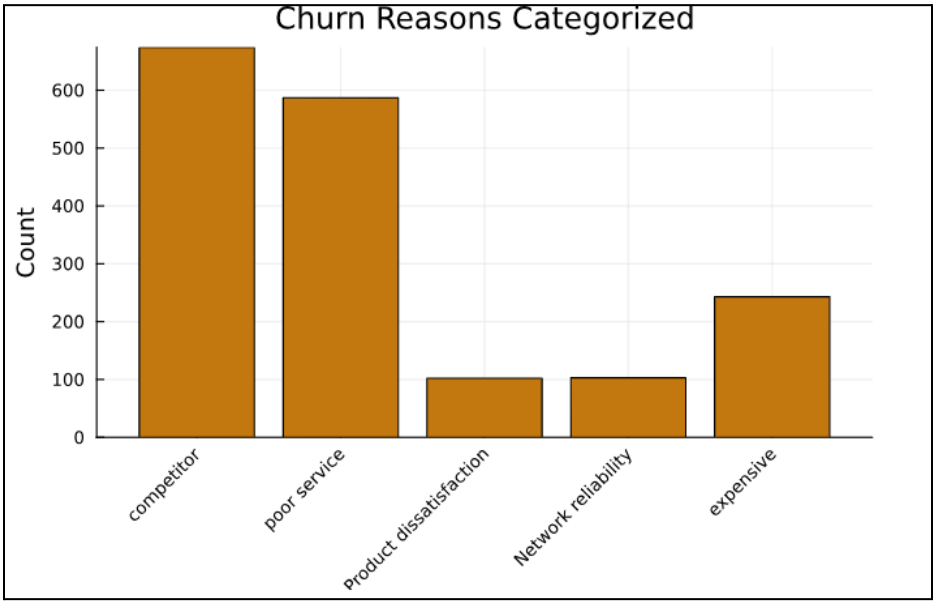


Figure 19: Reasons to Churn

The second most common reason for churn is poor service. This is a critical area for any service provider, as it directly impacts customer satisfaction and loyalty. The company might need to delve deeper into the specifics of what constitutes poor service for these customers — whether it

is customer support, technical issues, or service disruptions — and address these areas promptly to improve overall service quality.

Product dissatisfaction and network reliability have a lower count compared to the first two reasons but are still significant. Product dissatisfaction could stem from the telecom services not meeting customer expectations or needs, suggesting a gap in the product-market fit. Network reliability issues might indicate technical problems that lead to service interruptions or poor quality of service, both of which are crucial for customer retention.

Lastly, the category of 'expensive' indicates that some customers find the pricing of the services too high, which could be a factor in the earlier noted positive correlation between churn and monthly charges. This suggests price sensitivity among a segment of the customer base, and it might be beneficial for the company to review their pricing structures or introduce more varied service tiers.

Even though these reasons for churn cannot be used predictively as they are obtained at the point of churn, they are invaluable for retrospective analysis and strategy development. Understanding why customers leave enables the company to take proactive steps to rectify issues and prevent future churn.

## 5 Methods

We strived for a methodological approach to ensure a robust and thorough examination of the factors influencing customer retention and attrition. We begin by preprocessing the data to serve as a staging area for our model development, followed by the construction of baseline models spanning various machine learning families. Recognizing the inherent challenges posed by class imbalance, we then refine our models to address this disparity, enhancing the predictive accuracy across all models considered. Further model optimization is achieved through specific parameter tuning, tailored to the unique characteristics of each model. Throughout this process, we are guided by a dual objective: to extract meaningful insights from the data and to develop predictive models with practical applicability in mitigating customer churn. This section elucidates our methodological framework, detailing the rationale behind our choice of methods, the potential issues encountered, such as collinearity, and the strategies employed to address these challenges. Moreover, it outlines our approach to model evaluation, ensuring that our analysis is both rigorous and relevant to the business context.

### 5.1 Preprocessing

The first step in our analysis involves data preprocessing to ensure the quality and usability of the dataset for modeling. We began with a dataset comprising 7,043 observations and 33 features, encompassing information on customer interactions, demographic details, service usage, and churn status.

Our preprocessing involved the elimination of redundant features that contribute little to the predictive power of our models. Specifically, we removed customer identifiers and count metrics, which are unique to each observation and thus do not provide generalizable insights. Similarly, location-based features such as country and state were dropped due to their lack of variability (all observations pertained to customers in California, United States), as well as the textual representation of geographical coordinates (Latitude-Longitude), given that numerical representations (latitude and longitude) were available and more suitable for our analytical methods. Redundant features directly related to churn (e.g., churn label, churn score) were also excluded, as our goal is to predict churn without using information that would only be available after churn has occurred. After this step, our dataset was refined to 7,032 observations, each described by 21 features. This streamlined dataset is characterized by a predominance of categorical variables.

To address the challenges posed by the mix of categorical and numerical variables, our Machine Learning pipeline incorporated a one-hot encoding strategy for categorical variables. This technique transforms categorical variables into a series of binary columns, each representing the presence or absence of a category level. For numerical variables, we applied standardization to ensure that all features have a mean of zero and a standard deviation of one, thus mitigating potential biases arising from scale differences.

By addressing missing values, eliminating redundant features, and appropriately transforming

variable types, we aimed to construct a robust foundation for predictive modeling. These steps are critical for developing models that not only perform well on historical data but also generalize effectively to new, unseen data.

### 5.2 Building Baseline Models

To capture the nature of customer churn, we construct baseline models across a diverse array of machine learning families, including logistic regression, decision trees, and ensemble methods like random forests and gradient boosting machines. This approach enables us to benchmark performance across model types, offering a panoramic view of predictive capabilities and inherent biases. The choice of models is driven by the need for both interpretability and predictive power, ensuring our findings are accessible to stakeholders while maximizing accuracy.

### 5.3 Addressing Class Imbalance

Recognizing the potential skew in churn rates—a common challenge in churn prediction—we employ techniques to balance the classes, such as synthetic minority oversampling technique (SMOTE) and adaptive synthetic sampling (ADASYN). These methods help in augmenting the minority class, thereby preventing model bias towards the majority class and enhancing the sensitivity towards churn prediction. Balancing the classes is crucial for avoiding misleading

accuracy metrics and for ensuring that our models genuinely capture the nuances of customer churn.

| Data set | Minority Class | Majority Class | Total Observations |
|---|---|---|---|
| Imbalanced | 1869 (27%) | 5163 (73%) | 7032 (100%) |
| Undersampling | 1869 (50%) | 1869 (50%) | 3738 (100%) |
| Oversampling | 5163 (50%) | 5163 (50%) | 10326 (100%) |

Table 1: Counts and Distribution (%) of observations in minority and majority classes

### 5.4 Model-Specific Parameter Tuning

With baseline models established, we go into model-specific parameter tuning to refine and optimize performance. This involves hyperparameter optimization techniques such as grid search and random search, tailored to each model's unique architecture. This iterative process is critical for uncovering the optimal configuration that balances bias and variance, thereby enhancing the predictive accuracy and generalizability of our models.

### 5.5 Potential Challenges

We have identified three principal challenges: the presence of a large feature set, class imbalance, and the predominance of categorical features. We try to mitigate their impact by employing different methods as a remedy. Through these targeted strategies, we address the principal

challenges encountered in our analysis, ensuring that our methodological framework is both resilient and adaptive.

### 5.5.1 Large Feature Set

The IBM Telco data set encompasses a wide array of features. While this breadth of information is valuable, it also poses the risk of model overfitting and increased computational complexity. To counteract these issues, we employ feature reduction techniques such as regularization and subset selection methods. Regularization techniques - like L1 Lasso - add a penalty term to the loss function, encouraging simpler models and thereby reducing overfitting. Subset selection (e.g. forward stepwise selection,) on the other hand, involves selecting a subset of relevant features for model training, enhancing model interpretability and reducing computational demands. These approaches allow us to distill the most predictive features from the dataset, focusing our analysis on variables with the greatest impact on churn.

### 5.5.2 Class Imbalance

Class imbalance is a prevalent issue in churn prediction, where the number of customers who churn is typically much lower than those who stay. Precisely, only 27% of customers end up churning after the three month records. This imbalance can lead to model bias, where predictions are disproportionately skewed towards the majority class. To address this challenge, we

implement strategies such as undersampling the majority class and oversampling the minority class. In undersampling, we first select all members of the minority class, and then randomly select an equal number of members from the majority class. This reduces the total number of observations available for the training and testing sets, but gives us a perfectly balanced data set based on our target variable class. On the other hand, in oversampling, we select all members of the majority class and then try to match this number of observations with the minority class. While there are many methods to do so, we chose to implement Synthetic Minority Over-sampling Technique (SMOTE), which synthetically generates similar observations for the minority class (based on near neighbor algorithms) in order to increase the number of observations of the minority class. Naturally, oversampling will give us more observations than we started with. We believe that these methods will considerably enhance the model's ability to learn from both classes equally, improving predictive performance on the minority class without compromising overall accuracy.

### 5.5.3 Predominance of Categorical Features

The dataset's abundant categorical features, representing various customer demographics and service options, introduce complexity due to their non-numeric nature. Traditional models might struggle to interpret these features effectively, leading to suboptimal performance. To mitigate this issue, we prioritize tree-based models, such as Random Forests and Gradient Boosted Trees, which are inherently equipped to handle categorical data. Tree-based models split data along the features, making them adept at managing both numerical and categorical variables without the

need for extensive preprocessing. This approach ensures that the rich information encapsulated in categorical features is effectively utilized, enhancing the models' predictive accuracy and interpretability.

### 5.5.4 Evaluating Model Performance

The effectiveness of our predictive models is gauged through a multifaceted evaluation framework, incorporating accuracy, precision, recall, F1-score, and the area under the receiver operating characteristic curve (AUC-ROC). This comprehensive assessment ensures that we not only focus on overall accuracy but also consider the balance between sensitivity and specificity, crucial for actionable churn prediction.

By adhering to this methodological blueprint, we aim to provide a robust and nuanced understanding of customer churn dynamics, grounded in statistical rigor and tailored to the real-world context of the telecommunications industry. This endeavor not only illuminates the pathways to enhancing customer retention strategies but also showcases the potential pitfalls in churn prediction and the strategic measures taken to navigate these challenges.

## 6. Modeling

In this section of our paper we delve into the construction and evaluation of machine learning models designed to predict customer churn for a telecommunications company. Our objective is

to identify customers likely to leave the company within the next three months, using the `churn_status` variable as our target. This endeavor begins with the establishment of a baseline model, which serves as a point of reference for assessing the effectiveness of more sophisticated modeling techniques introduced later in our analysis.

### 6.1 Logistic Regression Models

The choice of a baseline model is guided by the need for a straightforward yet effective starting point that sets a benchmark for performance. In this instance, we selected Logistic Regression, a widely acknowledged method for binary classification problems due to its simplicity, interpretability, and efficiency in dealing with binary outcomes like our churn_status variable. To construct our baseline model, we allocated 70% of our dataset for training and reserved the remaining 30% for testing. This split ensures that our model learns from a substantial portion of the data while also being evaluated on a separate set to gauge its generalization capability.

#### 6.1.1 The Baseline Model

Utilizing Julia's MLJ framework, we orchestrated our modeling pipeline, incorporating the LogisticClassifier from the MLJLinearModels package to predict the probability of a customer churning (churn_status = 1). Logistic Regression models the probabilities of the default classes (churned or not churned) using a logistic function, thereby providing a probabilistic foundation for binary classification. The decision threshold was set at 0.5 as a typical starting point for

classification models. This threshold signifies that when the model predicts a churn probability of 50% or higher, the customer is classified as likely to churn.

The performance of our baseline Logistic Regression model was evaluated using accuracy, precision, and recall. Baseline accuracy of 74% indicates the proportion of total predictions (both churned and not churned) our model got right. While useful, accuracy alone can be misleading, especially in imbalanced datasets where one class significantly outnumbers the other. Precision, the model's ability to correctly identify true positives from all the positive predictions made, was at 86%. In our context, it means that when our model predicts a customer will churn, it is correct 86% of the time. Lastly, recall came at 60%, which tells us that our model successfully identifies 60% of all customers who will churn. Notably, the area under the curve is significantly high at 0.86.

In the next sections, we will explore advanced modeling techniques and strategies aimed at enhancing these initial results. Our focus will be on not only improving model performance but also ensuring that our approach remains grounded in solid statistical principles and is interpretable within the business context. Through iterative refinement and the application of machine learning best practices, we aim to develop a robust predictive model that can serve as a valuable tool in the company's customer retention efforts.

### 6.1.2 Moving the Threshold

The imbalance in our dataset, with 27% of the customers having churned and 73% not churned, introduces challenges in accurately predicting churn. Hence, adjusting the decision threshold becomes a strategic method to enhance model sensitivity towards the minority class without sacrificing overall accuracy.

To address this imbalance, we experimented with various thresholds, ranging from 0.1 to 0.8, to identify an optimal value that balances precision and recall in a manner conducive to our specific business context. This process involves running the logistic regression model multiple times, each with a different threshold for classifying a customer as likely to churn. The goal is to find a threshold that improves the model's ability to identify true positives (actual churn) without an unacceptable increase in false positives (incorrectly predicted churn).

Our investigation revealed that adjusting the threshold to 0.2 markedly enhanced the model's performance metrics. The figure below shows the precision vs. recall trade-off based on the value of the threshold chosen.

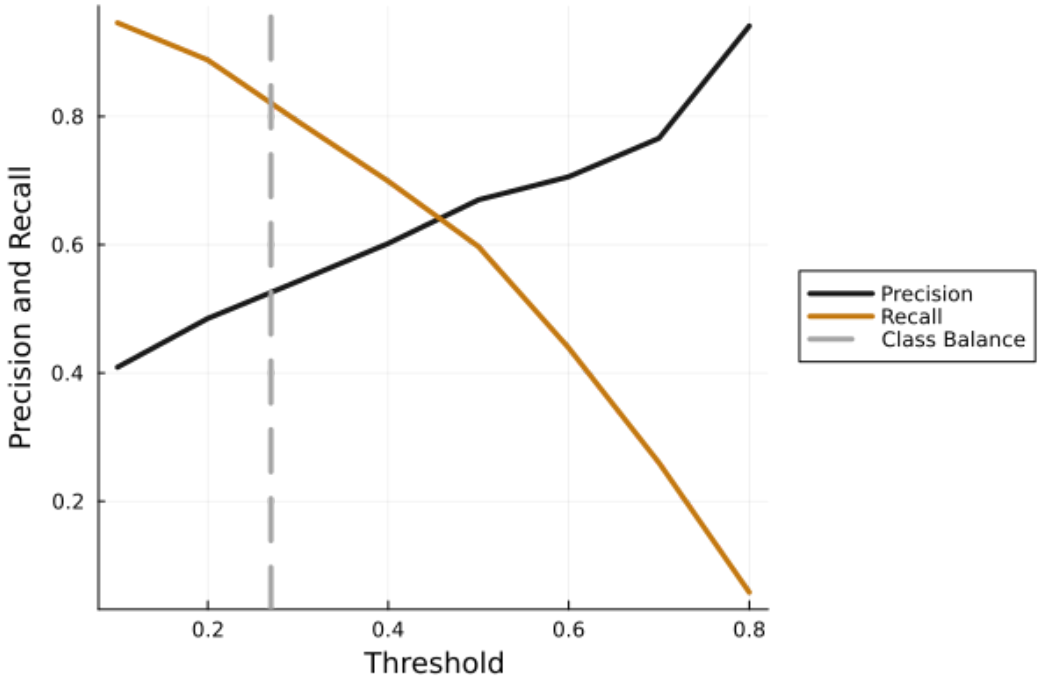


Figure 20: Precision and Recall for Various Threshold Values

Using the optimal threshold value at 0.2, we see accuracy improve to 80%, indicating a higher overall rate of correct predictions. Recall sees a notable increase to 69%, indicating that the model is now more adept at identifying customers who are likely to churn. However, we sacrifice precision for the gains we made earlier. At 60%, precision has decreased compared to the baseline model. This reduction is anticipated as lowering the threshold generally increases the number of positive predictions, which can lead to more false positives.

It is important to note here that we are not improving the model in the sense that the model is now ‘smarter’ or more capable. Simply moving the threshold always improves one metric

(precision or recall) at the expense of the other. The fact that the area under the curve stays the same at 0.86 is further evidence of this fact.

| Model (Threshold) | Accuracy | Precision | Recall | AUC |
|---|---|---|---|---|
| Baseline (0.5) | 74% | 86% | 60% | **0.86** |
| Optimal Threshold (0.2) | 80% | 60% | 69% | **0.86** |

Table 2: Evaluation Metrics for the Baseline Model at Various Thresholds

### 6.1.3 Addressing Class Imbalance

Addressing class imbalance is a critical step in enhancing the performance of machine learning models, especially in tasks where the outcome classes are unevenly distributed, as is the case with our customer churn prediction model. To mitigate this, we explore two widely recognized techniques: undersampling and oversampling. Undersampling involves randomly removing instances from the majority class to match the number of instances in the minority class. This approach can improve model sensitivity to the minority class but at the cost of losing potentially valuable data from the majority class. We got higher performance for precision and recall when using the undersampled dataset, and incurred a slight hit on accuracy as compared to the baseline model. Specifically, accuracy was at 77%, precision at 75%, and recall at 81%. This suggests that undersampling effectively makes the model more sensitive to the minority class.

Oversampling, in contrast, involves creating additional instances of the minority class, often through duplication or more sophisticated methods like SMOTE (Synthetic Minority Over-sampling Technique). This approach retains all majority class data while enhancing the representation of the minority class. Results from the model built with the oversamples dataset are similar, albeit slightly better. Accuracy of 78% indicates a slight but meaningful improvement over both the baseline and the undersampled model. This suggests that enriching the dataset with more minority class instances provides the model with a more comprehensive learning base. Precision and recall also show slight improvement over the undersamples model.

| **Model (Threshold)** | **Accuracy** | **Precision** | **Recall** | **AUC** |
|---|---|---|---|---|
| Baseline (0.5) | 74% | 86% | 60% | 0.86 |
| Optimal Threshold (0.2) | 80% | 60% | 69% | 0.86 |
| Balanced: Undersampling | 77% | 75% | 81% | 0.85 |
| Balanced: Oversampling | 78% | 77% | 83% | 0.87 |

Table 3: Model Performance using Balanced Classes

The improvement in model performance through both undersampling and oversampling techniques underscores the detrimental impact of class imbalance on predictive modeling. By balancing the class distribution, both methods significantly enhance the model's sensitivity to churn, as evidenced by the notable improvements in recall. This is particularly important in a business context, where identifying at-risk customers is crucial for implementing effective retention strategies.

Moreover, the similar performance of both models suggests that the choice between

undersampling and oversampling might depend on specific business contexts and constraints, such as the availability of computational resources and the importance of preserving data integrity. Oversampling, with its ability to maintain all original data and augment the minority class, might be preferred in scenarios where data loss is a concern. Conversely, undersampling could be advantageous in situations requiring faster model training times or when the dataset is large enough to afford some level of data reduction without losing critical information.

### 6.1.4 Before Leaving the Linear World

In the process of developing predictive models for customer churn, three logistic regression models were initially constructed: a baseline model, an undersampled model, and an oversampled model, utilizing more than 20 available features. The primary motivation for further refining these logistic regression models, before considering more complex algorithms such as random forests, is twofold: to maintain the models' interpretability and to enhance their predictive accuracy. To address these objectives, we build two additional logistic regression models, each employing a different feature selection technique. The first model uses Forward Stepwise Selection, a systematic method that incrementally selects features based on their contribution to predictive performance. The second model applies L1 Regularization (Lasso), which introduces a penalty proportional to the absolute value of the regression coefficients, encouraging a reduction in feature count by driving some coefficients to zero. These approaches are

aimed at reducing the complexity of the models by selecting a subset of features, thereby improving interpretability and potentially increasing predictive accuracy.

### 6.1.5 Forward Stepwise Selection

The Forward Stepwise Selection process has identified ten features as the most significant in predicting customer churn for the telecommunications company, with the resulting model achieving an accuracy of 79%, precision of 78%, recall of 83%, an F1-score of 0.8, and an AUC of 0.87. These metrics, particularly the F1-score, were chosen over accuracy alone to ensure a balanced consideration of both precision (the model's ability to correctly identify actual churns) and recall (the model's ability to capture as many actual churns as possible). The chosen metrics underscore the model's effectiveness in balancing these aspects, with an F1-score of 0.8 indicating a robust predictive capability.

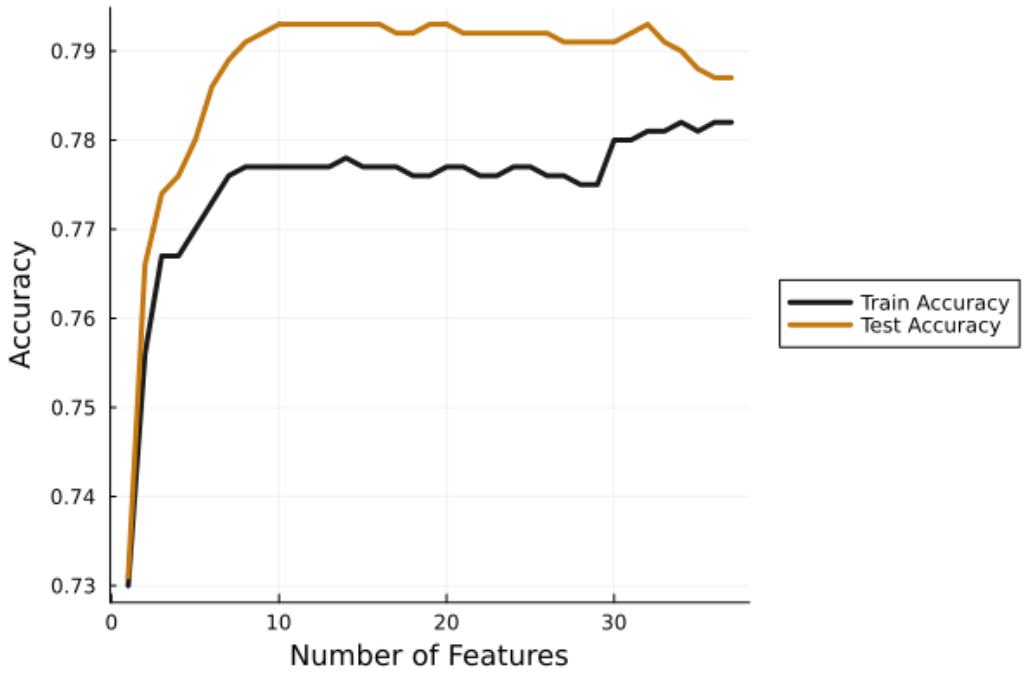


Figure 21: Accuracy by Number of Features with Forward Stepwise

The first 10 selected features offer insightful implications in the context of business operations and strategy. Below the features are given in order of importance:

**`Contract__Month-to-month**:` This suggests that customers on month-to-month contracts are more likely to churn, possibly due to the low commitment required to switch providers. This aligns with business understanding that longer-term contracts typically reflect higher customer retention.

**`Dependents__No**:` Customers without dependents might have fewer constraints or considerations when deciding to change services, making them more prone to churn. This could imply that customers with dependents prioritize stability in service offerings.

**`Internet_service__No**:` Customers without internet service may represent a segment less engaged with the company's value-added services, which could influence their loyalty and satisfaction levels.

**`Tenure_months**:` Shorter tenure may indicate a higher churn risk, underscoring the importance of early engagement and retention strategies for new customers.

**`Internet_service__DSL**:` This may reflect customer dissatisfaction or competitive disadvantages with DSL technology compared to other internet service types, such as fiber-optic.

**`Payment_method__Electronic check**:` This payment method might be associated with higher churn, possibly due to demographics that prefer this payment method or the lack of automatic payments, which can increase the churn risk.

**`Phone_service__No**:` Similar to customers without internet service, those without phone service may also be less engaged and thus more likely to churn.

**`Paperless_billing__No**:` Customers not opting for paperless billing might be less digitally engaged or satisfied with the service, influencing their likelihood to churn.

**`CLTV (Customer Lifetime Value)**:` Surprisingly, this feature's significance suggests that even customers with potentially high value are at risk of churning, highlighting the need for targeted retention strategies across all segments.

**`Senior_citizen__Yes**:` Senior citizens may have specific service needs or preferences that are not being met, leading to higher churn rates in this demographic.

### 6.1.6 Subset Selection with L1 Regularization (Lasso)

L1 Lasso Regularization works by adding a penalty equal to the absolute value of the magnitude

of coefficients to the loss function. This method encourages sparsity in the model, effectively shrinking some coefficients to zero, hence performing feature selection by retaining only the most significant features in the model.

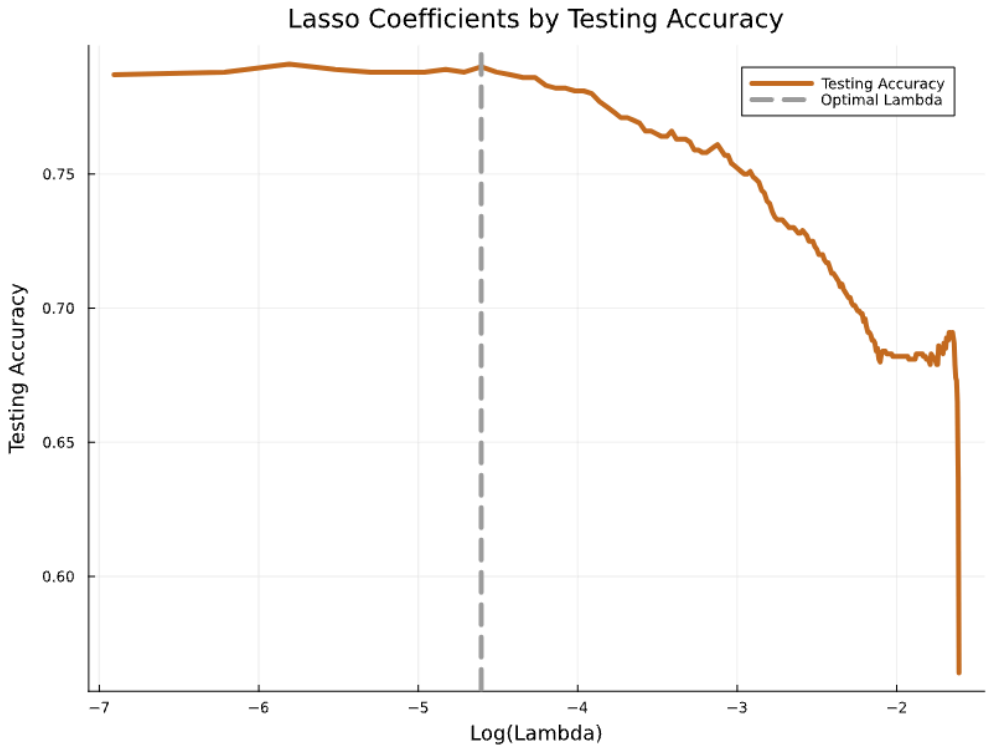


Figure 22: Testing Accuracy for Different Values of Lambda

For a nuanced analysis of the impact of L1 regularization on model performance, we developed custom code to systematically experiment with various values of the regularization parameter, lambda (λ). This approach allowed us to capture detailed information on how each coefficient is influenced by changes in λ, facilitating a more comprehensive understanding than what is typically achievable with the predefined Grid objects found in MLJ packages for Julia. Through this exploration, we determined that the optimal value of lambda, which maximizes the test

accuracy, is $\lambda = 0.01$.

We ran the model for each lambda value in the range and recorded evaluation metrics for each lambda. Below we map the Testing Accuracy by log(lambda) for visual clarity. Notice how accuracy plunges when lambda gets closer to 0.2, which from our experiments sets most coefficients to zero. This means that the model is essentially the intercept. We also plot the lambda that gave the optimal testing accuracy in the gray dashed line. Employing this optimal $\lambda$ value in the Lasso model yielded the following performance metrics: an accuracy of 79%, a precision of 78%, and a recall of 82%. These results indicate a high level of predictive accuracy and balance between precision and recall, demonstrating the effectiveness of L1 Lasso Regularization in enhancing the model's predictive capabilities while maintaining a focus on the most impactful features.

| **Positive Coefficients** | **Value** | **Negative Coefficients** | **Value** |
|---|---|---|---|
| contract__Month to Month | 0.82 | dependents__Yes | -0.92 |
| payment_method__Electronic | 0.46 | tenure_months | -0.74 |
| tech_support__No | 0.41 | multiple_lines__No | -0.12 |
| monthly_charges | 0.36 | | |
| internet_service__Fiber Optic | 0.26 | | |
| paperless_billing__Yes | 0.22 | | |
| dependents__No | 0.10 | | |

Table 4: Left, coefficients that increase churn probability. Right, coefficients that decrease churn probability. Only features with significant coefficients are shown.

The model built with the optimal lambda value of 0.01 yields insights into the factors influencing customer churn probability through the magnitude and direction of its coefficients. These coefficients can be interpreted as the degree of influence each feature has on the likelihood of a customer deciding to leave the company, with positive coefficients indicating an increase in churn risk and negative coefficients indicating a decrease.

Below we show how the coefficients evolve as we increase the value of lambda. Note how more coefficients collapse to zero as we increase lambda. In a gray, dashed vertical line we again plot the optimal value for lambda at 0.01.

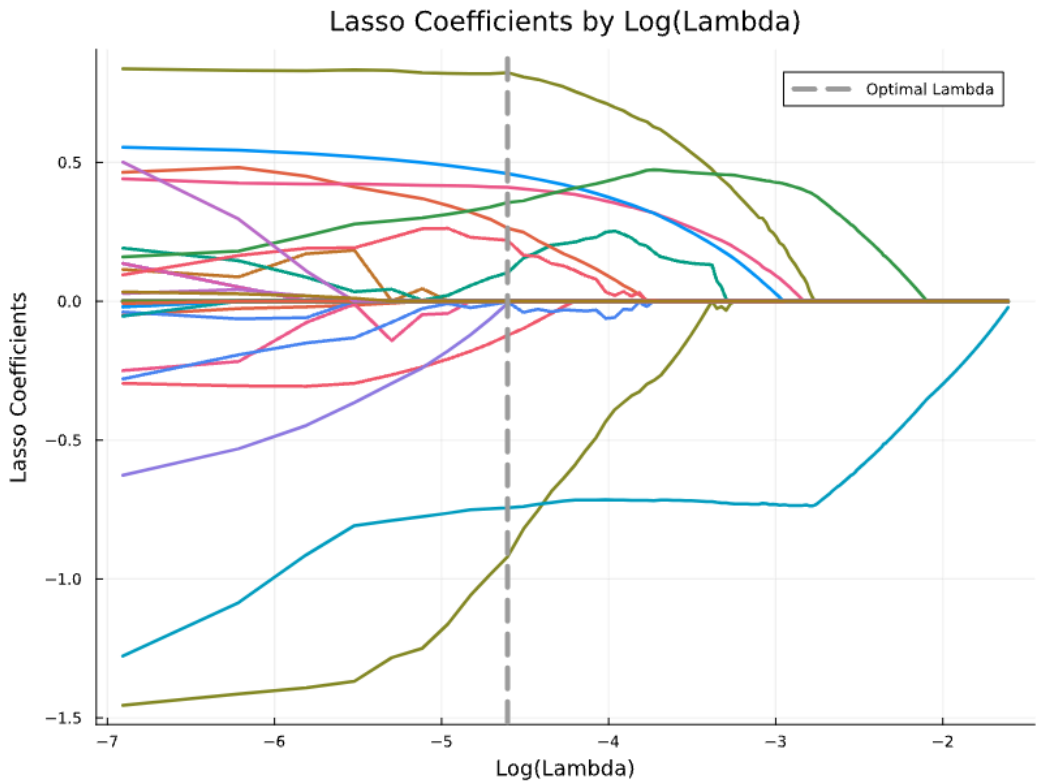


Figure 23: Coefficient Values by Log(Lambda)

#### 6.1.6.1 Negative Coefficients Analysis (Reducing Churn Probability)

Having dependents significantly decreases the likelihood of churn. This may reflect a higher need for stability or a more complicated process for changing services for customers with family responsibilities.

Also**,** longer tenure correlates with reduced churn risk, possibly due to increased customer loyalty or satisfaction over time. Customers without multiple lines are slightly less likely to churn, which might indicate that those not utilizing extensive services find their current plans satisfactory.

#### 6.1.6.2 Positive Coefficients Analysis (Increasing Churn Probability)

Customers with month-to-month contracts are much more likely to churn, likely due to the low barriers to changing providers compared to those in longer-term contracts. Preference for electronic payment methods is associated with a higher churn probability, which could reflect a demographic that is more comfortable with technology and, therefore, more likely to explore other options. Lack of tech support significantly increases churn risk, underscoring the importance of customer service in retaining clients.

Higher monthly charges are linked to increased churn, suggesting price sensitivity among customers.

Customers with fiber optic service are more likely to churn, possibly due to higher expectations or issues related to the service quality or price. Opting for paperless billing slightly increases

churn risk, which might indicate a segment of customers that are more digitally engaged and thus have higher service expectations. Interestingly, the absence of dependents, while having a smaller effect size compared to having dependents, also contributes positively to churn risk, suggesting that customers without family obligations may find it easier to switch providers.

These coefficients highlight the nuanced dynamics of customer behavior and preferences. Features associated with contract flexibility, payment methods, service offerings, and customer support play significant roles in influencing churn. Identifying these factors allows for targeted strategies to mitigate churn risk, such as improving service quality, offering competitive pricing, enhancing customer support, and tailoring offerings to meet the needs of different customer segments.

### 6.2 Moving Beyond the Linear World

We spent considerable time trying to select the best linear model. We found a good subset of features (we say good because stepwise subset selection does not guarantee optimal model selection) and manually engineered a few features from theory. Our efforts were not in vain, as we did achieve a slightly better predictive logistic regression model while still maintaining interpretability through coefficients. However, it is important to note that manual feature engineering is limited for obvious reasons. With thousands of observations there is little chance we humans can capture the true nuances in the way features interact, and there is a high chance that the linear model is a gross simplification of the true model for this dataset. Therefore, it is worth exploring other non-linear models in pursuit of gaining more predictive powers. This does

come at a cost, however,  in two major ways. One, we compromise interpretability at best, and lose it at worst. Two, as our models become more flexible, we run into the variance-bias tradeoff, which is unavoidable in statistics. Our linear models are highly biased, as they assume the true model is linear, which is rarely the case in the real world (although we could still make relatively accurate predictions.) On the other hand, the linear model suffers less from variance than the more flexible models, which tend to have high variance because of their inherent flexibility.

### 6.3 K-Nearest Neighbors Classifier

The K-Nearest Neighbors (K-NN) classifier is a non-parametric, instance-based learning algorithm used for classification and regression. In classification tasks, it operates by identifying the 'k' instances in the training dataset closest to the input feature vector, and the output class is determined by a majority vote among the 'k' nearest neighbors. This method is intuitive and straightforward, as it assumes similar instances are likely to be found in close proximity in the feature space.

#### 6.3.1 The Baseline Model

For implementing the K-NN classifier in this analysis, we utilized the NearestNeighborModels package within the Julia programming environment, incorporating it into an MLJ pipeline equipped with cross-validation. This approach ensures a robust evaluation of the model's

performance across different subsets of the data, enhancing the reliability of the results. The baseline K-NN model, configured with k=5 and applied to a balanced dataset, yielded improved performance metrics for accuracy, precision, and recall at 80%, 75%, and 92% respectively.

Compared to the previously discussed logistic regression models, the K-NN classifier exhibits a higher recall rate (92% vs. up to 83% in logistic regression), suggesting a superior ability to identify true positives, or actual churn cases. The accuracy is also slightly higher or comparable (80% for K-NN vs. up to 79% in logistic regression), indicating a strong overall performance. However, precision is slightly lower for the K-NN model compared to the logistic regression model (75% vs. up to 78%), which suggests that while the K-NN model is better at capturing actual cases of churn, it does so at the expense of misclassifying some non-churn cases as churn.

One possible explanation for the K-NN model's performance, particularly its high recall, is related to the use of SMOTE (Synthetic Minority Over-sampling Technique) or similar techniques for balancing the dataset. Since SMOTE generates synthetic samples for the minority class by considering the nearest neighbors of those samples, this preprocessing might inherently align well with the K-NN algorithm's approach of classifying based on the proximity of neighbors. The balanced dataset, enhanced by techniques leveraging nearest neighbor calculations, could therefore be particularly well-suited for K-NN, enabling it to perform effectively in identifying cases of churn.

This synergy between the dataset preprocessing method and the K-NN classifier's underlying principle highlights the importance of matching the characteristics of the data preprocessing techniques with the assumptions and strengths of the predictive model being used.

### 6.3.2 Fine-Tuned with Cross-Validation Model

The fine-tuning process of the K-Nearest Neighbors (K-NN) classifier, through cross-validation, identified the optimal number of neighbors (k) as 1. Adjusting the model to this optimal parameter setting led to a notable improvement in performance metrics compared to the baseline K-NN model. The metrics for the improved model are as follows: Accuracy of 85%, Precision of 80%, and Recall of 94%.

This adjustment in metrics, when compared to both the baseline K-NN model and the previously reported results of the fine-tuned model, offers a new perspective on the model's performance. The updated accuracy of 85% represents a substantial improvement over the baseline model's 80% accuracy. This indicates a higher overall rate of correctly predicted outcomes (both churn and no churn) when using the fine-tuned model. The model's ability to generalize and accurately classify new data points has notably been enhanced. The precision remains at 80%, as previously noted with the incorrect metrics. This consistency in precision, despite changes in other metrics, suggests that the model's ability to predict churn accurately without increasing false positives remains strong and unaffected by the adjustments in k. The most notable change is in recall, which has increased from 92% in the baseline model to 94% in the fine-tuned version. This improvement indicates that the model is now even more effective at identifying actual churn cases, missing only a very small fraction of these cases.

The fine-tuning process, specifically the selection of k=1, implies that the model's predictions are highly sensitive to the nearest data points. The success of such a model configuration suggests that the dataset likely features well-distinguished classes, allowing individual instances to be

strong predictors of their respective class labels. The increase in recall, in particular, is advantageous for a churn prediction model, as identifying the majority of churn cases (true positives) is typically more valuable for a business than mistakenly identifying a few non-churn cases as churn (false positives). This is because interventions designed to retain customers can then be more accurately targeted, potentially leading to more effective retention strategies and reduced unnecessary expenditure on customers who were not at risk of churning.

### 6.3.3 KNN Model Results

Below is the comparison between the logistic regression baseline model and the two KNN models:

| Model | Accuracy | Precision | Recall |
|---|---|---|---|
| Logistic Regression: Baseline | 74% | 86% | 60% |
| KNN: Baseline (k=5) | 80% | 75% | 92% |
| KNN: Tuned (k=1) | 85% | 80% | 94% |

Table 5: Model Performance for K-Nearest Neighbors Models

## 6.4 Tree-Based Models

Tree-based models are particularly adept at handling datasets with a large number of categorical features because they naturally segment data into branches based on feature

values. This segmentation approach mirrors decision-making processes, making tree-based models intuitive and effective for classification tasks where categorical variables play a significant role.

In our analysis of the IBM Telco dataset, which includes numerous categorical features related to customer demographics and service usage patterns, we decided to employ two prominent tree-based models: Random Forest and XGBoost. These models were chosen for their complementary strengths in dealing with complex, real-world datasets.

### 6.4.1 Random Forest Classification

Random Forest is particularly beneficial for its robustness against overfitting, thanks to the averaging of multiple trees. This makes it a reliable choice when dealing with a dataset that has the potential for high variance or when the relationship between features and the target variable is complex.

#### 6.4.1.1 The Baseline Model

In our analysis of the IBM Telco dataset, we constructed a baseline model using the `RandomForestClassifier` from the MLJ Decision Tree Interface package, a comprehensive framework for machine learning in Julia. We utilize all available features in our dataset and use a prediction threshold of 50%. Upon evaluation, the baseline

RandomForestClassifier achieved an accuracy of 80%. Accuracy alone can be misleading in such contexts, as it doesn't differentiate between the types of errors made by the model. The precision and recall metrics, standing at 66% and 52% respectively. Precision at 66% indicates that when the model predicts churn, it is correct about two-thirds of the time. However, recall at 52% reveals that the model is only able to identify slightly more than half of the actual churn cases. These figures are considerably lower than those of other models, suggesting that while the RandomForestClassifier is generally reliable in classifying customers, it struggles to consistently identify the subtler patterns associated with churn.

#### 6.4.1.2 Balanced Class Model

Transitioning to an improved model involved addressing one of the critical challenges highlighted in our baseline model evaluation: the class imbalance inherent in the IBM Telco dataset. Using SMOTE to balance the classes before training the model, without any hyperparameter tuning at this stage, allowed us to focus on the impact of class balance on model performance. The model was then re-evaluated using the same metrics as before: accuracy, precision, and recall. Accuracy improved to 87%, indicating a higher overall rate of correct predictions across both classes. This improvement suggests that balancing the classes enabled the model to make more accurate predictions, reducing the bias towards the majority class that was observed in the baseline model. Precision rose to 86%. This significant improvement from the baseline model indicates that with a balanced

dataset, the model became much more reliable in predicting churn. When the model predicts that a customer will churn, it is correct approximately 86% of the time, which is crucial for targeting interventions effectively without unnecessarily expending resources on customers who are unlikely to churn. Recall saw an even more impressive improvement, reaching 89%.

This step in our modeling process demonstrates a significant leap forward in our ability to predict customer churn accurately. The insights gained here will guide further refinements, including hyperparameter tuning and exploring additional modeling techniques, to develop an even more effective predictive model for the Telco company's customer churn challenge.

#### 6.4.1.3 Hypertuned Model

Hyperparameter tuning is a critical step in refining machine learning models to enhance their performance by systematically searching for the optimal set of parameters that governs the training process. For our Random Forest classifier, applied to the IBM Telco dataset, we engaged in hyperparameter tuning using a grid search approach within the MLJ framework. The parameters we chose to tune were number of trees, maximum depth of a tree, and minimum samples split. Number of Trees specifies the number of decision trees that should be built and combined within the Random Forest. Increasing the number of trees generally improves the model's stability and accuracy, as it averages more decisions from more diverse perspectives. However, beyond a certain point, gains in

performance can diminish, and computational costs can increase significantly. Our tuning ranged from 10 to 100 trees. The maximum depth of each tree in the forest. A deeper tree can model more complex patterns by creating more specific (conditional) branches. However, too much depth can lead to overfitting, where the model captures noise in the training data instead of underlying patterns, reducing its generalization to new data. We considered depths ranging from 1 to 30. Minimum sample split determines the minimum number of samples required to split an internal node. Higher values prevent the model from learning overly specific patterns, which could be noise in the training data, thus mitigating overfitting. We explored values from 2 to 20 for this parameter.

After applying grid search cross-validation with these parameters, we found the optimal settings for our Random Forest model to be 97 for the number of trees and 28 for maximum tree depth. We stop our grid search because the optimal parameters are now below the maximum values in the ranges we tried.

With these optimized parameters, we constructed a Random Forest model and evaluated its performance. The results showed an accuracy of 87%, which matches the performance of the model trained on the balanced dataset using SMOTE. The precision slightly decreased to 85%, indicating a small trade-off where the model, despite being highly accurate, has a marginally lower proportion of true positive predictions out of all positive predictions. Conversely, recall improved to 90%, which is an important metric in the context of churn prediction. The Area Under the Curve (AUC) of 0.94 remained unchanged, indicating that the overall ability of the model to distinguish between the churned and retained customers remained excellent throughout the tuning process. The

AUC is a performance measurement for the classification problems at various threshold settings, and a score of 0.94 signifies a high level of model performance.

### 6.4.2 XGBoost Classifier

#### 6.4.2.1 The Baseline Model

Moving on to the XGBoost classifier represents a strategic shift towards leveraging gradient boosting, a powerful approach known for its predictive accuracy and efficiency on structured datasets like the one we're analyzing from the IBM Telco dataset. Like with the Random Forest model, we first addressed the issue of class imbalance by employing SMOTE (Synthetic Minority Over-sampling Technique) to balance the classes. This preparation ensures that our model training is not biased towards the majority class and can learn the minority class's characteristics more effectively.

With the class-balanced dataset, the baseline XGBoost model yielded an **accuracy** of 86%, precision of 86%, and recall of 87%.

These results demonstrate the model's strong overall performance, with a high accuracy indicating that a significant majority of predictions match the actual values. The precision and recall metrics are particularly noteworthy because they suggest that the model not only makes accurate predictions about customer churn but also does so with a high degree of reliability (precision) and sensitivity (recall). This balance is crucial for practical

applications, ensuring that interventions to prevent churn are both appropriately targeted and comprehensive.

#### 6.4.2.2 Important Feature Analysis

The analysis of feature importance further enriches our understanding of the factors driving customer churn. The features identified as most influential in predicting churn are given below:

| Important Feature | XBoost Score |
|---|---|
| contract__Month to Month | 117 |
| internet_service__Fiber Optic | 22 |
| dependents__No | 22 |
| payment_method__Electronic | 14 |
| tech_support__No | 9.7 |
| contract__Two Year | 8.7 |
| paperless_billing__No | 7.6 |
| contract_One Year | 6.2 |
| multiple_lines | 5.3 |
| payment_method__Credit Card | 5 |

Table 6: Important Features as Identified by the XGBoost Model

Customers with month-to-month contracts are more likely to churn, likely due to the lower barriers to changing services compared to longer-term contracts. The presence of

fiber optic service is a significant predictor, possibly indicating higher expectations or different experiences that influence churn. Customers without dependents may have more flexibility or different needs, affecting their churn risk. The choice of payment method might reflect customer preferences or satisfaction levels that correlate with churn. Lack of tech support could lead to dissatisfaction, increasing churn risk. Longer-term contracts might have a protective effect against churn, though less so than month-to-month contracts increase it. Preferences for billing methods may also be indicative of customer engagement or satisfaction. Similarly to two-year contracts, one-year contracts suggest a commitment that may reduce churn risk. Customers with a simpler service setup may have different churn drivers compared to those with multiple lines. As with electronic checks, the choice of payment method may reflect broader preferences or satisfaction levels.

These features provide actionable insights into customer behavior and preferences, suggesting areas where interventions could be targeted to reduce churn. For example, improving support services, offering incentives for longer-term contracts, or addressing the specific needs of fiber optic service users could be effective strategies.

The performance of this baseline XGBoost model, combined with the insights from feature importance, lays a solid foundation for further optimization, including hyperparameter tuning.

#### 6.4.2.3 Model with Hyperparameter Tuning

We use the XGBoost Classifier from the XGboost package in Julia, and feed it into an MLJ tuning pipeline with 5-folds cross-validation. Focusing on tuning two crucial hyperparameters. One, the learning rate (ETA.) This parameter controls the step size at each iteration while moving toward a minimum of the loss function. A smaller ETA value means that the model will take smaller steps during gradient descent, potentially leading to a more precise convergence at the cost of requiring more iterations. Conversely, a larger ETA value could lead to faster convergence but at the risk of overshooting the minimum. The range explored was from 0.01 to 0.3. Two, maximum depth, a parameter that specifies the maximum depth of a tree, affecting the model's complexity. A deeper tree can model more complex patterns, but it also risks overfitting to the training data. Conversely, a shallower tree might not capture the intricacies of the data, leading to underfitting. The depth range considered was from 3 to 10.

After performing hyperparameter tuning, the optimal parameter value for ETA and maximum depth were 0.247, and 8, respectively. Using these optimized parameters to fit the XGBoost model, the performance metrics remained unchanged from the baseline model across all metrics measured: accuracy, precision, and recall. The unchanged metrics post-tuning indicate a few possible scenarios in the context of this analysis. One, the baseline model was already performing optimally within the constraints of the dataset and the model's capability. This suggests that the initial configuration was quite effective, and the data does not allow for much improvement beyond what was already achieved. Two, the selected parameters for tuning, while important, may not be the most sensitive to

changes in this particular dataset or model configuration. It's possible that other parameters, not included in the tuning process, could have a more significant impact on the model's performance. Three, the dataset might have a level of complexity or noise that prevents further improvement from hyperparameter tuning alone. Essentially, the model may have reached a point where additional complexity (through deeper trees or adjusted learning rates) does not translate into better generalization due to the inherent limitations of the available information.

### 6.5 Model Evaluation

We have evaluated a diverse array of strategies across four primary types of models: Generalized Linear Models (GLM), Random Forest, K-Nearest Neighbors (KNN), and XGBoost. Each model was subjected to various adjustments and optimizations, including baseline configurations, oversampling, undersampling, optimal threshold adjustments, regularization, subset selection, and hyperparameter tuning. Our objective was to enhance the models' performance metrics—accuracy, precision, recall, F1 score, and the area under the receiver operating characteristic curve (AUC)—to identify the most effective model for predicting customer churn.

| Model | Accuracy | Precision | Recall | F1 Score | AUC |
|---|---|---|---|---|---|
| GLM: Baseline | 0.817 | 0.67 | 0.597 | 0.632 | 0.859 |
| GLM: Optimal Threshold | 0.8 | 0.602 | 0.699 | 0.647 | 0.859 |
| GLM: Undersampling | 0.767 | 0.752 | 0.809 | 0.78 | 0.847 |
| GLM: Oversampling | 0.787 | 0.773 | 0.827 | 0.799 | 0.87 |
| GLM: L1 Regularization | 0.79 | 0.78 | 0.822 | 0.8 | 0.867 |
| GLM: Forward Stepwise | 0.793 | 0.783 | 0.825 | 0.804 | 0.868 |
| K-Nearest Neighbors: Oversampling | 0.805 | 0.752 | 0.926 | 0.83 | 0.877 |
| K-Nearest Neighbors: Tuned | 0.848 | 0.801 | 0.937 | 0.864 | 0.845 |
| Random Forest: Baseline | 0.804 | 0.661 | 0.523 | 0.584 | 0.853 |
| Random Forest: Oversampling | 0.865 | 0.856 | 0.885 | 0.87 | 0.942 |
| Random Forest: Tuned | 0.865 | 0.847 | 0.899 | 0.873 | 0.94 |
| XGBoost: Oversampling | 0.858 | 0.855 | 0.871 | 0.863 | 0.938 |
| XGBoost: Hyperparameter Tuning | 0.859 | 0.856 | 0.871 | 0.864 | 0.943 |

Table 7: Model Evaluation Metrics

Accuracy, a measure of the overall correctness of the model across both classes, varied notably across the models and techniques used. The Random Forest model with hyperparameter tuning and oversampling exhibited the highest accuracy (0.865), indicating its superior ability to correctly classify both churned and retained customers. This performance suggests that Random Forest, especially when fine-tuned and with adjusted sample distribution, is highly effective for this prediction task. On the other hand, models like the GLM with undersampling showed lower accuracy, suggesting that while they may excel in other metrics, they might misclassify more often than the more

sophisticated Random Forest and XGBoost models.

Precision, which quantifies the accuracy of the model in predicting positive (churned) cases, was led by the XGBoost and Random Forest models after oversampling and hyperparameter adjustments, with precision scores peaking at 0.856. This indicates a strong capability of these models to minimize false positives, a crucial aspect in customer churn prediction to ensure that interventions are targeted appropriately. The high precision of these models demonstrates their efficiency in identifying customers at high risk of churn without unnecessarily targeting those likely to stay, thus optimizing resource allocation for retention efforts.

Recall, measuring the model's ability to capture actual positive cases, was outstanding in the K-Nearest Neighbors model following hyperparameter tuning, reaching up to 0.937. This suggests that, despite other models showing balanced performance across metrics, KNN, particularly when finely tuned, excels at identifying nearly all customers who are likely to churn. High recall is essential in a business context to ensure minimal missed opportunities for intervention with at-risk customers. However, the balance between recall and precision, especially in the context of limited resources for intervention strategies, underscores the importance of choosing a model that aligns with business priorities and the cost implications of false positives versus false negatives.

### 6.6 Feature Analysis, Revisited

The comprehensive feature analysis across Logistic Regression with Forward Stepwise Selection and L1 Regularization (Lasso), alongside XGBoost, presents a fascinating picture of the determinants influencing customer churn in the telecommunications sector.

Common across all models is the prominence of the `Contract__Month-to-month` feature, which suggests that customers on month-to-month contracts exhibit a higher propensity to churn. This consensus highlights the volatility associated with short-term commitments. This highlights a universal business insight: the stability offered by longer-term contracts significantly bolsters customer retention. Another broadly acknowledged feature is the payment method, with a particular focus on `Payment_method__Electronic` check, flagged by both Logistic Regression models and noted by XGBoost. This indicates a demographic segment that, while potentially tech-savvy, may be more fickle or attracted to competing offerings that leverage more modern or user-friendly payment solutions.

Notably, the analysis diverges in the emphasis on service types and customer demographics. For instance, `Internet_service__No` and `Phone_service__No` are critical in the Logistic Regression models, pointing towards a segment less engaged due to the absence of these services. However, XGBoost places substantial weight on `Internet_service__Fiber Optic`, revealing a nuanced perspective: while fiber optic service may attract customers with its high-speed offering, it might also come with higher expectations and competition,

leading to churn if these are not met. Similarly, demographic features like `Senior_citizen__Yes` and `Dependents__No` underscore varied vulnerabilities across customer segments, with Logistic Regression models highlighting these aspects more than XGBoost. This variance could be attributed to the models' differing capacities to capture complex, non-linear relationships and interactions among features.

The holistic view from this feature analysis presents a compelling narrative on customer churn, where contract type, engagement level with services, payment methods, and demographic factors interplay in determining churn risk. The models agree on several key features, underscoring their importance across different analytical techniques.

## 7 Conclusion

In conclusion, our comprehensive analysis not only advances the understanding of customer churn dynamics but also provides a strategic framework for leveraging predictive analytics in support of nuanced and effective customer retention strategies.

The rigorous methodology, encompassing preprocessing, model selection, tuning, and evaluation, ensures the reliability of the findings. Moreover, the use of statistical measures like accuracy, precision, recall, and the area under the receiver operating characteristic curve (AUC) across different models substantiates the robustness of the analysis. Through careful model selection and the application of insights derived from our analysis,

businesses can better navigate the challenges of customer churn, fostering sustained engagement and loyalty.

### 7.1 Trade-offs

Our journey in predictive modeling for customer churn reveals the essential trade-offs between precision and recall, as well as accuracy and interpretability. The Generalized Linear Models (GLM), Random Forest, K-Nearest Neighbors (KNN), and XGBoost models each exhibit strengths and weaknesses that underscore the trade-offs faced in predictive analytics. For instance, while GLM offers interpretability, it may lag in predictive accuracy compared to more complex models like Random Forest and XGBoost. Conversely, models with high predictive power often sacrifice interpretability, complicating the communication of findings to non-technical stakeholders.

### 7.2 Key Inferences

In examining customer churn within the telecommunications sector, our analysis reveals a nuanced interplay of contractual, service-related, and demographic factors that significantly influence customer retention decisions. Central to these findings is the discernible impact of contract type on churn likelihood, with month-to-month contracts identified as a key risk factor for increased churn rates. Equally important is the influence

of service quality and offerings, particularly the presence of fiber optic internet service, which emerges as a significant churn driver. This suggests that customer expectations around service offerings and their satisfaction with these services are critical determinants of churn, highlighting the need for telecommunications companies to continuously evaluate and improve their service quality to meet customer demands. Moreover, demographic characteristics such as being a senior citizen or having dependents also play a consequential role in churn dynamics. These demographic insights indicate varied needs and preferences across different customer segments, further emphasizing the importance of tailored engagement and retention strategies that cater to the specific circumstances and expectations of diverse customer groups. Collectively, these inferences shed light on the multifaceted nature of customer churn, offering a foundation for interventions to mitigate churn risk and enhance customer loyalty.

### 7.3 Implications for Business

The implications of our findings for business strategies are multifaceted. When customer attrition is particularly costly, models with high recall should be prioritized to ensure that as many at-risk customers as possible are identified, even at the expense of some false positives. In contrast, if customer retention strategies are expensive or if interventions are limited, models with high precision become more valuable to ensure that efforts are concentrated on customers most likely to churn. This strategic focus aligns model selection with business objectives, optimizing resource allocation and maximizing the

impact of retention initiatives.

### 7.4 Limitations

Despite the comprehensive nature of our analysis in predicting customer churn, several limitations are noteworthy. The complexity of models like Random Forest and XGBoost, while enhancing predictive accuracy, presents a trade-off in terms of interpretability, potentially complicating the communication of findings to stakeholders who may lack technical expertise. Additionally, efforts to address class imbalance through techniques such as SMOTE cannot entirely eliminate the inherent challenges posed by skewed datasets, which might still influence the performance of predictive models. Moreover, the models deployed in this analysis do not account for temporal dynamics, neglecting the potential impact of changes in customer behavior or market conditions over time. This oversight suggests that the predictive accuracy of the models might wane without regular updates and adaptations to reflect evolving trends, thereby highlighting the importance of continuous model evaluation and adjustment in response to emerging data patterns.

### 7.5 Future Scope

Looking ahead, the evolution of our predictive modeling efforts will embrace emerging data sources, model refinement techniques, and the integration of machine learning

advancements. The exploration of deep learning for pattern recognition in high-dimensional data, and the application of natural language processing to analyze customer feedback, represent promising avenues for enhancing the predictive accuracy and utility of churn models. Additionally, ongoing validation of model performance in the face of changing customer behavior and market conditions will ensure that our approaches remain relevant and impactful.

## 8 References


Alan Turing Institute Contributors. (n.d.). *MLJ: A Julia package for machine learning* (Development version). Retrieved 03-19-2024, from https://alan-turing-institute.github.io/MLJ.jl/dev/

James, G., Witten, D., & Hastie, T. (2023). *An Introduction to Statistical Learning: With Applications in Python* (1st ed.). Springer.

JuliaLang Contributors. (n.d.). *Julia documentation (Version 1.10.0)*. Retrieved 03-19-2024, from https://docs.julialang.org/en/v1.10.0/

JuliaDataScience Contributors. (n.d.). *JuliaDataScience*. Retrieved 03-19-2024, from https://juliadatascience.io

Tan, P.-N., Steinbach, M., & Kumar, V. (2018). *Introduction to Data Mining* (2nd ed.). Pearson.