Predictive Analysis of Customer Churn in the Telecom Sector

Blake Tindol, Yesul Song, Sean Deery

IST-718 Big Data Analytics

Telecom Customer Churn Analysis

Contents

Introduction

In the competitive landscape of the telecommunications industry, retaining customers is equally as critical as acquiring new ones. Customer churn, the phenomenon where customers switch from one service provider to another, poses a significant challenge and incurs considerable costs. This project aims to leverage data science methodologies to analyze customer data, understand the underlying factors of churn, and develop predictive models that forecast potential churn, enabling telecom companies to implement targeted retention strategies.

Telecom companies face the ongoing challenge of customer churn, which affects their revenue and long-term growth. The reasons behind churn can be multifaceted, including poor service quality, better offers from competitors, customer dissatisfaction, or changes in customer needs. Identifying these factors early and accurately predicting which customers are at risk of churning can empower telecom companies to take proactive measures, thereby reducing churn rates and enhancing customer loyalty. The ability to predict churn allows telecom companies to reduce operational costs, improve customer experience, increase revenue, competitive advantage.

The objectives of this analysis are to complete an analysis of the data, create a predictive model, and develop and propose a strategy for customer churn. The data analysis will include exploratory data analysis and feature engineering to identify key predictors of churn. Creating a predictive model will include developing and training machine learning models like logistic regression, random forest, and a support vector machine to predict churn. Based on the insights gained, a strategy can be developed that could include targeted intervention strategies for customer retention.

Addressing customer churn is crucial for the sustainability and growth of telecom companies in a highly competitive market. By applying data science techniques to understand and predict churn, this project aims to provide actionable insights that can significantly impact customer retention strategies. The outcomes of this project will not only benefit the telecom companies by improving their financial health and customer satisfaction levels but also contribute to the broader field of data science by showcasing the applicability and impact of predictive analytics in solving real-world business challenges

Data

Data: https://www.kaggle.com/datasets/shilongzhuang/telecom-customer-churn-by-maven-analytics/data

The dataset acquired from Kaggle contains 7,043 rows and 38 columns. Each row represents a customer from a Telecommunications company in California in Q2 2022. The features contain details about their demographics, location, tenure, subscription services, status for the quarter (joined, stayed, or churned), and more. An additional table of population by zip code is provided. It contains 1,671 rows and 2 columns.

Data Cleaning

The data was cleaned following the steps below:

- *Customer ID* was removed because it does not hold any valuable information.
- The *Zip Code* columns for both datasets were converted from an integer to a string, and the zip codes with leading zeros were cleaned to make sure the zeros were included.
- Population was added to the main data frame by matching the zip code between the two initial datasets.

Missing Values

Missing values were dealt with as follows:

- *Avg Monthly Long Distance Charges*
    - Missing values were set to zero. This column indicates the customer's average long-distance charges, calculated to the end of the quarter specified above (if the customer is not subscribed to home phone service, this will be 0).
- *Multiple Lines*
    - Missing values were set to 'No'. This column indicates if the customer subscribes to multiple telephone lines with the company: Yes, No (if the customer is not subscribed to home phone service, this will be No)
- *Internet Type*
    - Missing values were set to 'None'. This column indicates the customer's type of internet connection: DSL, Fiber Optic, Cable (if the customer is not subscribed to internet service, this will be None)
- *Avg Monthly GB Download*
    - Missing values were set to 0. This column indicates the customer's average download volume in gigabytes, calculated to the end of the quarter specified above (if the customer is not subscribed to internet service, this will be 0)
- *Online Security*
    - Missing values were set to 'No'. This column indicates if the customer subscribes to an additional online security service provided by the company: Yes, No (if the customer is not subscribed to internet service, this will be No)
- *Online Backup*
    - Missing values were set to 'No'. This column indicates if the customer subscribes to an additional online backup service provided by the company: Yes, No (if the customer is not subscribed to internet service, this will be No)
- *Device Protection Plan*
    - Missing values were set to 'No'. This column indicates if the customer subscribes to an additional device protection plan for their Internet equipment provided by the company: Yes, No (if the customer is not subscribed to Internet service, this will be No)
- *Premium Tech Support*

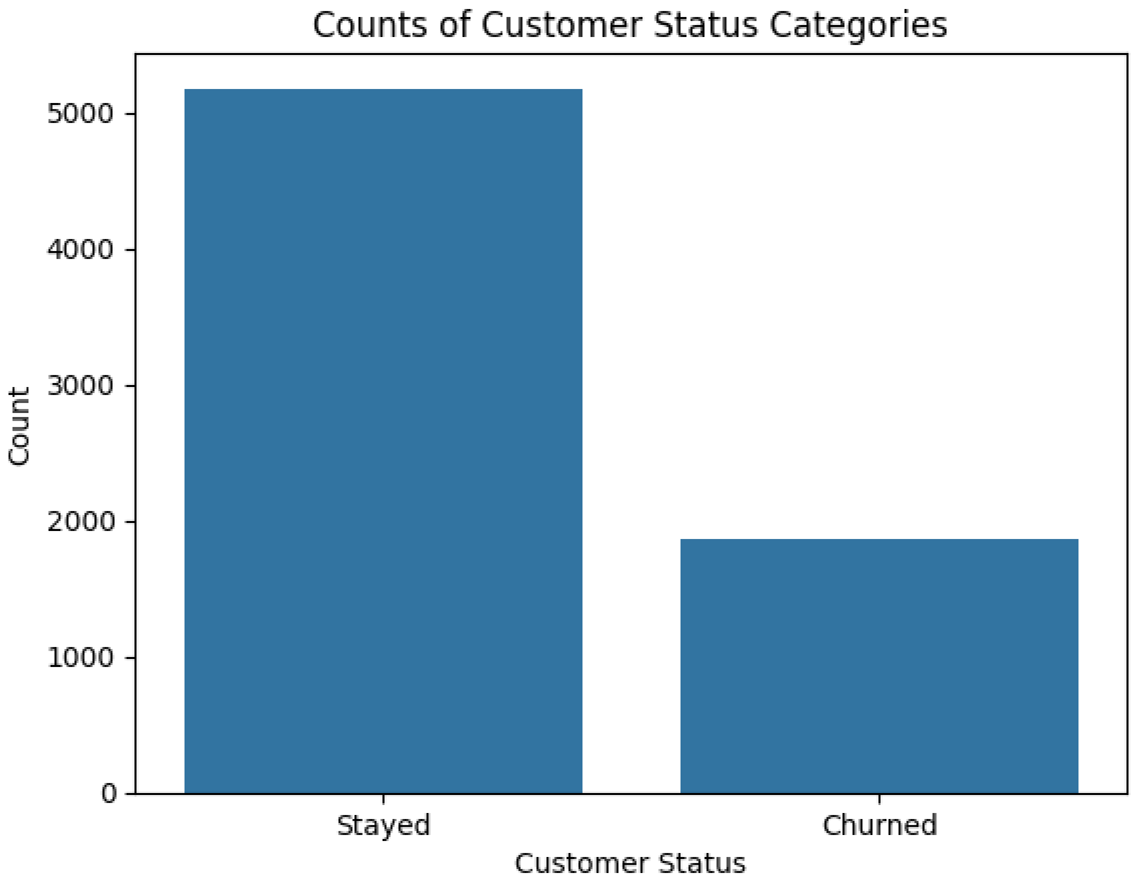
  - Missing values were set to ‘No’. This column indicates if the customer subscribes to an additional technical support plan from the company with reduced wait times: Yes, No (if the customer is not subscribed to internet service, this will be No)
- *Streaming TV*
  - Missing values were set to ‘No’. This column indicates if the customer uses their Internet service to stream television programming from a third party provider at no additional fee: Yes, No (if the customer is not subscribed to Internet service, this will be No)
- *Streaming Movies*
  - Missing values were set to ‘No’. This column indicates if the customer uses their Internet service to stream movies from a third party provider at no additional fee: Yes, No (if the customer is not subscribed to Internet service, this will be No)
- *Streaming Music*
  - Missing values were set to ‘No’. This column indicates if the customer uses their Internet service to stream music from a third party provider at no additional fee: Yes, No (if the customer is not subscribed to Internet service, this will be No).
- *Unlimited Data*
  - Missing values were set to ‘No’. if the customer has paid an additional monthly fee to have unlimited data downloads/uploads: Yes, No (if the customer is not subscribed to internet service, this will be No)

Exploratory Data Analysis

Our comprehensive Exploratory Data Analysis (EDA) dove into the telecom dataset, scrutinizing the relationship between various customer attributes and their churn status. The analysis employed a mix of visual and quantitative methods to uncover underlying patterns that might elucidate churn drivers. Below, we expand upon the key relationships identified between customer attributes and churn, enhanced with insights and potential implications for customer retention strategies.

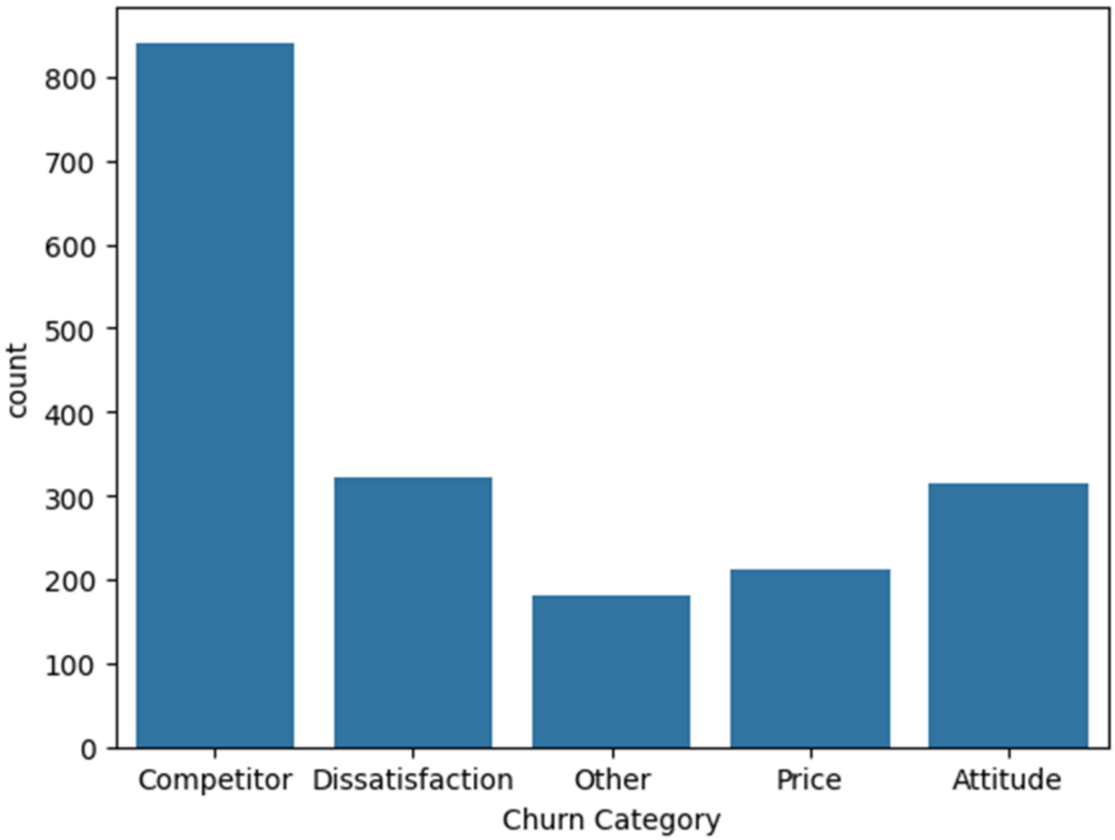
Target Variable: Customer Status

The target variable in the customer dataset is the *Customer Status* column. The values indicate the status of the customer at the end of the quarter: Churned, Stayed, or Joined. Since this analysis is focused on predicting customer churn, Joined and Stayed were combined to represent customers who did not leave. The count plot below shows that about 26.5% of the customers churned in Q2.



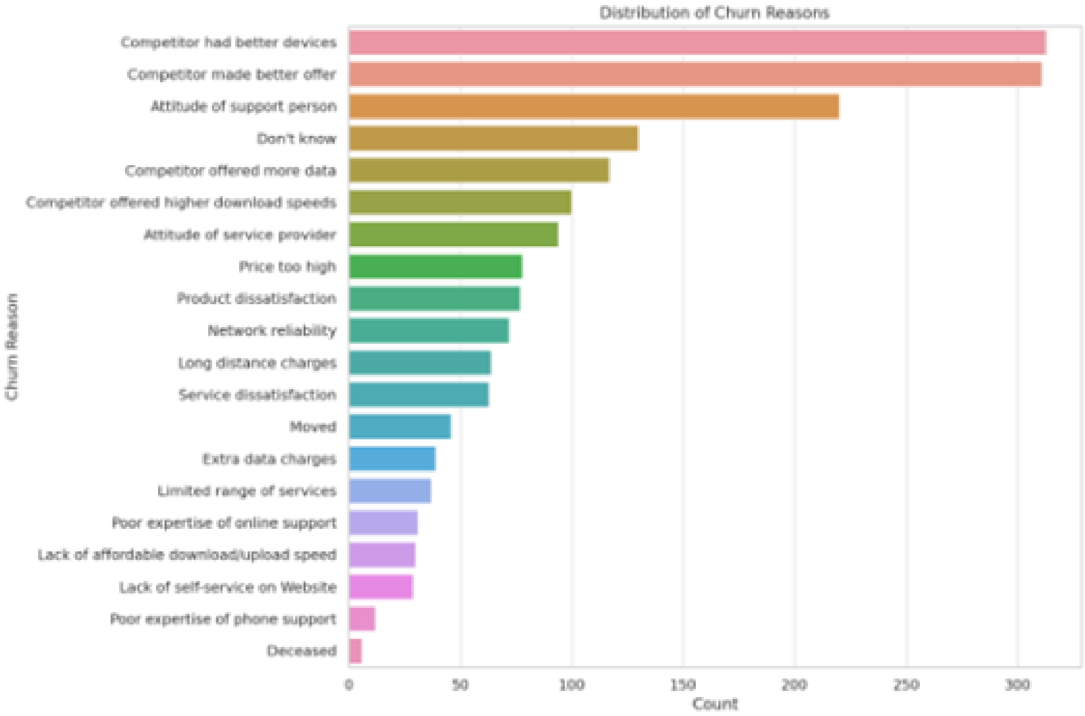
Churn Category

The Churn Category was also provided in the dataset. The count plot below shows that most of the churn is in the Competitor category. It is interesting to note that Price was less frequent than Attitude and Dissatisfaction.



Churn Reason

The Churn Reason column expands on the Churn Category. The count plot below shows that Competitor devices and offers drew much of the churn. The third most frequent churn reason was the attitude of the support person. These are interesting to note because they indicate where the company's interventions might make the most difference. This might include a marketing campaign about the company's offers and how their devices stack up against the competition. Customer-facing departments can also be trained to provide the best possible customer experience while effectively consulting customers about the company's products that fit their needs.



Location Data

Regional Cluster from Latitude and Longitude

The *Latitude* and *Longitude* columns were converted to a single column that represents one of three regional clusters: North California, Central California, and Southern California. The clusters were created by passing Lattitude and Longitude through a K-Means Clustering model. The optimal k value was found using the elbow method and taking into account the nature of the problem and the location. Figure 2 shows the results of the elbow method. We chose to use a k value of 3 because it is where the within-cluster sum of squares levels off, and it splits the state of California up into intuitive regions.

Figure 1

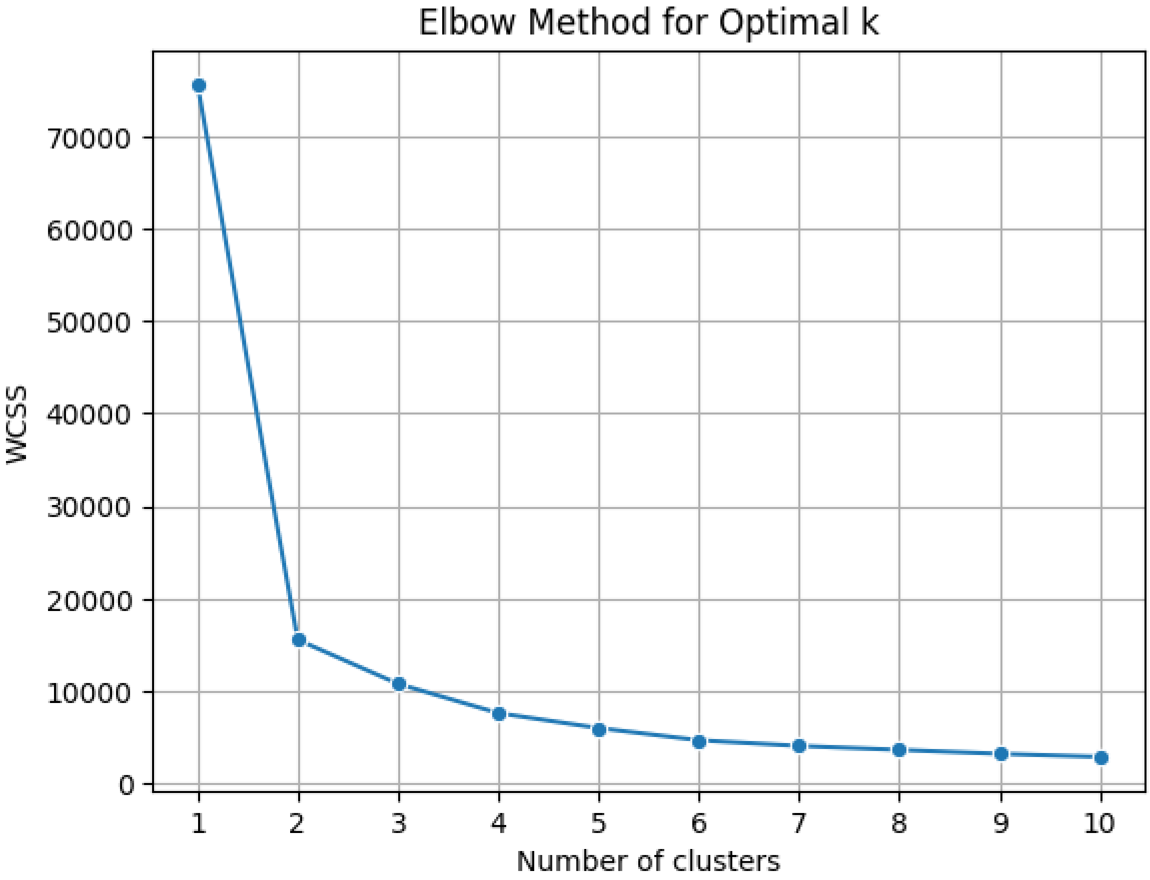
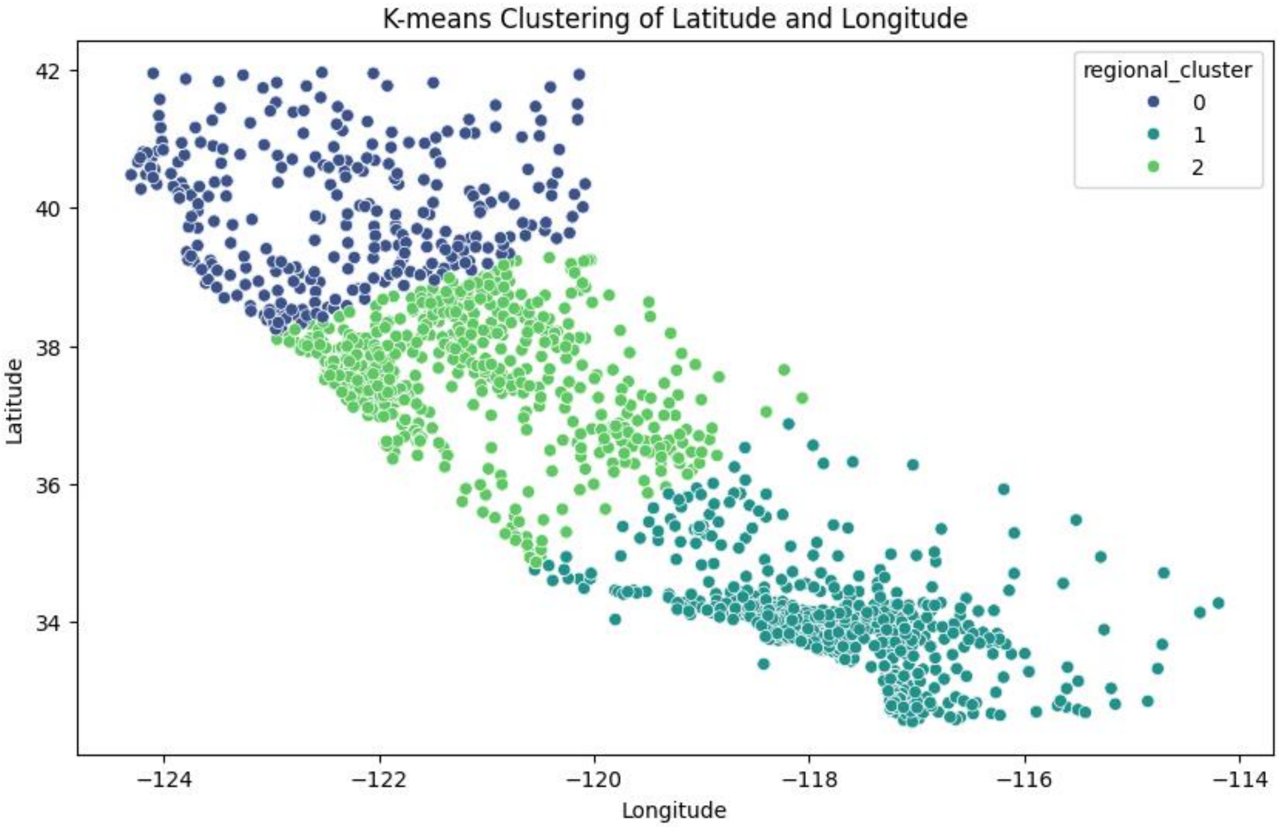
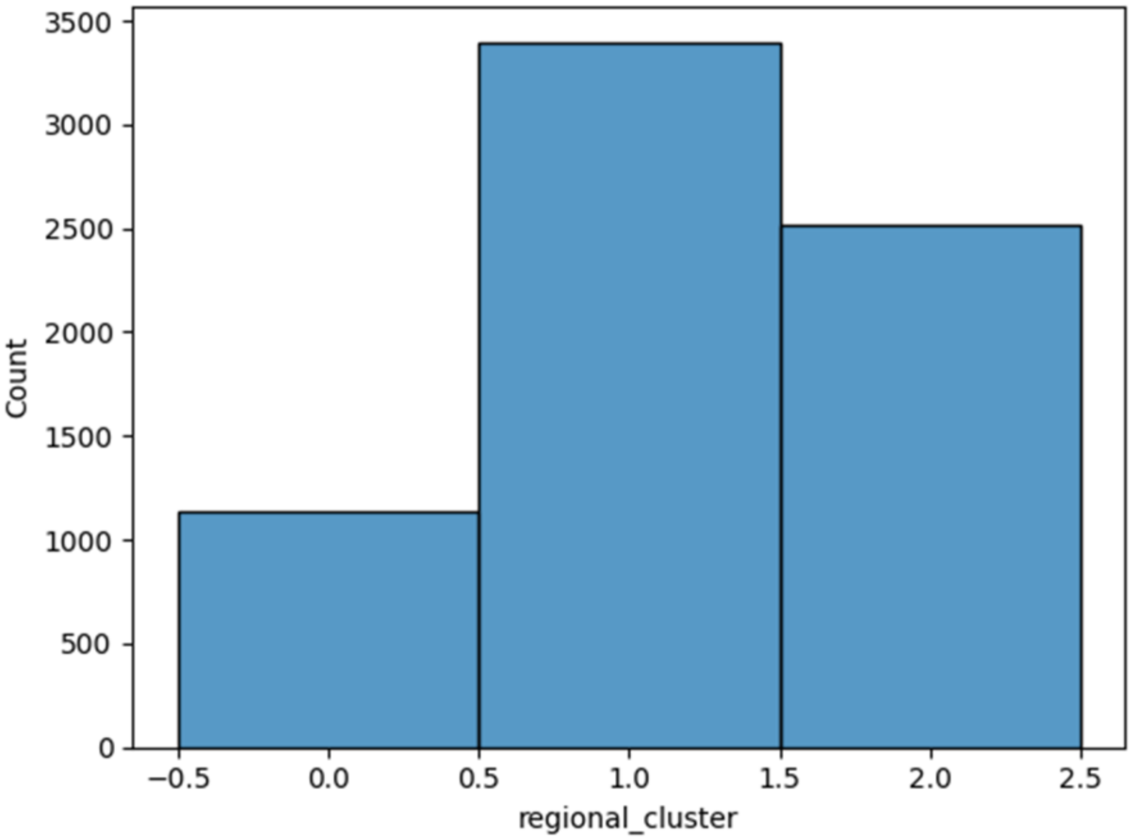
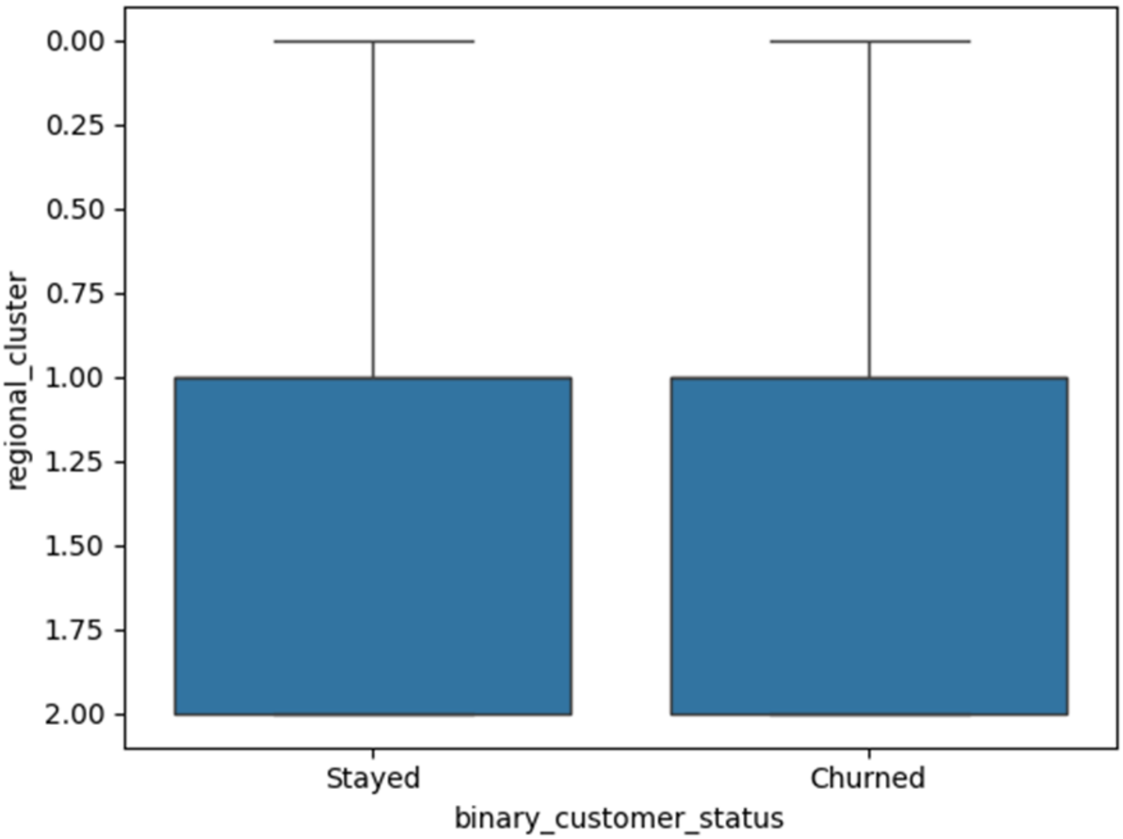


Figure 3 shows the results of using K-Means Clustering to convert Latitude and Longitude into regional clusters.

Figure 2

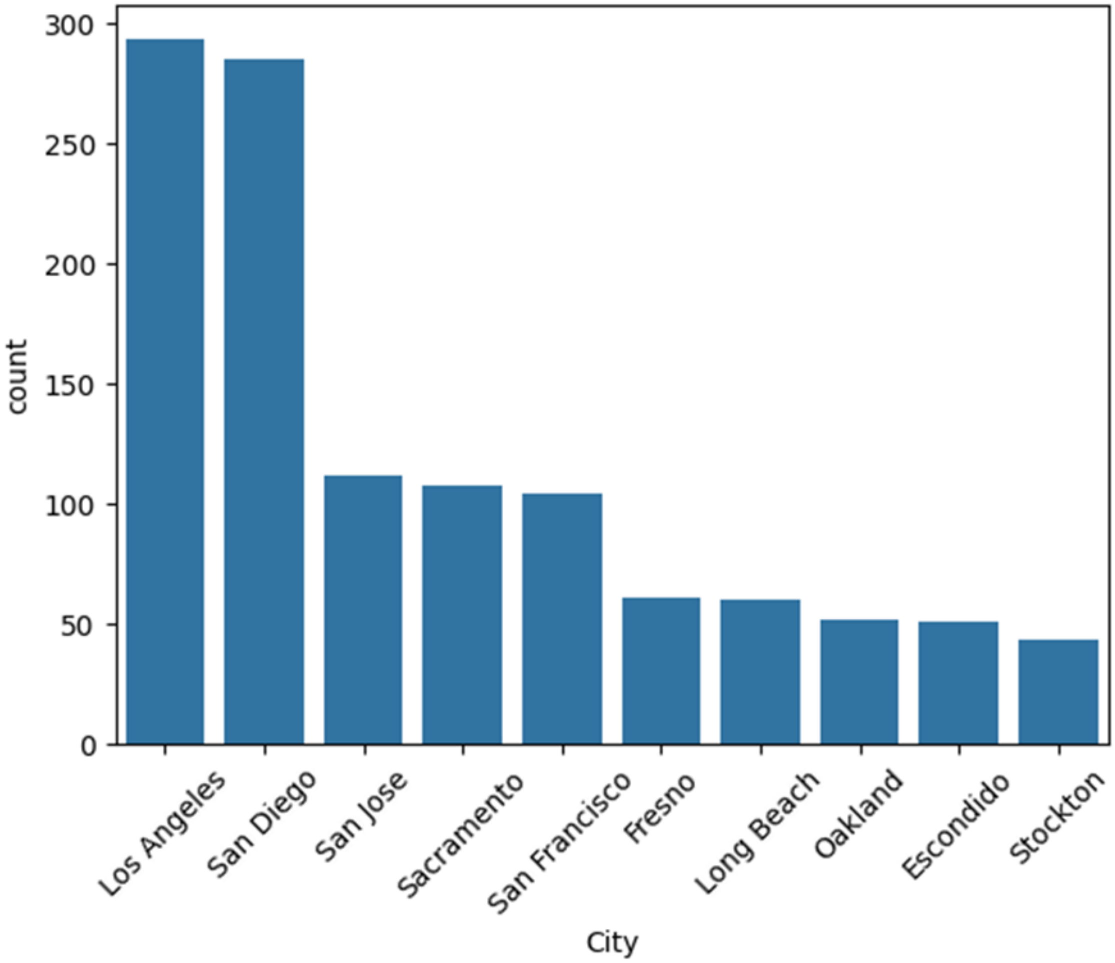

K-means Clustering of Latitude and Longitude
regional_cluster
0
1
2
Latitude
42
40
38
36
34
Longitude
−124
−122
−120
−118
−116
−114


3500
3000
2500
2000
1500
1000
500
0
Count
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
regional_cluster



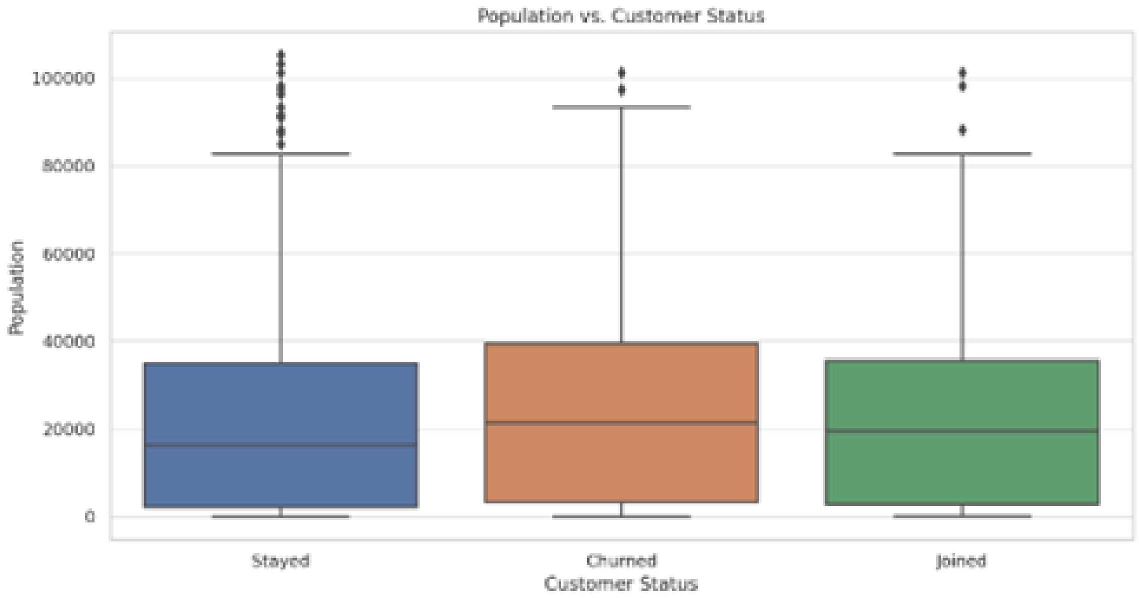
City

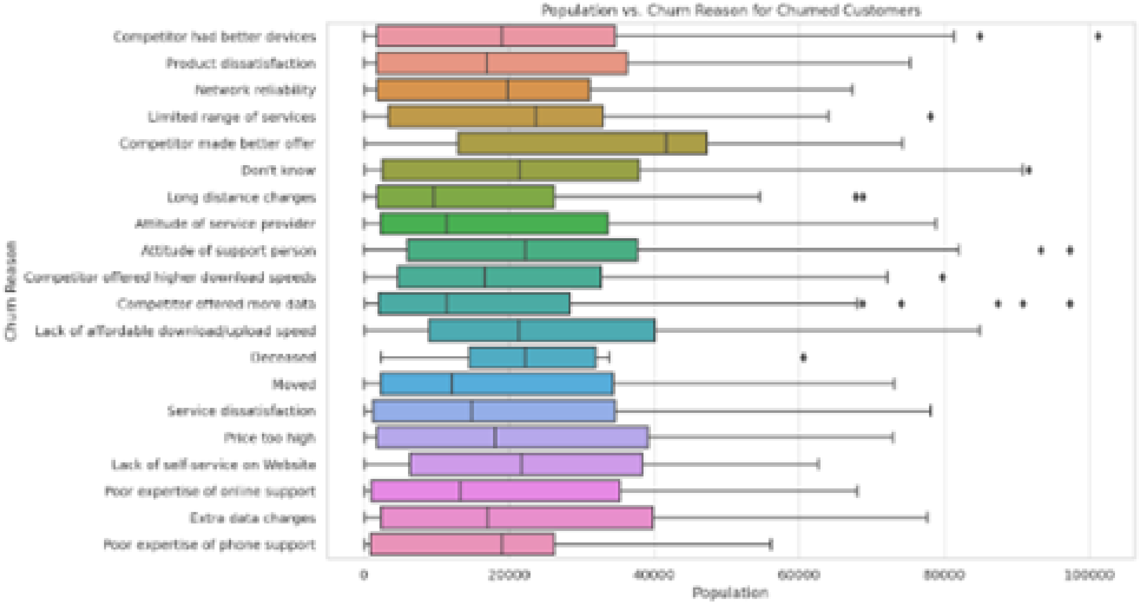
Since the customers' location spans California, the City column has too many unique values to include them in the machine learning models. The countplot below however shows that the cities with the highest number of customers are the major cities Los Angeles, San Diego, San Jose, Sacramento, and San Fransisco.



Population

The Population column shows the population of the zip code area each customer lives in. The boxplots below show that the population does not have a significant impact on customer churn alone. One thing that stands out is that the Population is significantly larger for the churn reason 'Competitor made a better offer'. This indicates that a potential intervention could be to create better offers that might keep people living in highly populated areas.

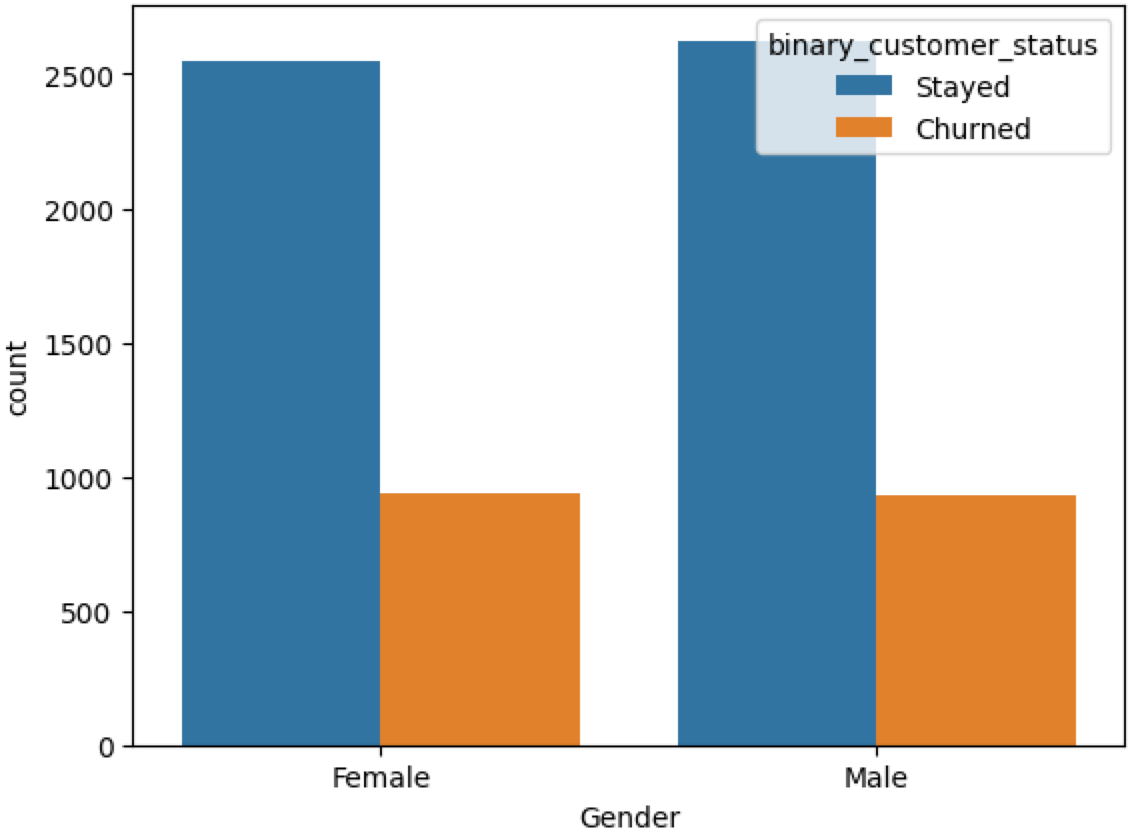




Demographic Data

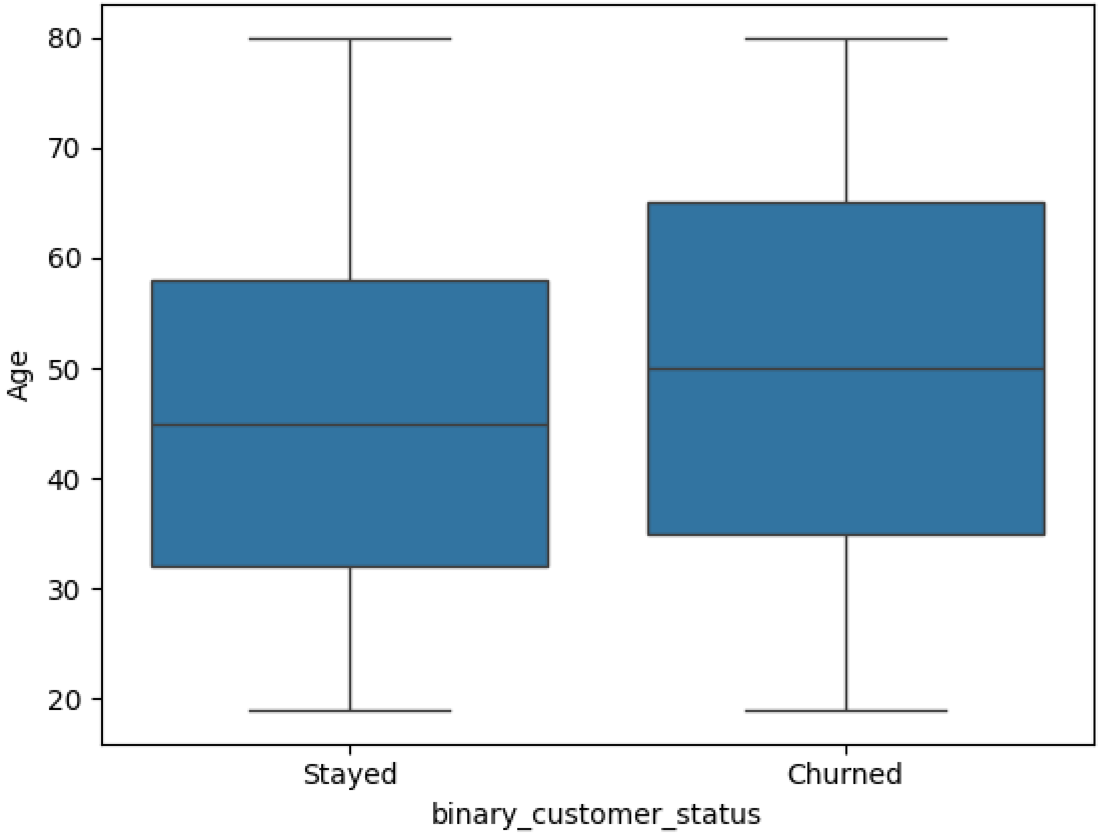
Gender

The distribution of genders among stayed and churned customers do not exhibit significant differences, indicating that gender might not be a decisive factor in predicting churn. This observation reinforces the idea that churn drivers are likely more related to service experiences, pricing, and value perception rather than demographic factors such as gender.



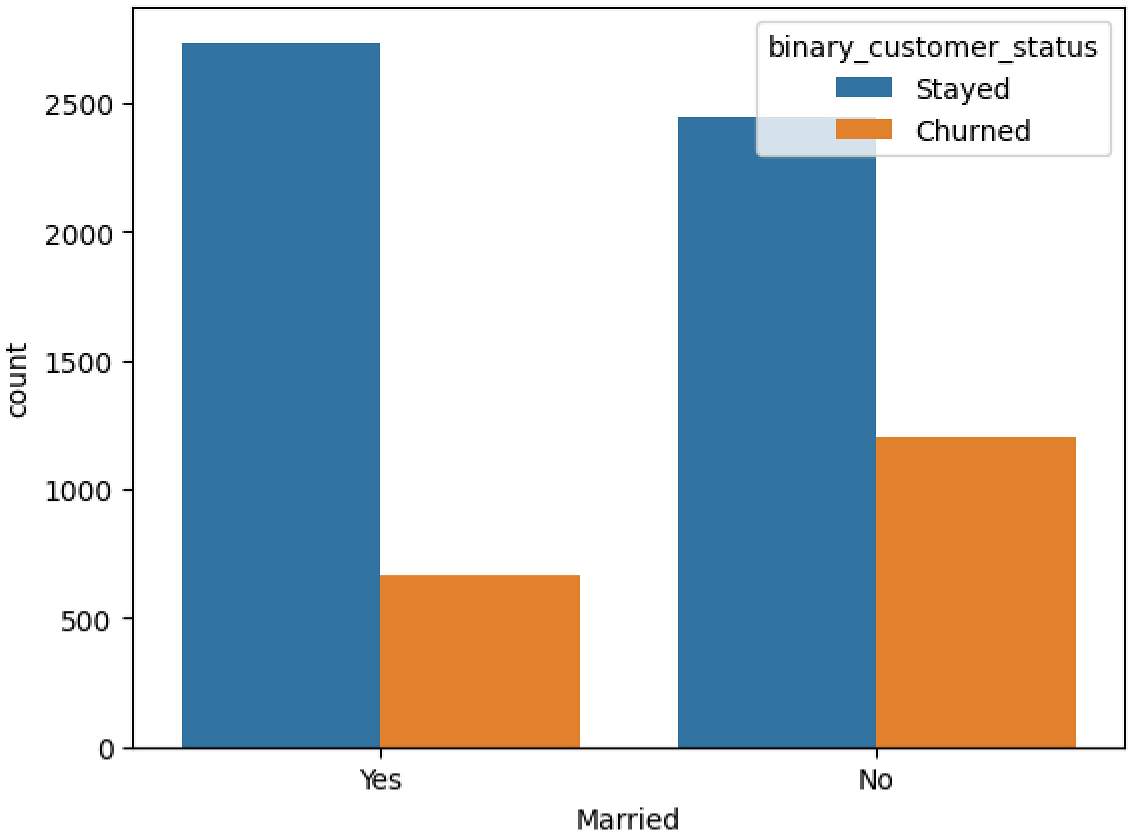
Age

The analysis indicated that the age distributions between customers who stayed and those who churned are notably similar, suggesting that age, as a standalone factor, might not strongly predict churn. This observation posits that factors beyond mere age demographics are at play in influencing customer decisions to leave or stay with the service provider.



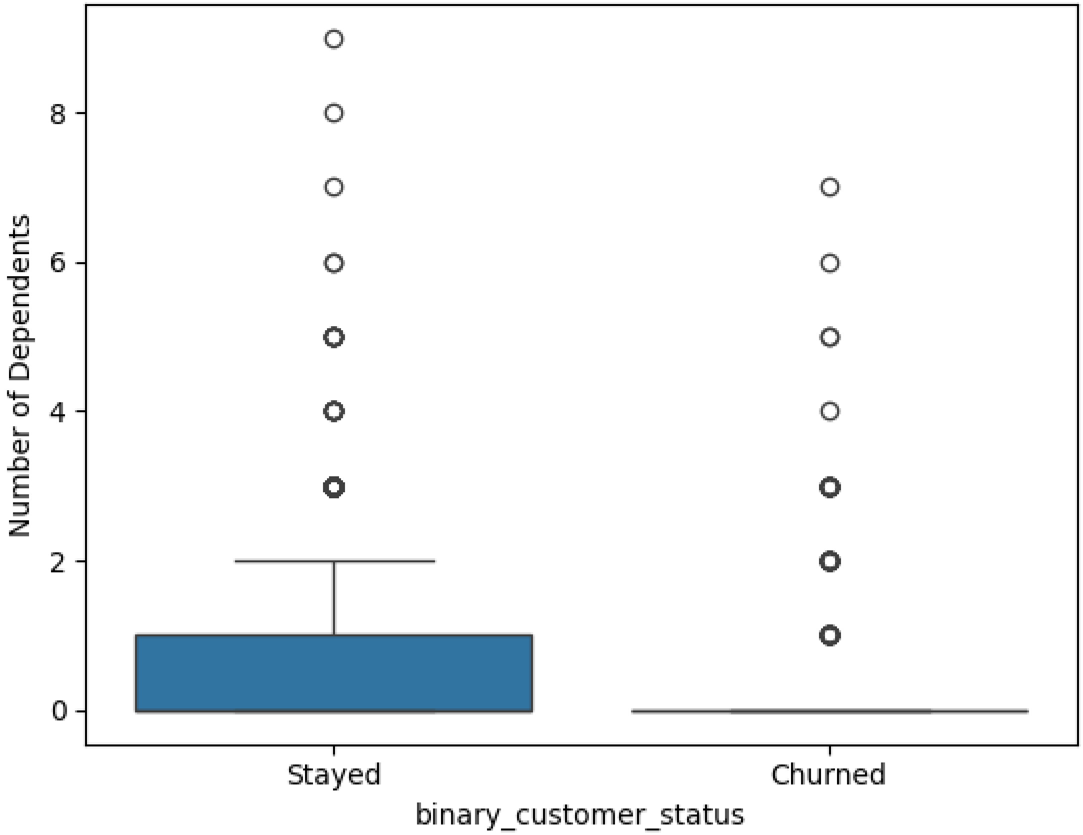
Married

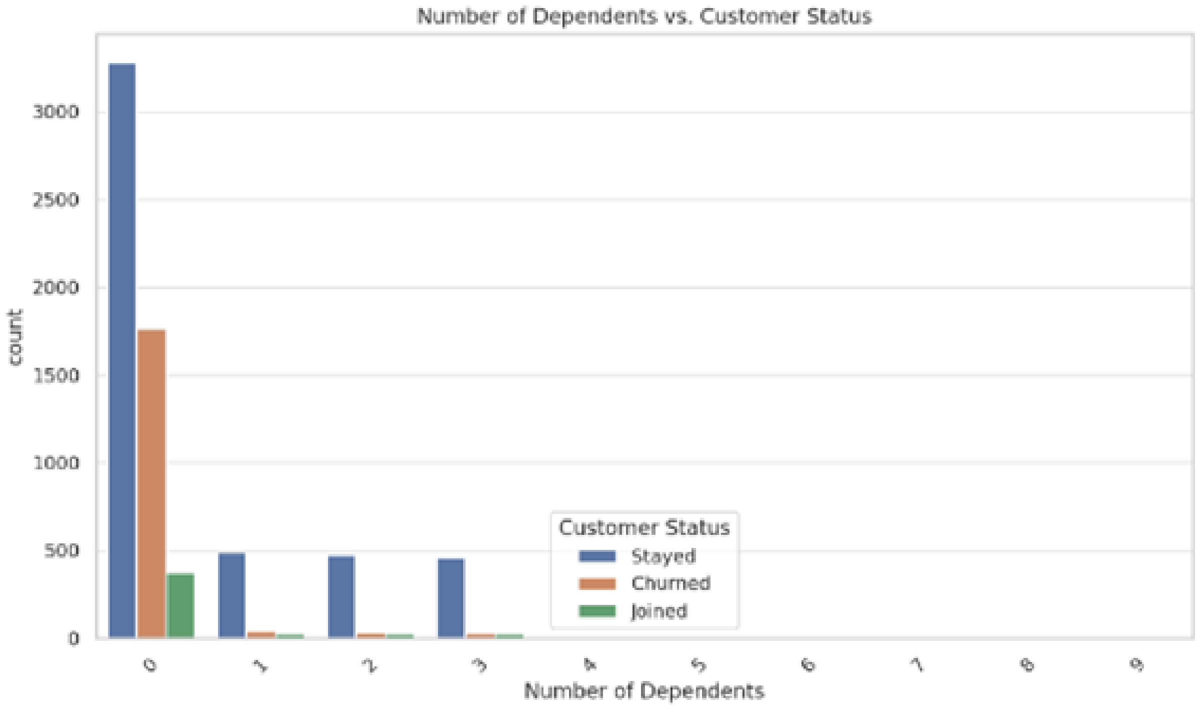
The distribution of married versus unmarried among stayed and churned customers shows some differences, indicating that being married might be a decisive factor in predicting churn.



Number of Dependents

Although the variation in the number of dependents between stayed and churned customers do not present a stark churn indicator, there is a subtle suggestion that customers with fewer dependents might be marginally more prone to churn. This aspect could point towards lifestyle or life stage factors influencing churn decisions, warranting further investigation into tailored marketing or service offerings that resonate with different customer segments.

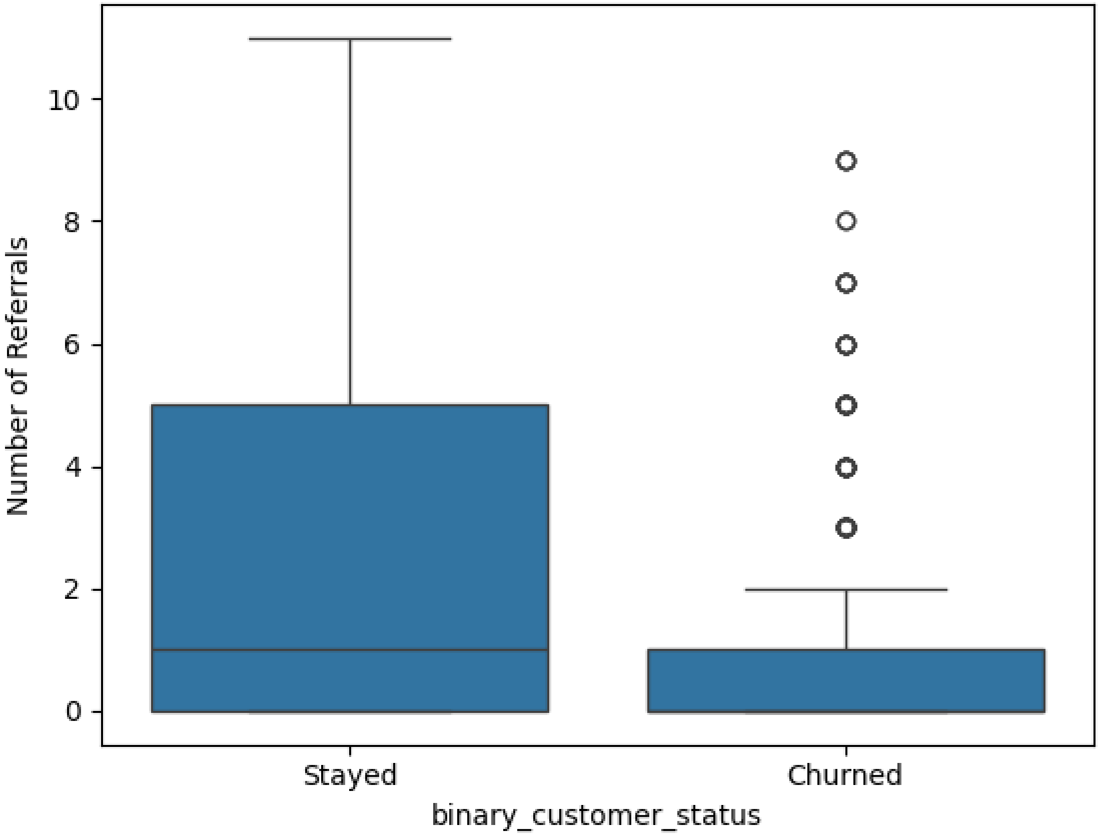

Number of Dependents
0
2
4
6
8
Stayed
Churned
binary_customer_status


Number of Dependents vs. Customer Status
count
0
500
1000
1500
2000
2500
3000
Customer Status
Stayed
Churned
Joined
0
1
2
3
4
5
6
7
8
9
Number of Dependents

Customer Relationship Data

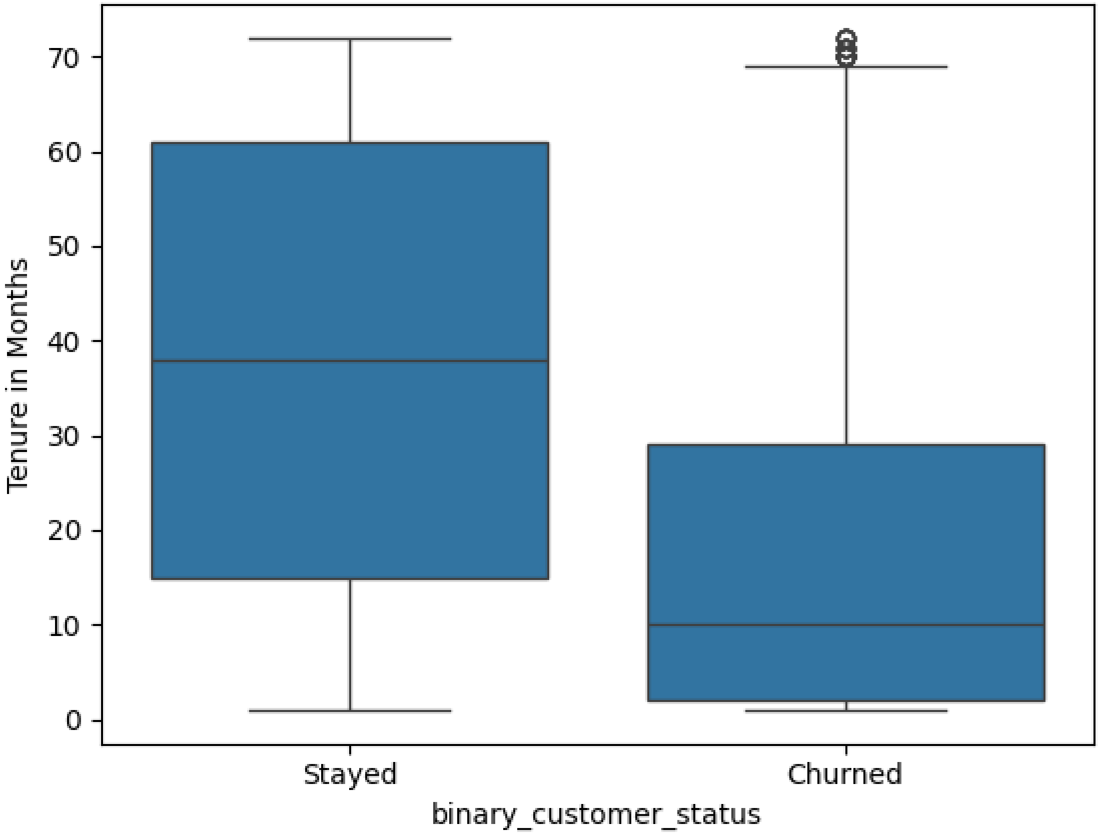
Number of Referrals

The Number of Referrals appears to offer some information to help determine churn. The boxplot below shows that there were not many churned customers that referred over 1 other person, while a significant number of those who stayed have referred 2 or more.



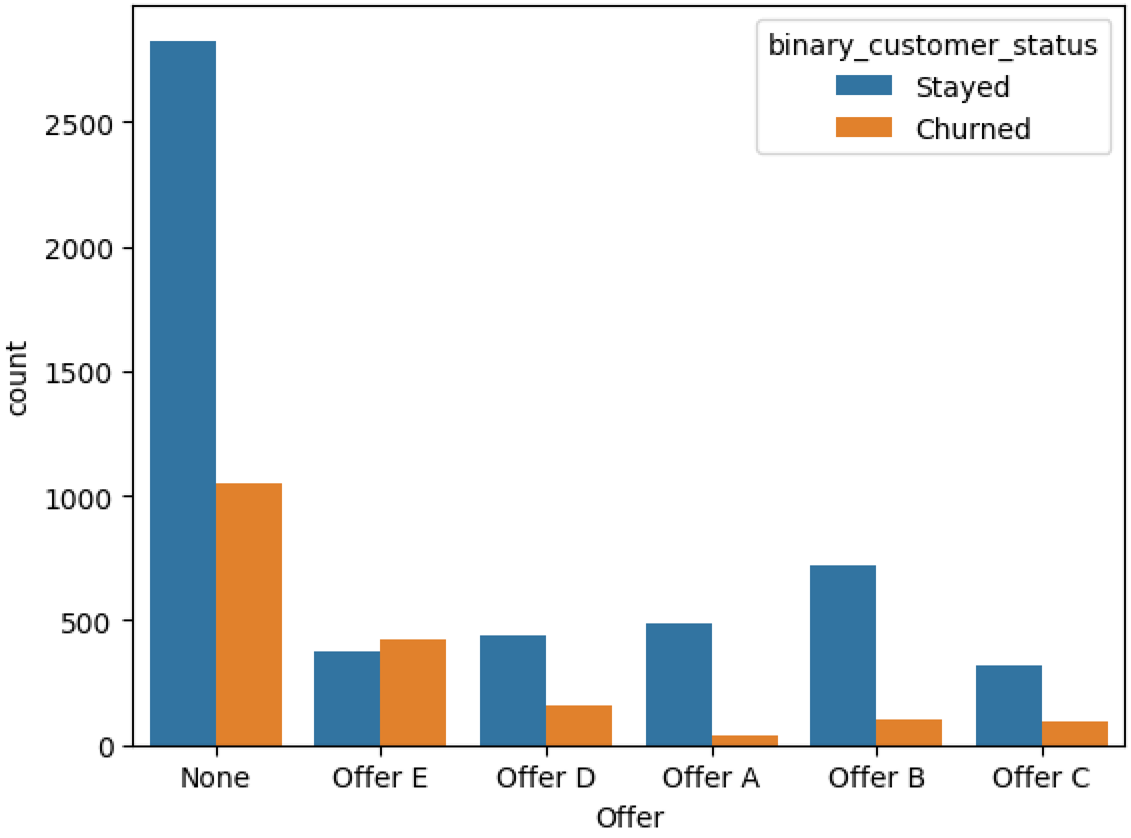
Tenure in Months

Tenure in Months shows that most of the customers who churned were customers for up to 2 years. It appears customers who stay longer than that are more likely not to stay with the company's services.



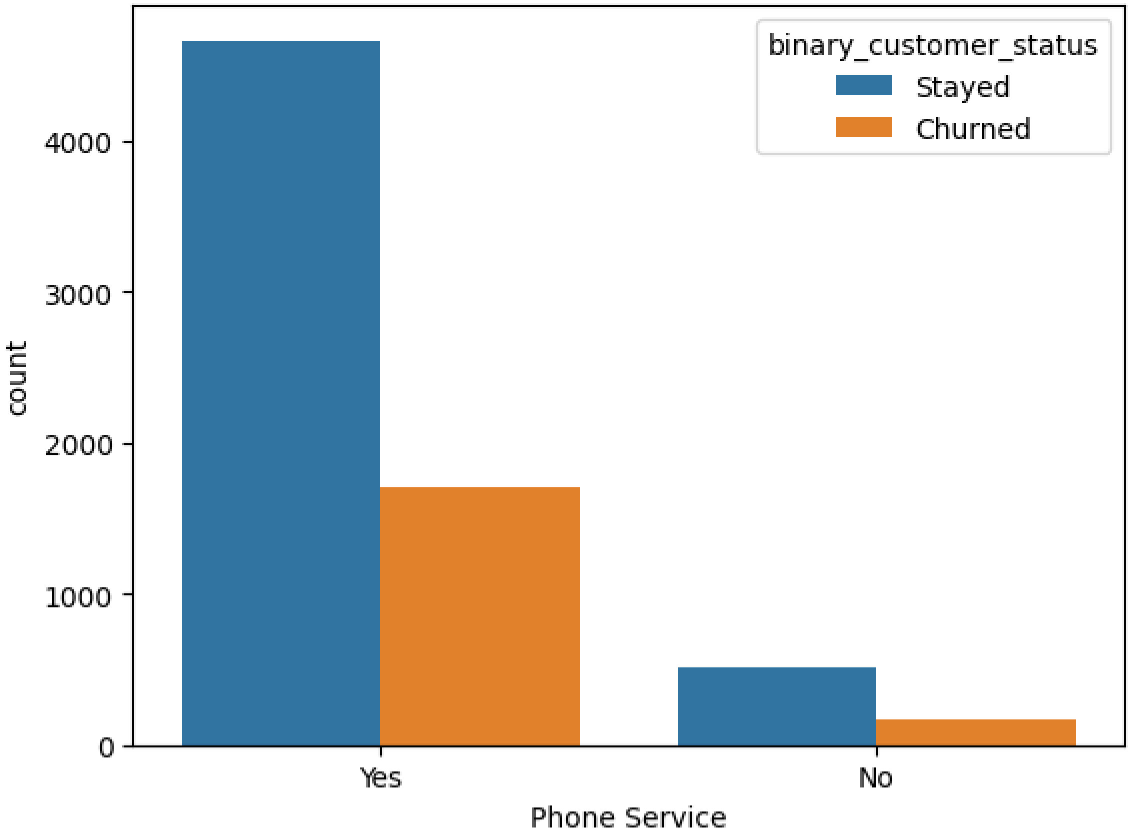
Offer

The Offer category shows that most customers did not have an offer. The count plot below shows that the customers taking part in Offer E had a significant number of churned customers, with more customers churned than stayed. This could be because this offer was old and coming to an end, or perhaps it was just not as valuable as customers initially thought.



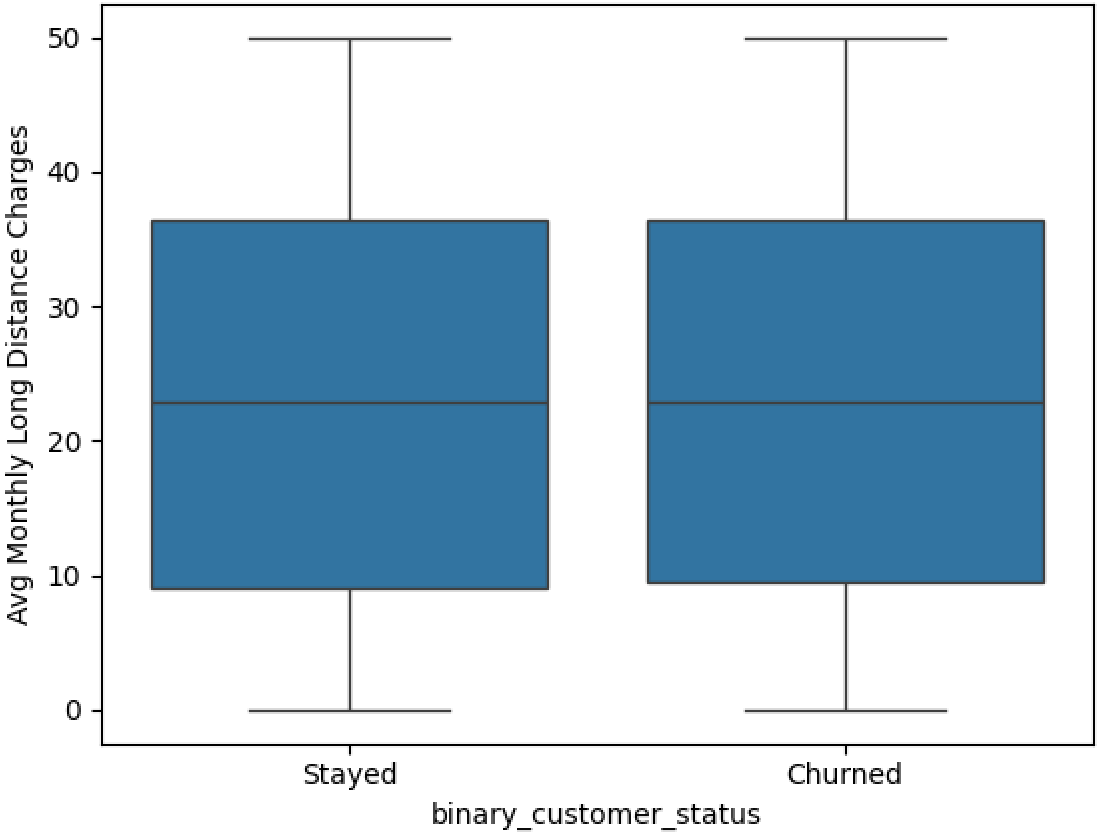
Phone Service

Whether or not the customer had phone service does not seem to make a difference when it comes to churn. The count plot below shows similar proportions of churn for those who had phone service and those who did not.



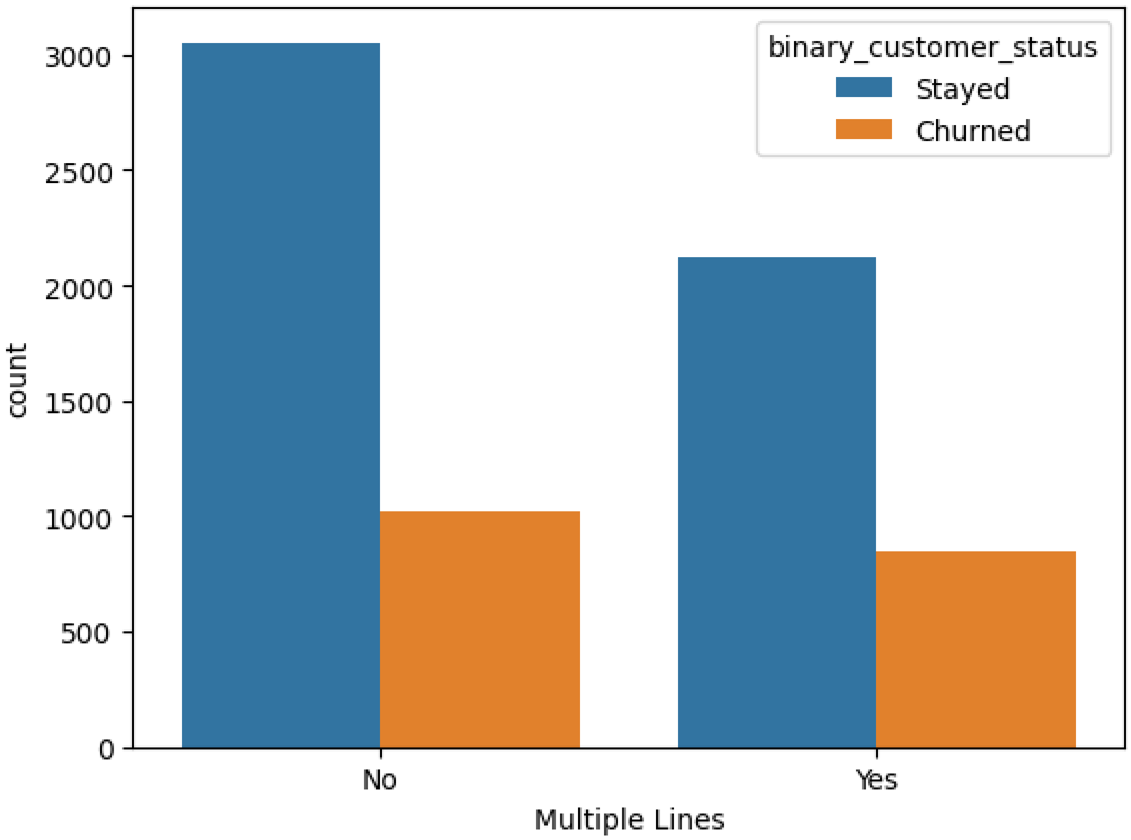
Avg Monthly Long Distance Charges

The Avg. Monthly Long Distance Charges do not appear to be a good predictor of churn. The boxplot below shows an almost identical distribution between churned customers and customers who stayed.



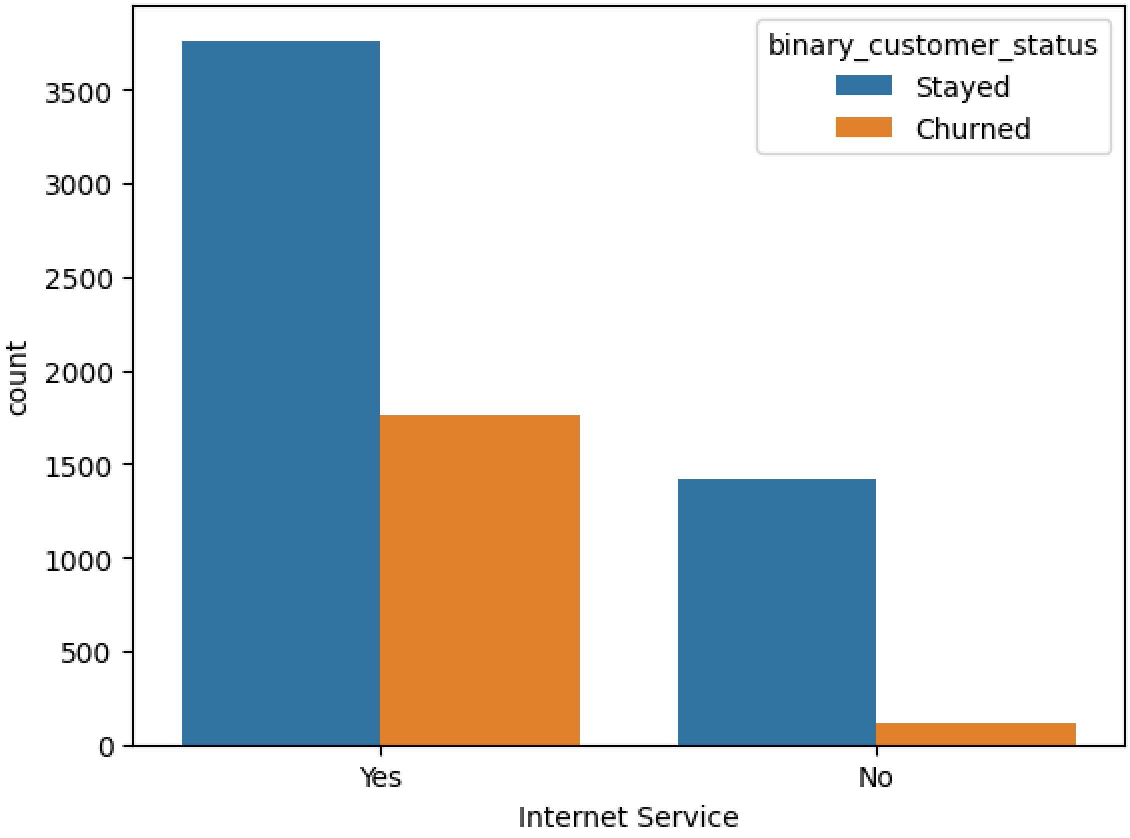
Multiple Lines

The Multiple Lines columns show that a greater proportion of customers who have multiple lines churned. This could indicate that competitors perhaps have better deals on multiple-line setups that the company could explore.



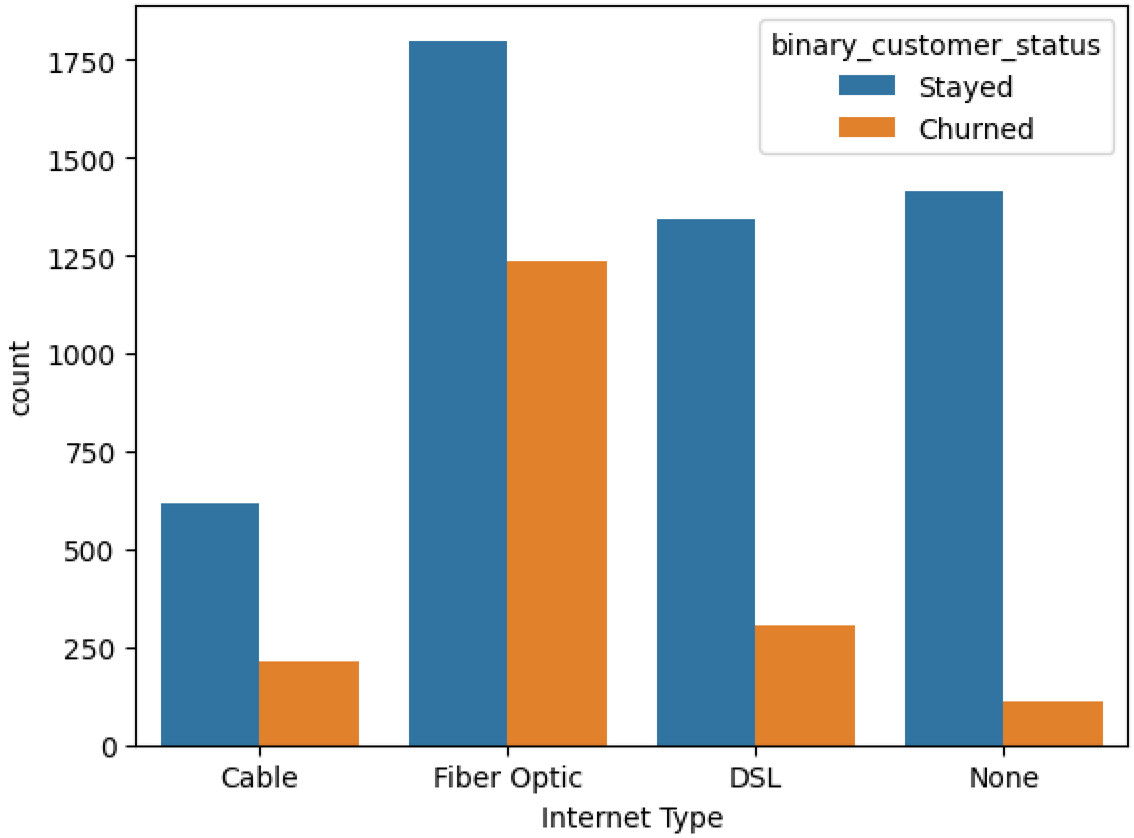
Internet

The Internet column shows that very few of the customers that did not have internet service through the company churned. The customers that had internet service showed a significantly larger churn rate.



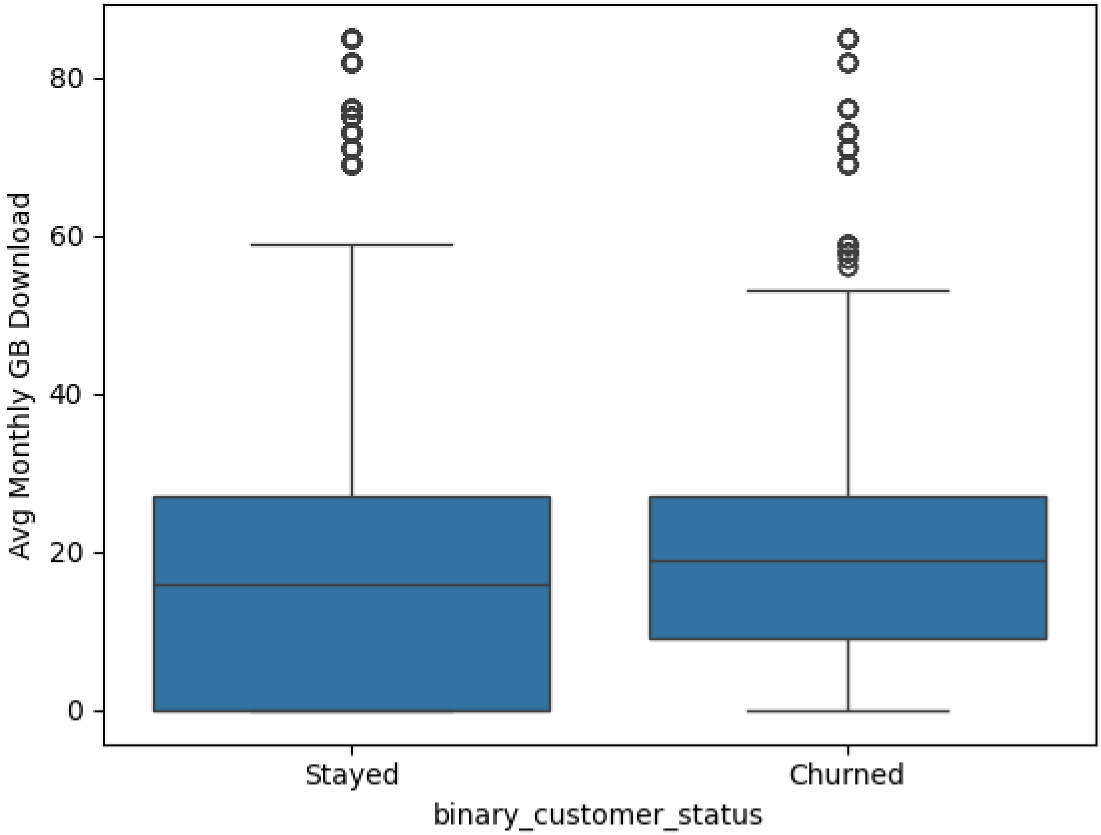
Internet Type

The Internet Type column offers some information to determine churn. It looks like customers with Fiber Optic service churned at a significantly greater rate than Cable, DSL, and those without internet service.



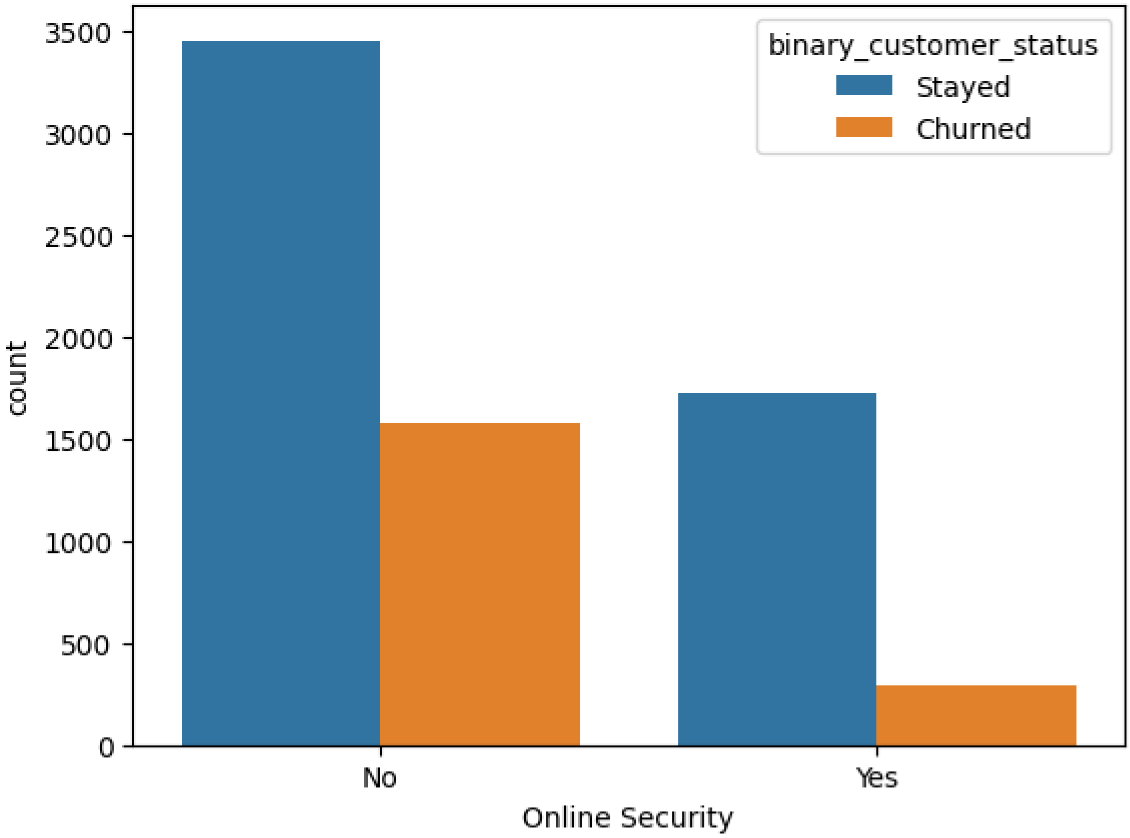
Avg Monthly GB Download

Avg Monthly GB Download offers little information that determines the churn rate. The distribution of Avg. Monthly GB Download is almost identical between customers who stayed and customers who churned.



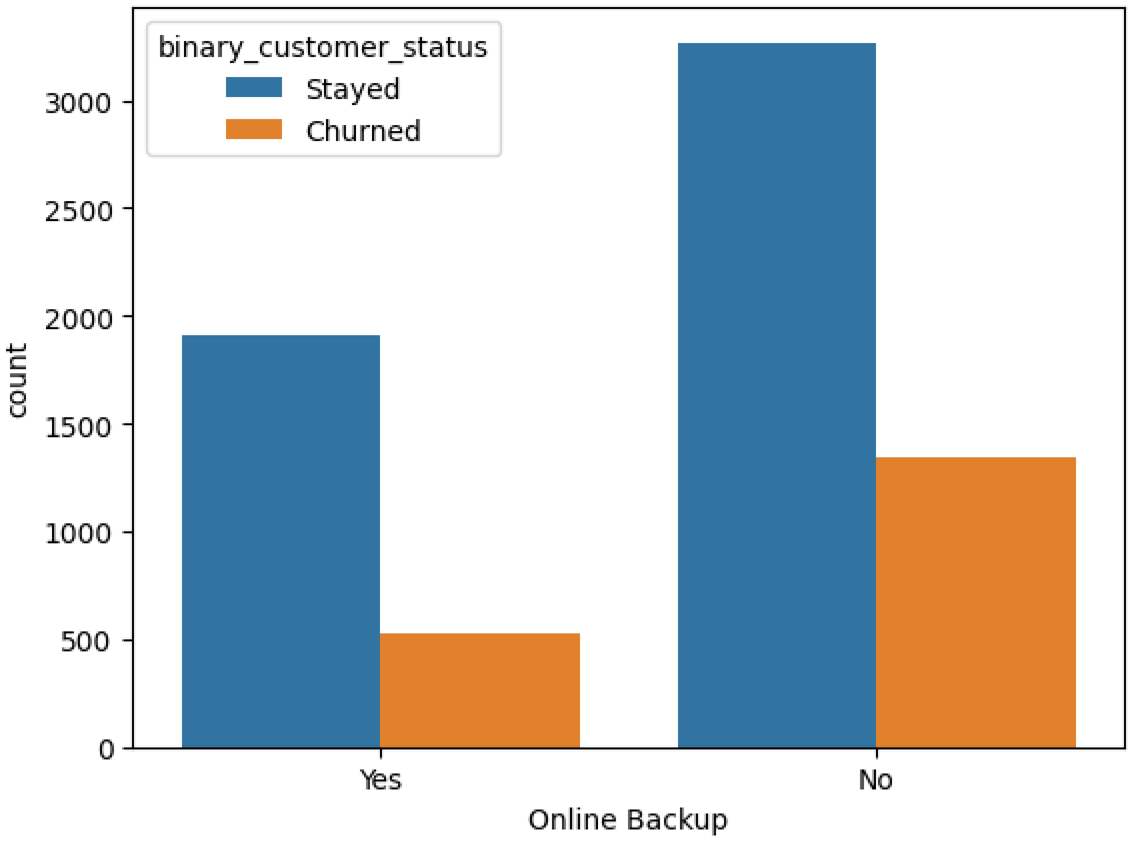
Online Security

The Online Security columns appear to show some differences between customers who stayed and those who churned. It looks like customers who do not have this service churned at a significantly greater proportion than those who had Online Security.



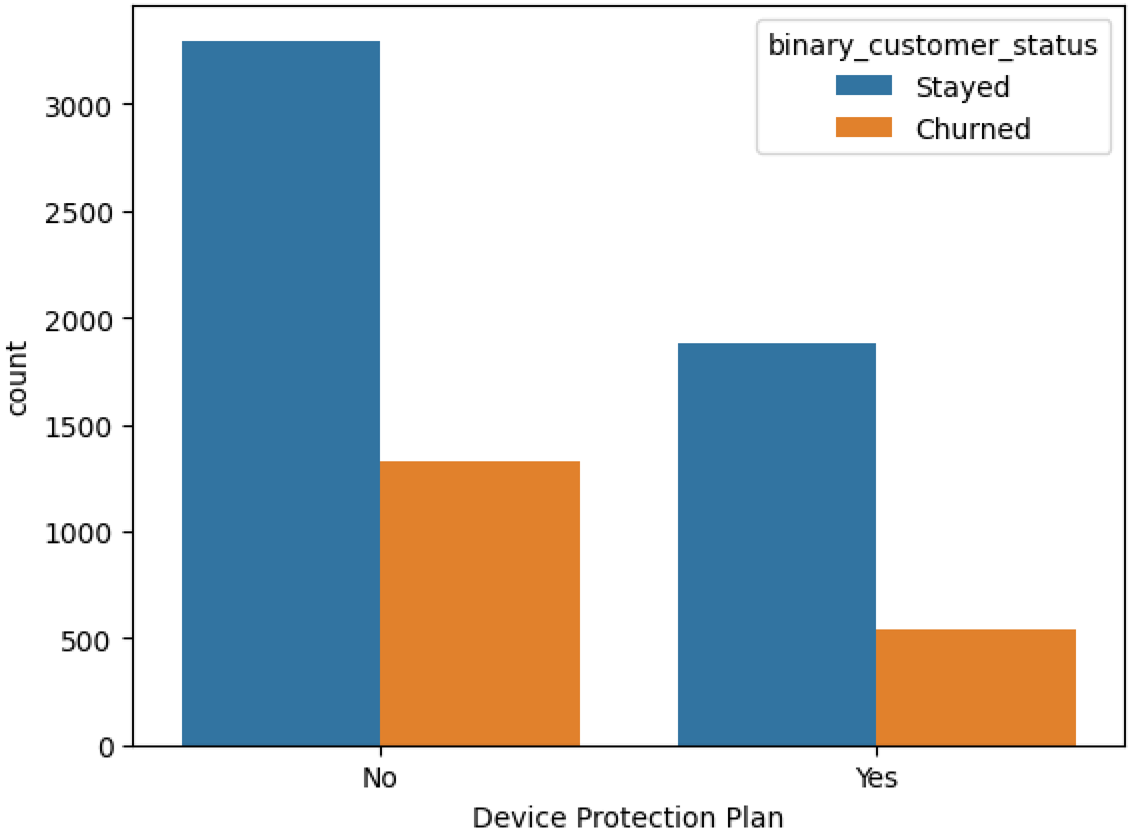
Online Backup

The Online Backup column shows only a small difference in churn rates. The proportion of customers who churned was almost the same between those who had Online Backups and those who did not.



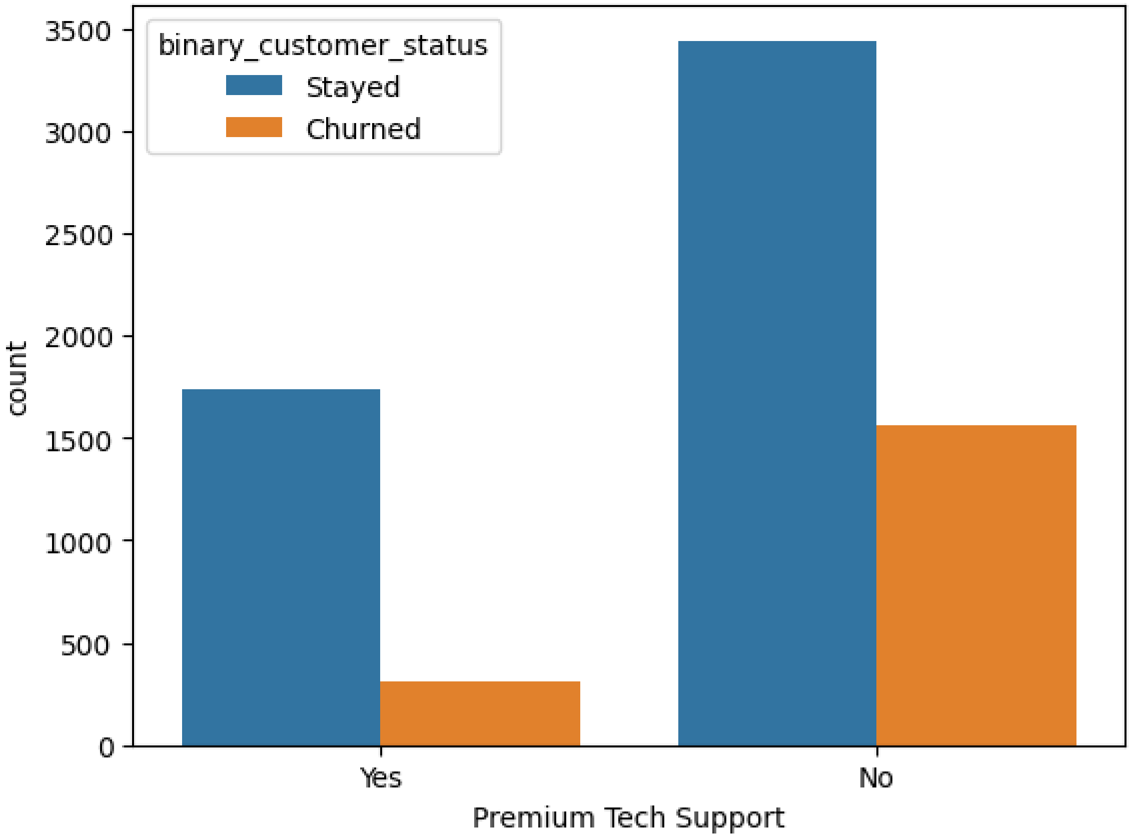
Device Protection Plan

The Device Protection Plan also offers little to no information to determine churn. The proportion of customers who churned was almost identical between customers who had a device protection plan and those who did not.



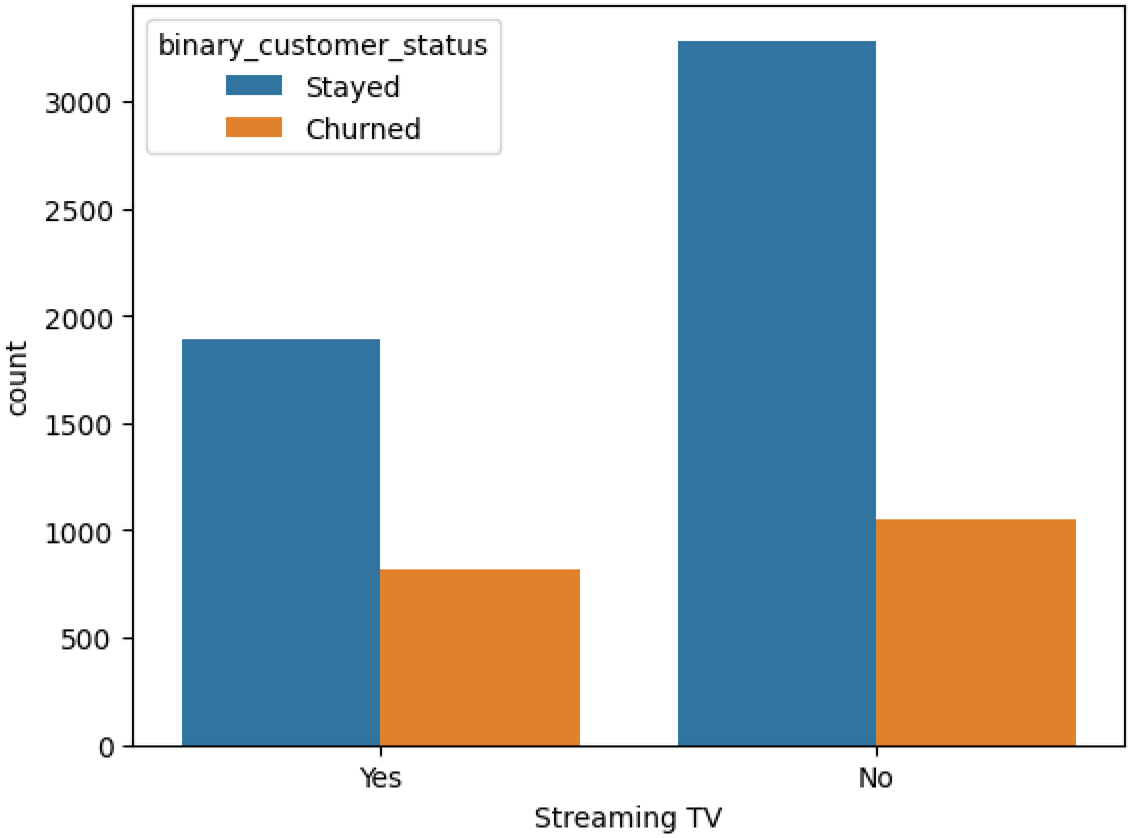
Premium Tech Support

The Premium Tech Support columns show that customers who had the service did not experience as high a churn rate as customers who did not have this service. This indicated the importance of providing a good customer experience.



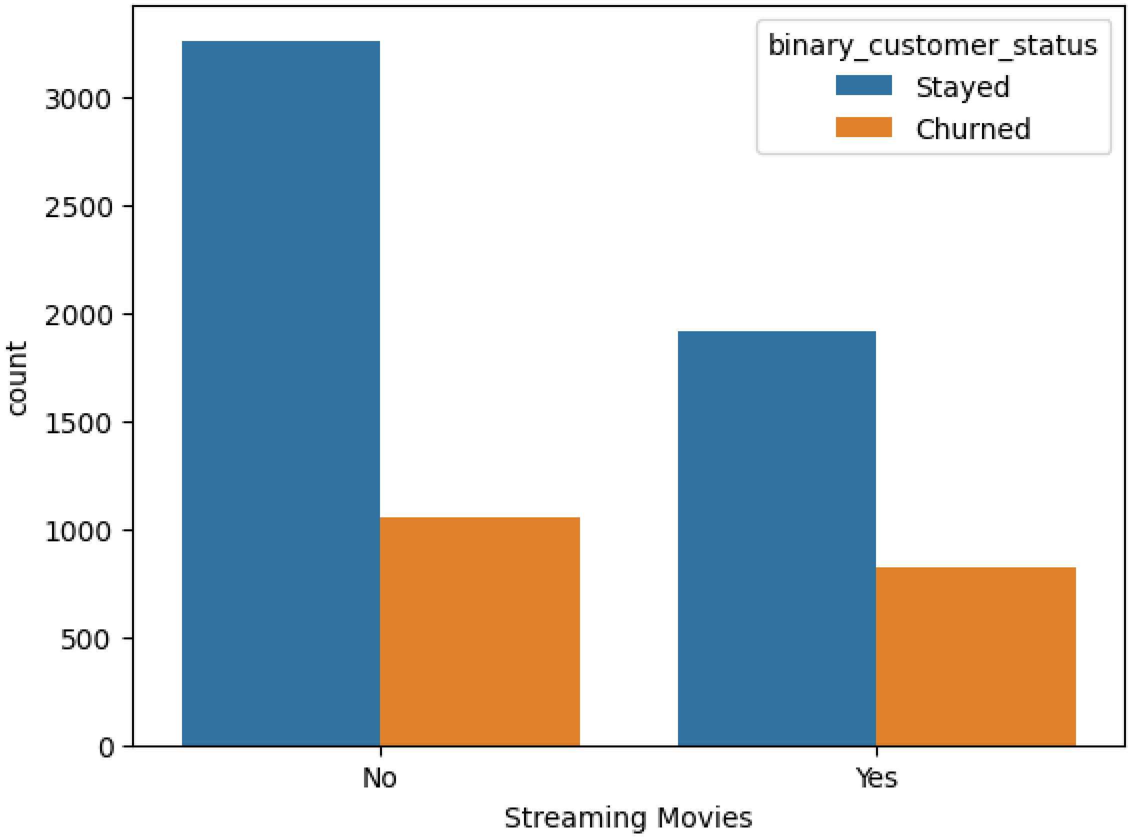
Streaming TV

The Streaming TV columns show that a greater proportion of customers who had Streaming TV churned. This might be a newer service the company is providing. Perhaps this same as part of a bundle or the customers that went for this service were just trying it out.



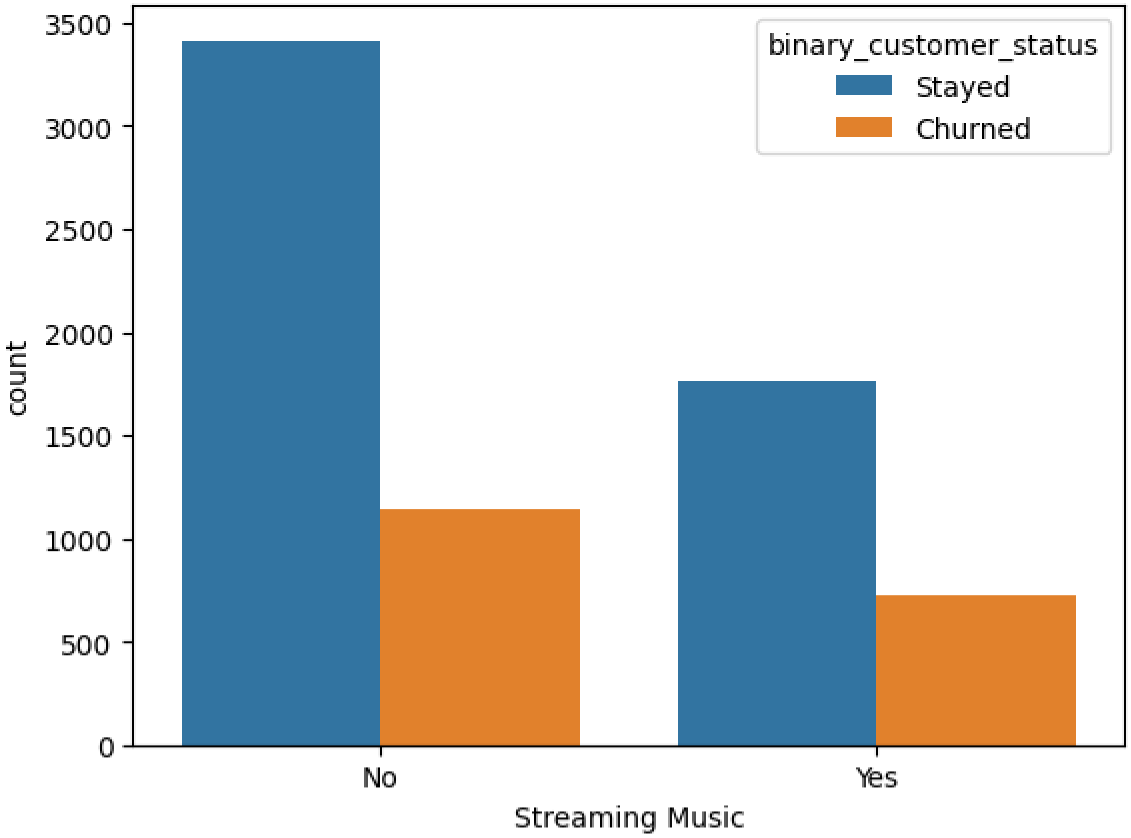
Streaming Movies

The Streaming Movies column is similar to Streaming TV in that it shows a greater proportion of customers with Streaming TV churned.



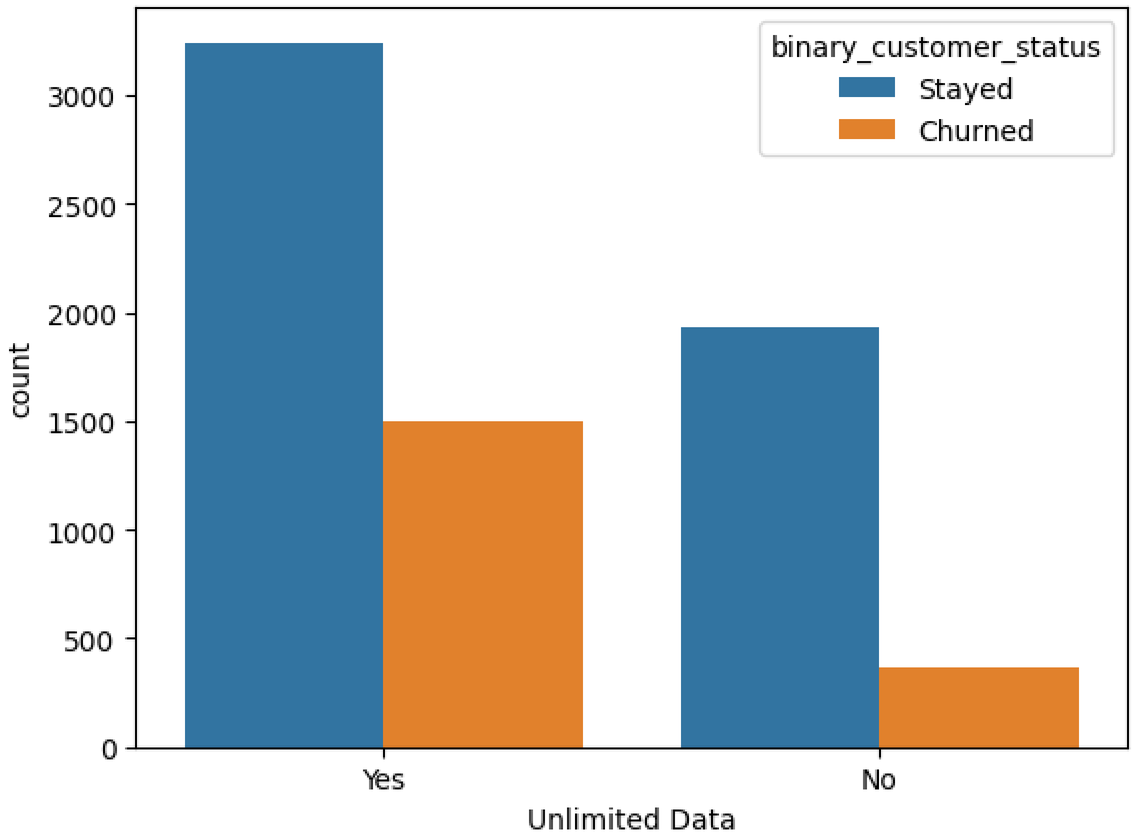
Streaming Music

The Streaming Music column shows a similar effect on churn as Streaming TV and Movies. It appears a greater proportion of customers who have Streaming Music churned.



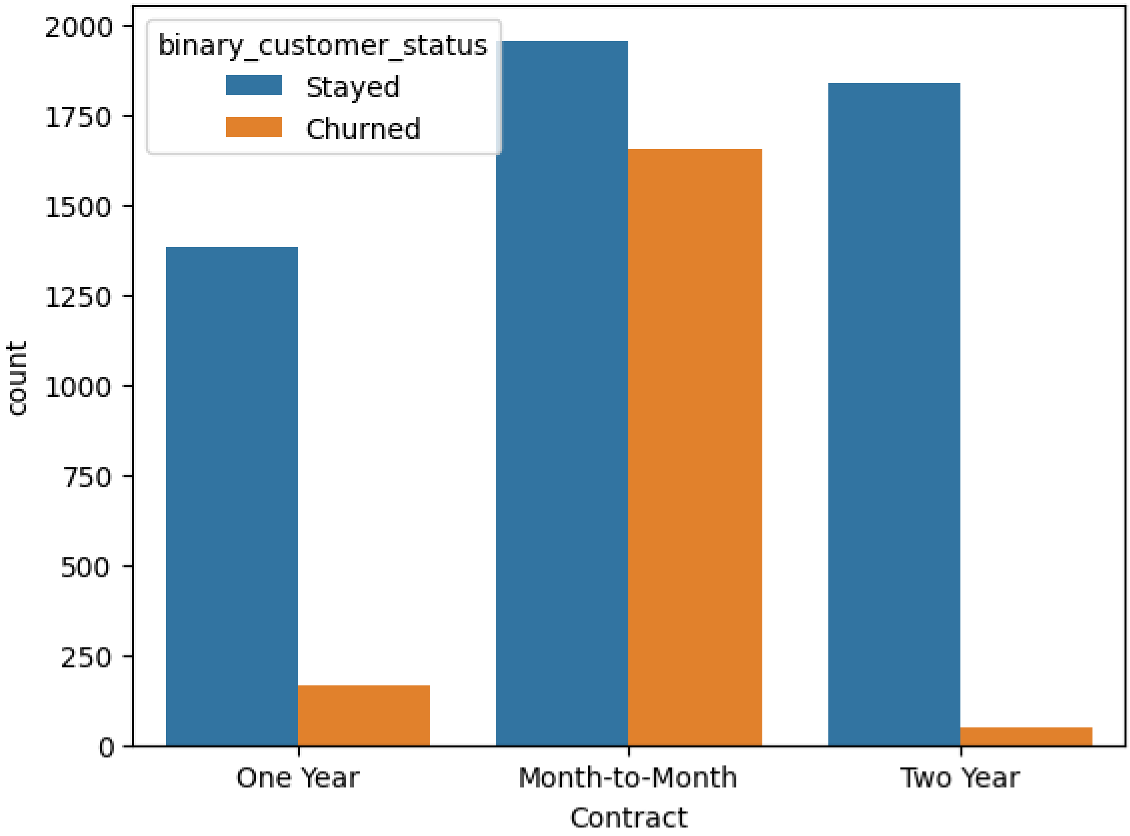
Unlimited Data

The Unlimited Data column shows that the customers who had Unlimited data churned at a greater rate than customers who did not have Unlimited data.



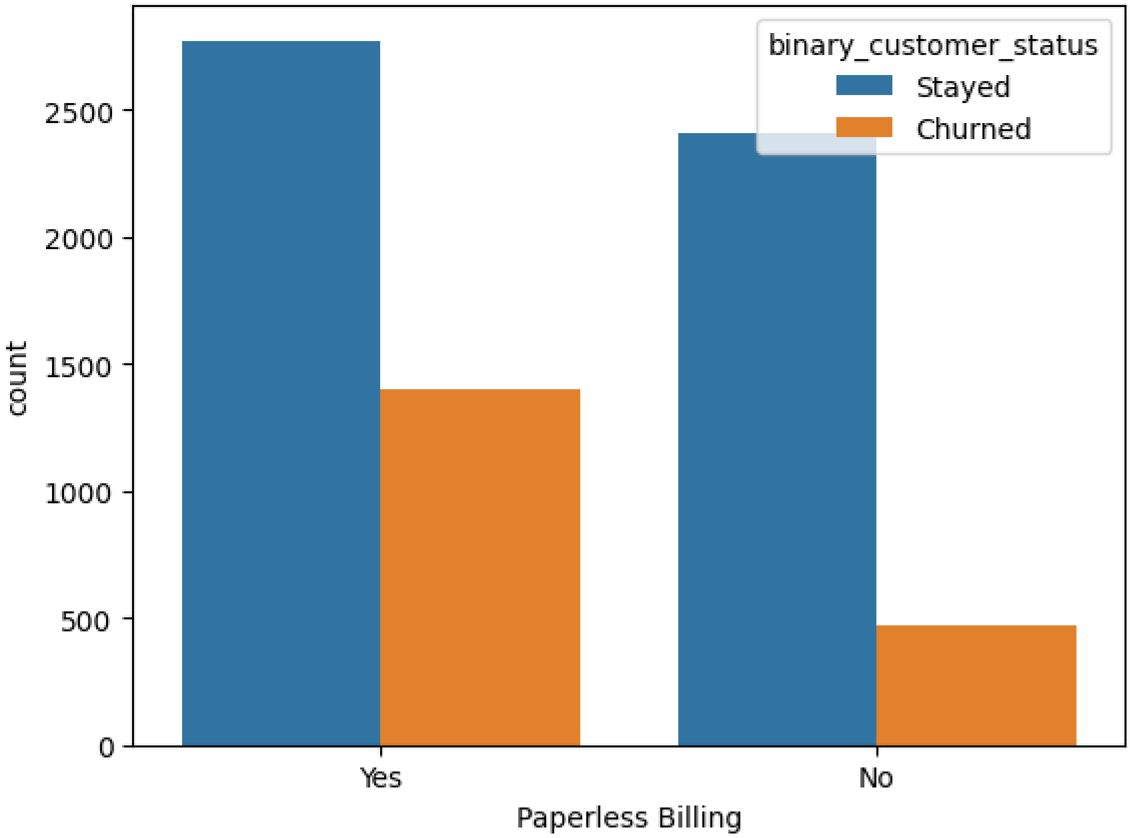
Contract

The Contract columns appear to be a good predictor for Churn. The customers who are on a Monthly plan have a significantly higher churn rate than those on a One or Two Year contract.



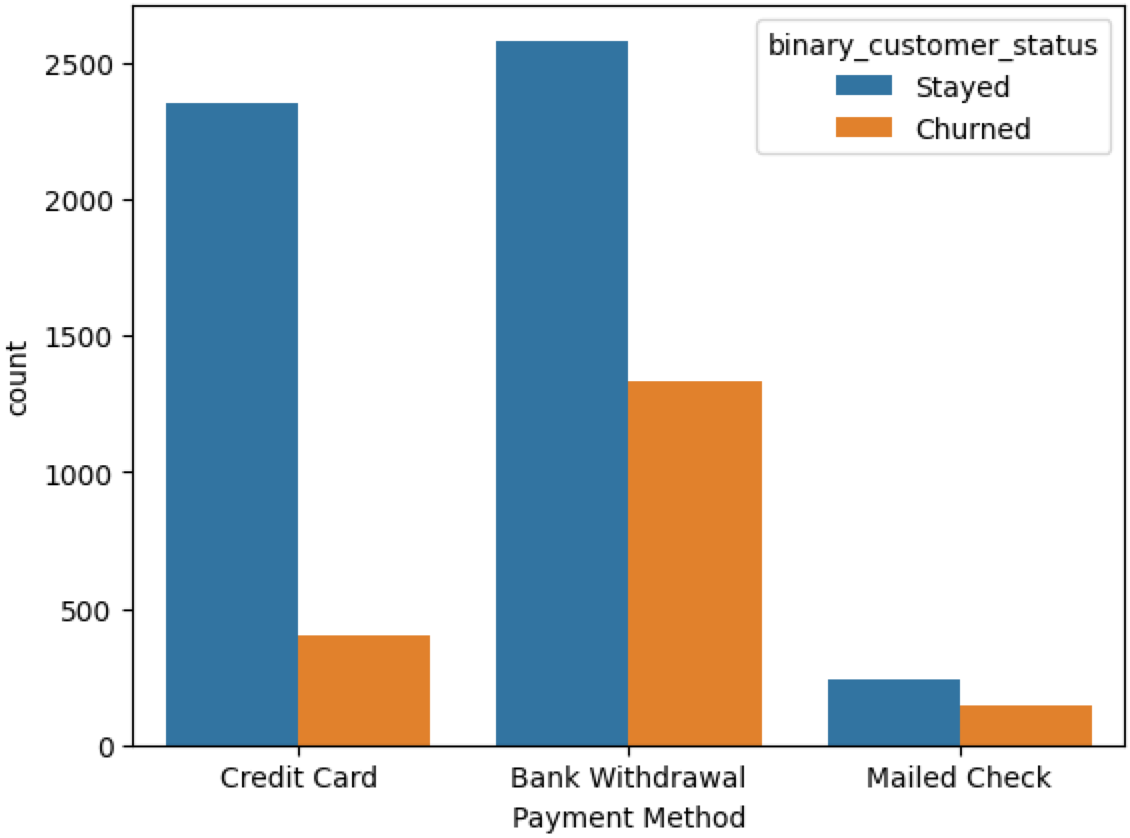
Paperless Billing

The Paperless Billing column shows a slight difference in churn rate. Customers who had paperless billing churned at a greater rate than those who did not.



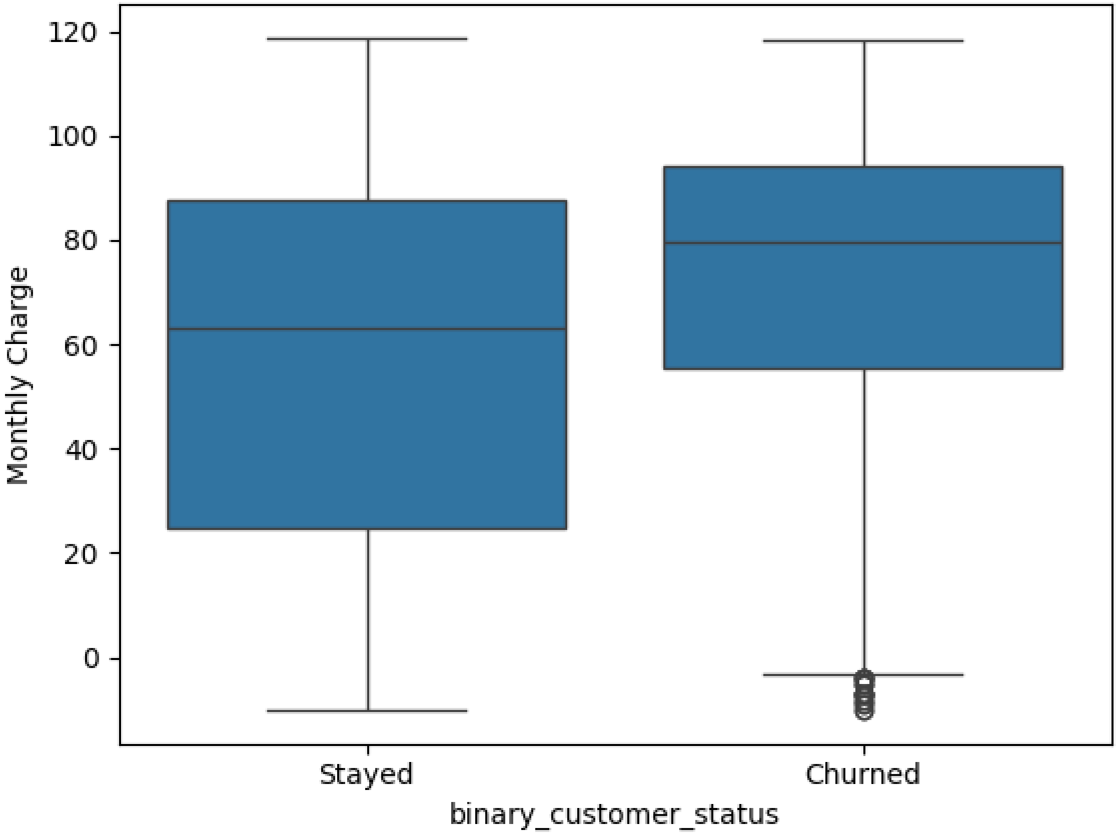
Payment Method

The Payment Method column shows that customers who pay with a credit card churn significantly less than customers who pay with a bank withdrawal or mail-in check.



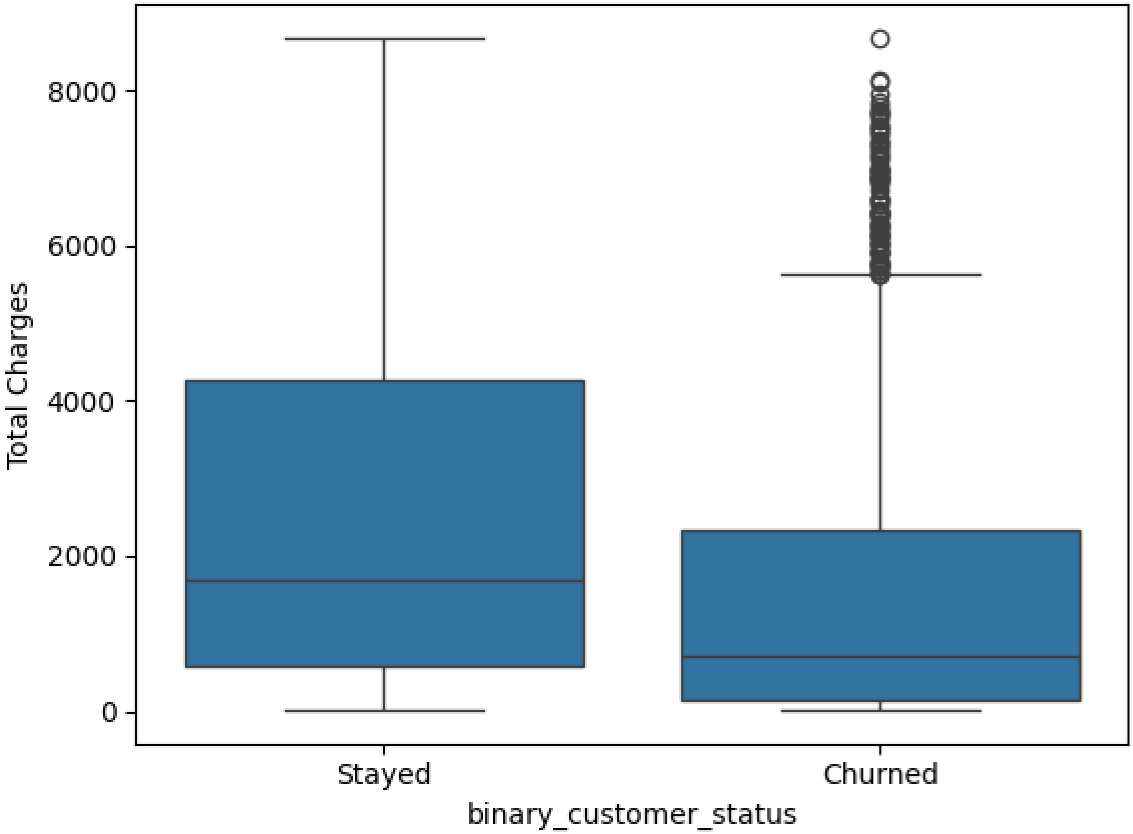
Monthly Charge

A pronounced difference in the distribution of monthly charges between customers who stayed and those who churned was observed. Customers who churned tend to face higher monthly charges, signifying that pricing plays a pivotal role in churn. This finding underscores the necessity for the telecom provider to reassess its pricing strategies, perhaps considering more competitive or flexible pricing models to enhance customer retention.



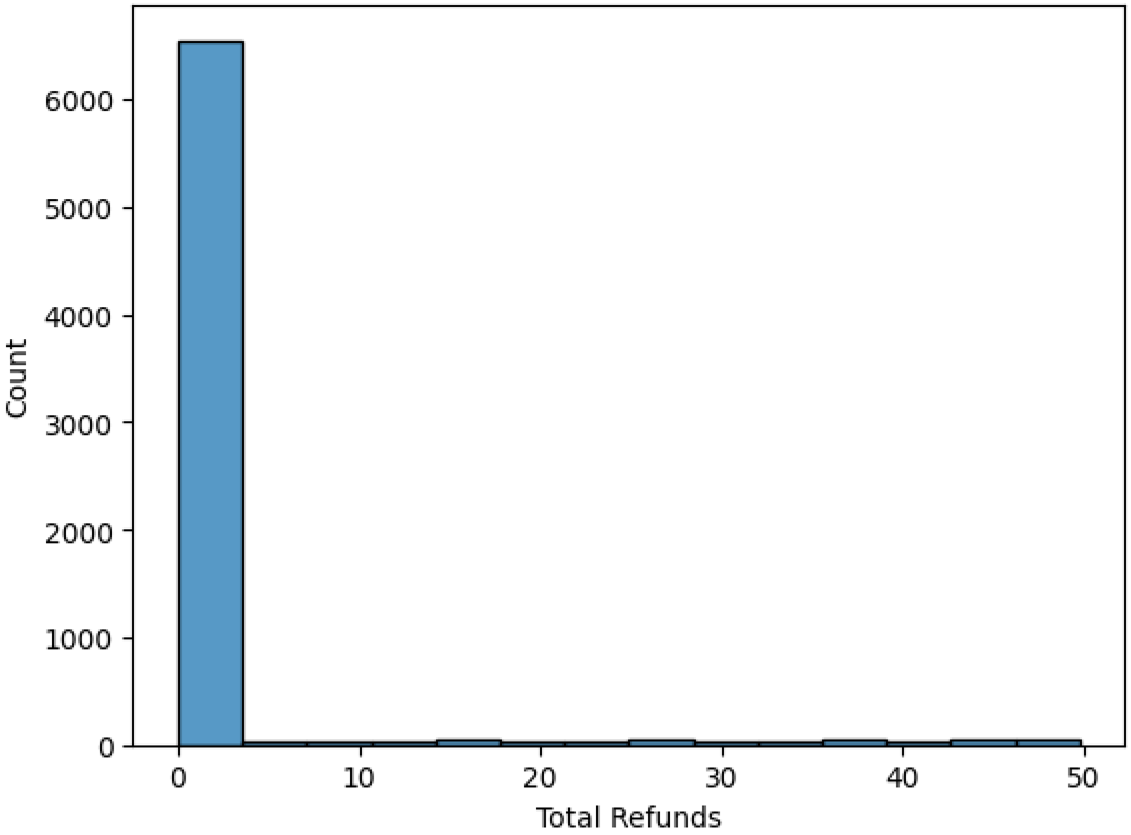
Total Charges

The analysis revealed that stayed customers generally have higher total charges, implying that longer tenure or a broader engagement with the service (through subscriptions or additional services) correlates with a reduced likelihood of churn. This insight suggests that fostering long-term relationships and encouraging deeper engagement with the service could be key strategies for minimizing churn rates.



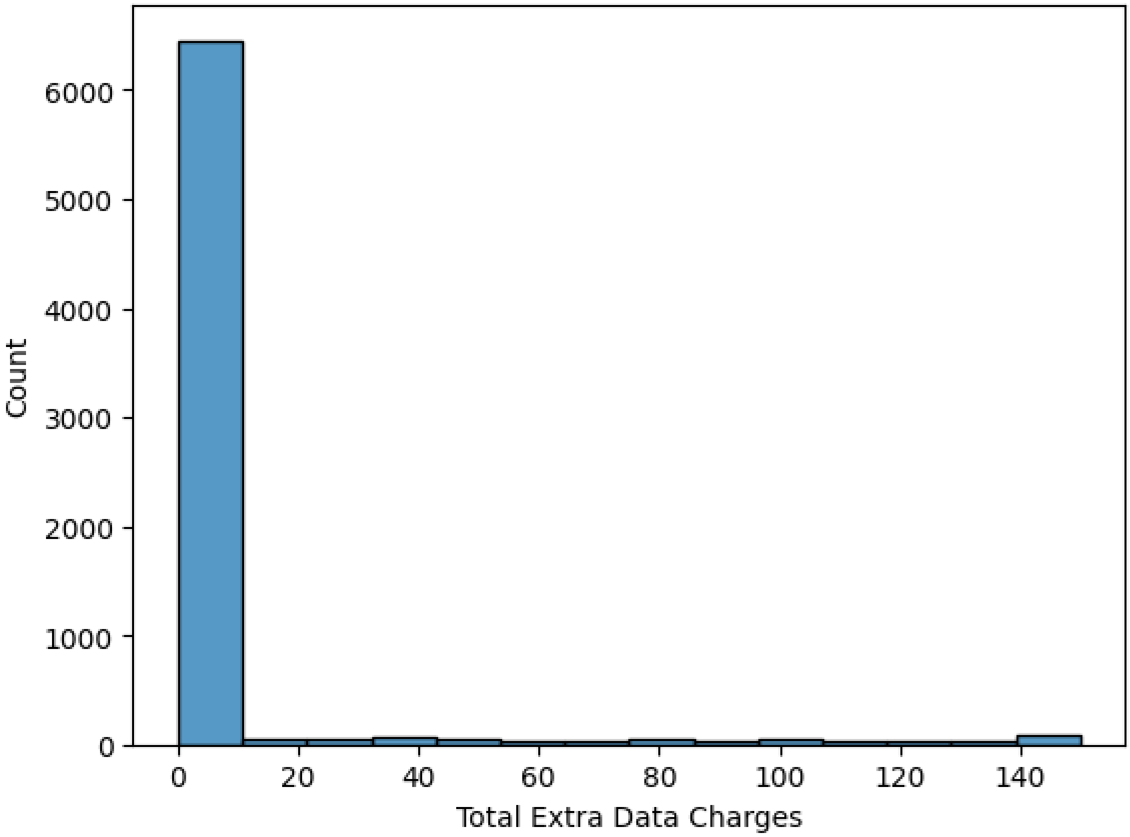
Total Refunds

The Total Refunds column does not offer much information because the majority of customers don’t have any refunds. There are simply not enough customers with refunds to be able to use it to predict churn.



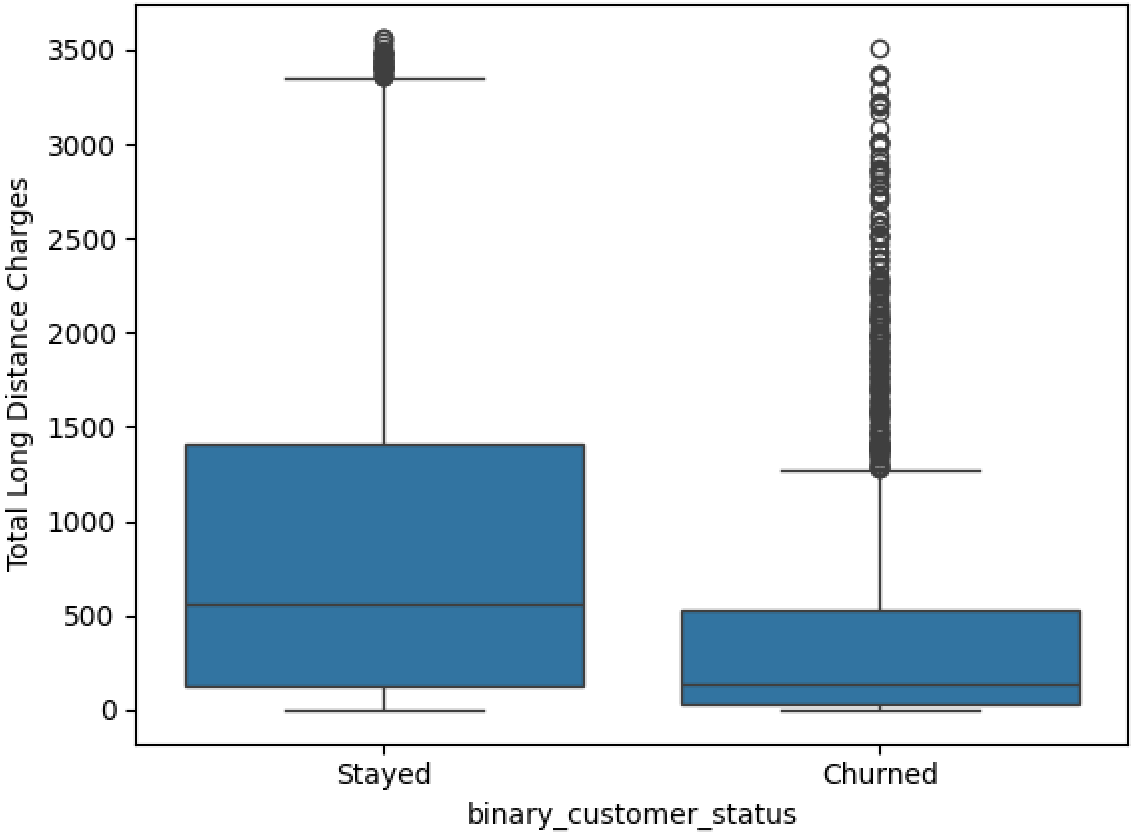
Total Extra Data Charges

Like Total Refunds, the Total Extra Data Charges column is so heavily skewed that it would not be an effective feature to predict churn with.



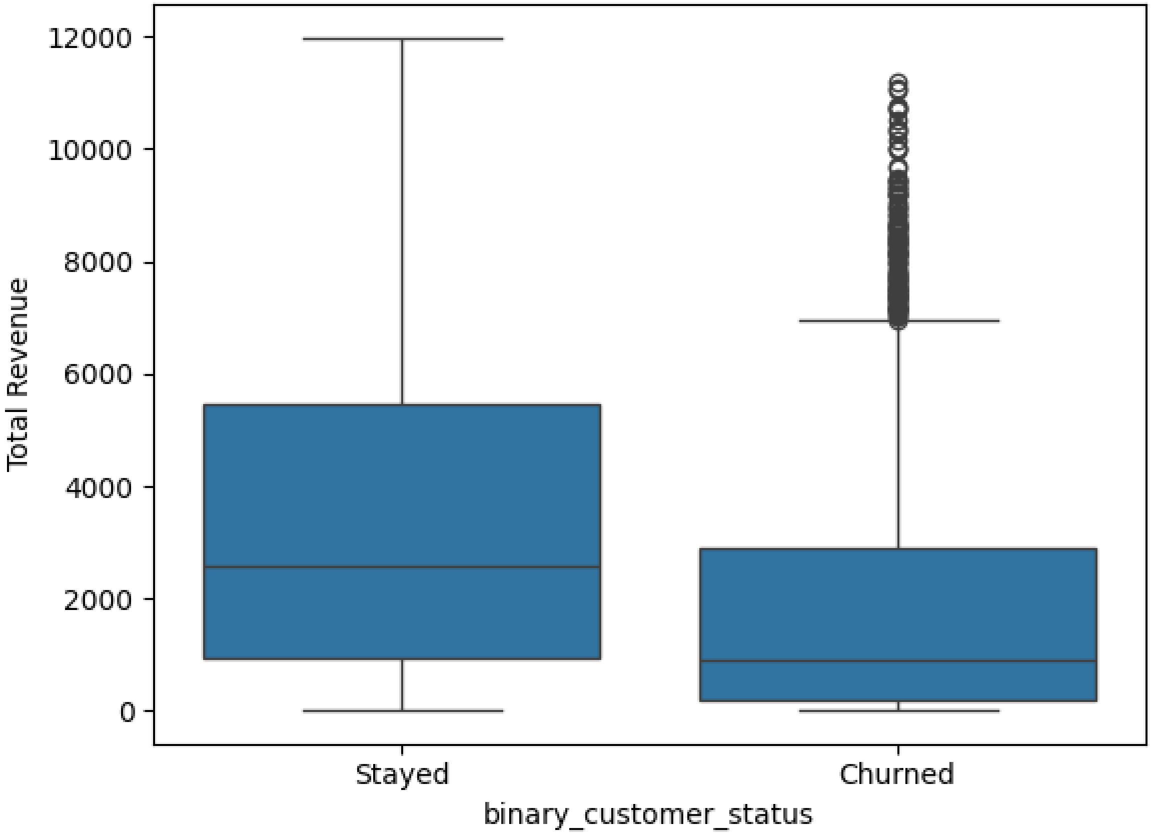
Total Long Distance Charges

The Total Long Distance Charges column shows some difference in the customers who churned versus those who stayed. It seems like the customers who stayed experienced more total long-distance charges.



Total Revenue

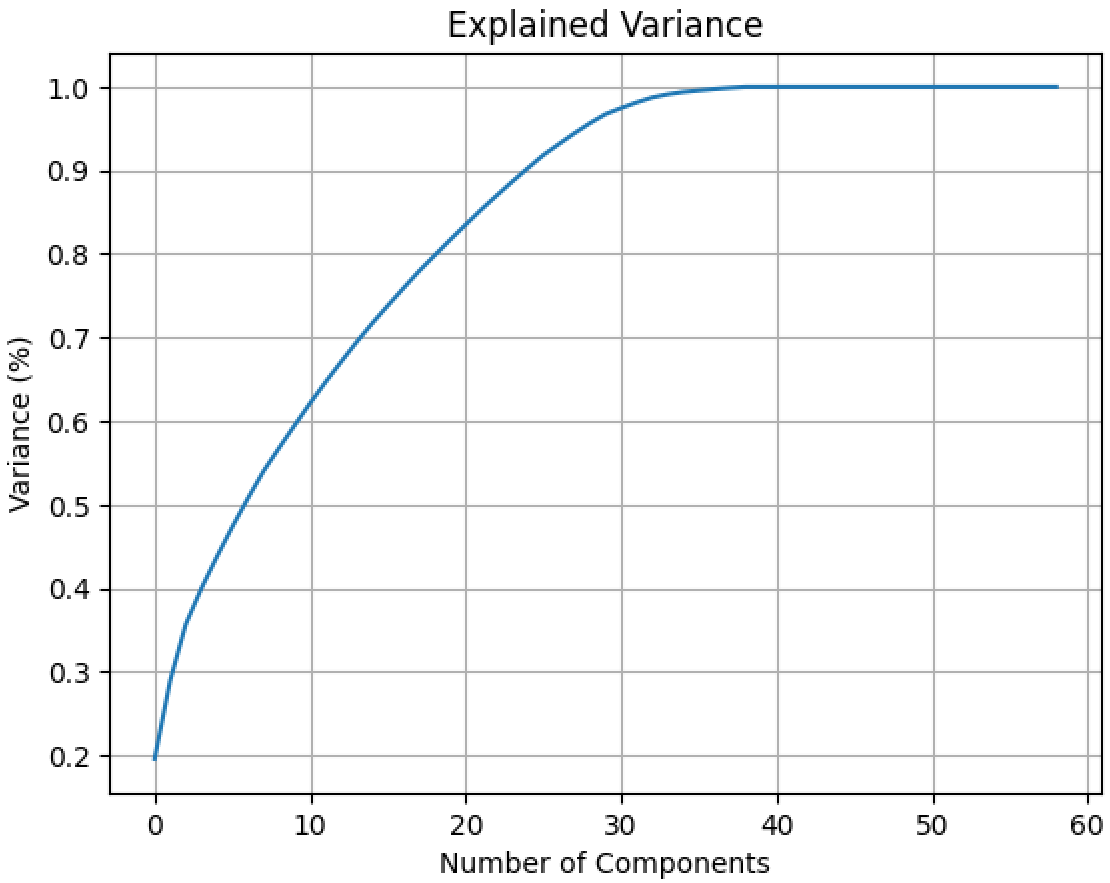
The Total Revenue column shows that customers who stay contribute a significant amount more in total revenue than customers who have churned. This correlates with tenure in that the customers who stay have been customers for longer.



Principle Components Analysis

Due to the high number of features included in the dataset, we used Principle Components Analysis (PCA) to reduce the number of dimensions while keeping as much variability as possible. PCA is a powerful statistical technique primarily used for dimensionality reduction in data analysis. By transforming a large set of variables into a smaller one that still contains most of the information in the large set, PCA helps to simplify the complexity in high-dimensional data while retaining the variability present in the dataset. This is achieved through identifying the directions, or 'principal components,' that maximize the variance of the data. These principal components are orthogonal to each other and represent the most significant underlying structure of the data. As a result, PCA is widely used in exploratory data analysis, predictive modeling, and the visualization of genetic distance and patterns in large-scale data sets, making it an indispensable tool in the fields of machine learning, pattern recognition, and data compression.

The plot below shows the change in explained variance as the number of principal components increases. It shows that the explained variance reaches the maximum at about 38 components. This indicates that our models should perform better using PCA to decrease the dimensionality of the dataset.



Exploratory Data Analysis Conclusion

Our exploratory data analysis sheds light on the complex dynamics of customer churn in the telecom industry, revealing that service experience and financial considerations play pivotal roles in influencing churn decisions. While demographics such as age and gender have minimal impact, the analysis distinctly highlights the sensitivity of customers to monthly charges. Elevated charges are strongly associated with increased churn, emphasizing the importance of competitive and flexible pricing strategies in customer retention efforts.

Furthermore, the correlation between total charges and lower churn rates underscores the significance of long-term customer relationships. This suggests that customers with longer tenures, who likely engage more with the service, exhibit lower churn tendencies. Therefore, telecom companies should focus on enhancing service quality and customer engagement to cultivate deeper, more enduring relationships with their customers.

By prioritizing competitive pricing and fostering long-term customer engagement, telecom companies can address the key drivers of churn identified in our analysis. Implementing targeted interventions in these areas promises to not only reduce churn rates but also enhance overall customer satisfaction and loyalty.

Modeling

The models we chose to predict which customers are likely to churn include Logistic Regression, Random Forest Classification, and Support Vector Machine. We chose a range of different models as our exploratory data analysis suggests that churn will be most accurately predicted by combining the effects of each of the columns. Running the data through different linear and nonlinear models increases the likelihood of finding the model that best fits the data.

Preprocessing

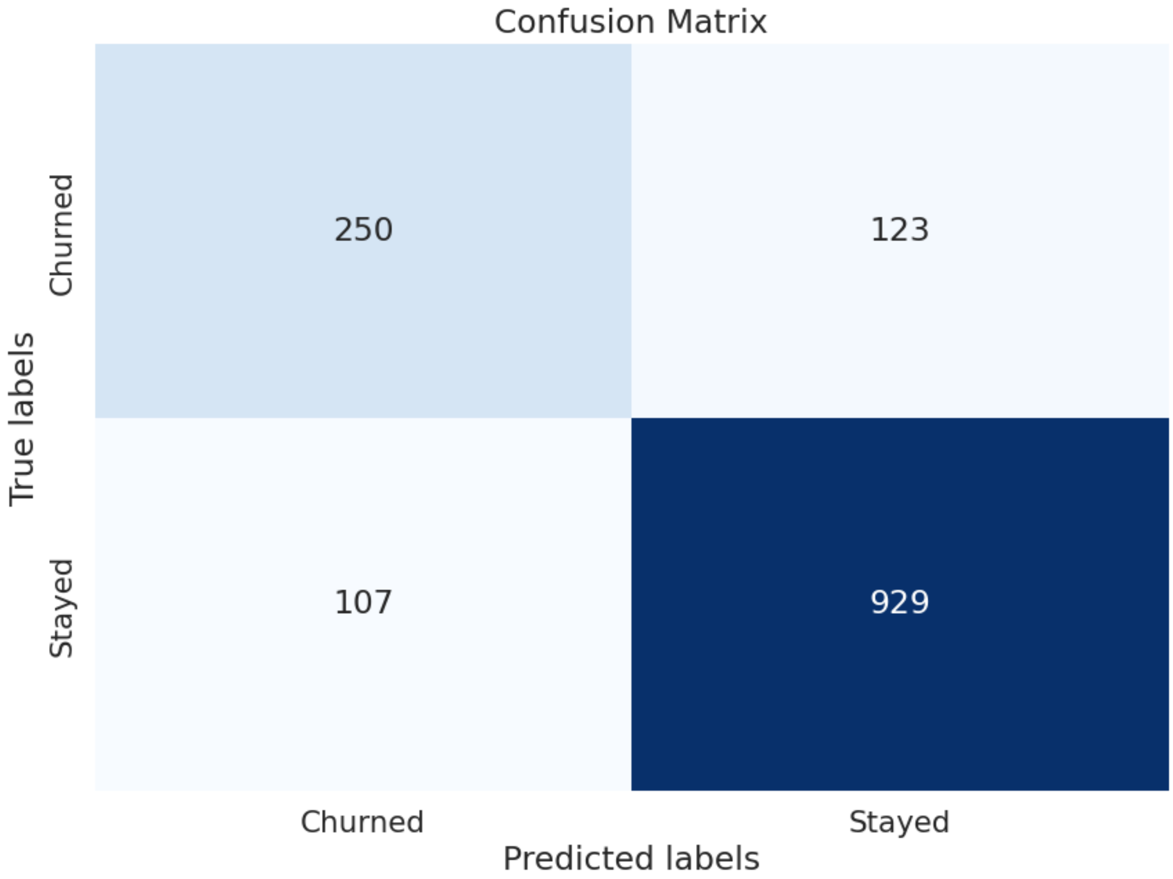
The modeling was completed by setting up a pipeline that included preprocessing the data and training the model. The pipeline was then run with different parameters for the model for tuning purposes. Preprocessing included one-hot encoding of the categorical features and scaling the data. Since the target variable is unevenly distributed, the models were also run with a down-sampled and an up-sampled version of the training data to see if a balanced dataset performed better.

Logistic Regression

Logistic Regression serves as a foundational model in the realm of customer churn analysis due to its simplicity and efficiency in handling binary classification problems. At its core, logistic regression estimates the probabilities of binary outcomes, making it an ideal candidate for predicting customer churn—where the outcome is whether a customer will leave or stay. Applying a logistic function to linear combinations of the input features provides a straightforward yet powerful means to model the relationship between customer attributes and their likelihood of churning. This model's interpretability is particularly beneficial, allowing analysts to understand the impact of various factors on customer retention decisions.

The results of the logistic regression models showed the best results were from the unbalanced dataset with an accuracy of 83.7%. The confusion matrix below shows the distribution of errors across the churned and stayed categories.

- Unbalanced Data: 83.7%
- Down-sampled Data: 78.1%
- Up-sampled Data: 78.9%

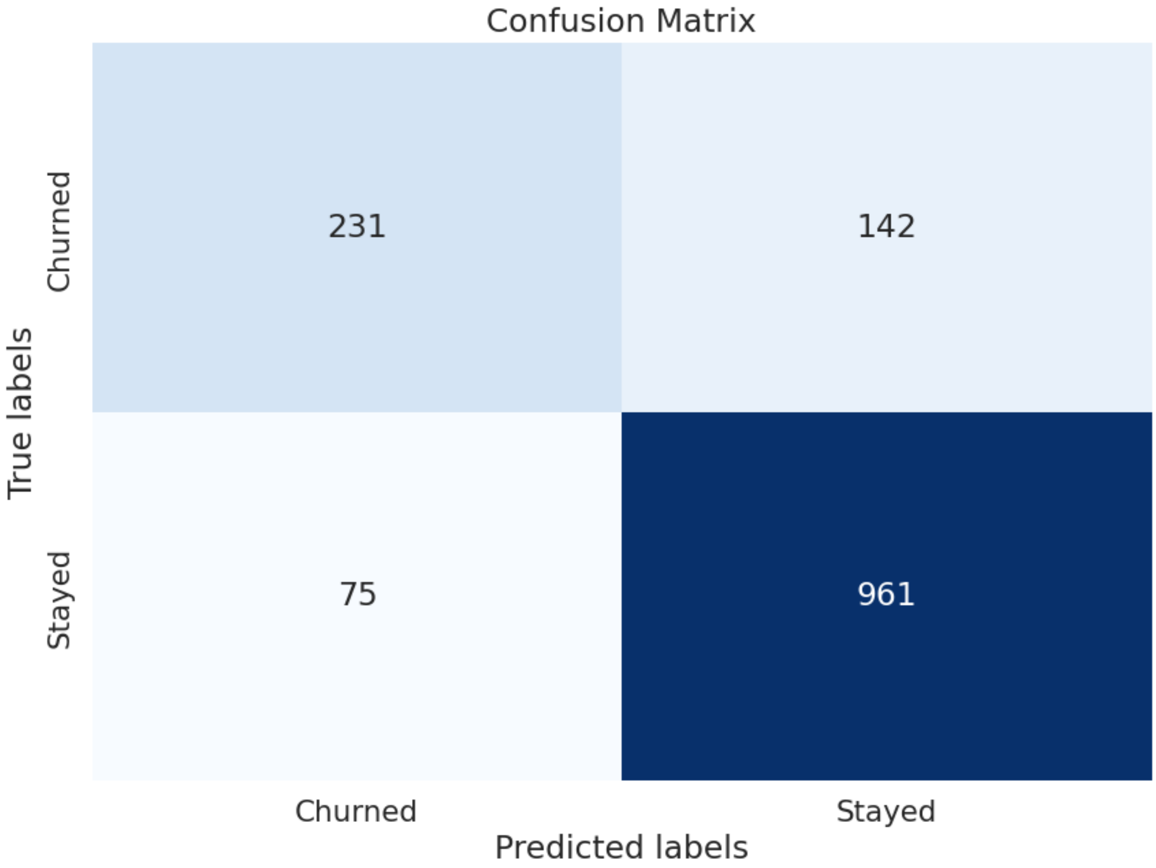


Random Forest Classification

The Random Forest Classifier stands out in customer churn analysis for its robustness and ability to handle complex, nonlinear relationships between customer characteristics and churn behavior. As an ensemble learning method, it operates by constructing a multitude of decision trees during training and outputting the class that is the mode of the classes of individual trees. This approach not only helps in capturing intricate patterns within the data but also in reducing overfitting, making it highly effective in improving prediction accuracy. Random Forest's capacity to deal with the high dimensionality of data and provide feature importance scores further aids in identifying key predictors of churn.

The results of the Random Forest Classifier show that the best results were obtained from the unbalanced dataset at 84.6%. The confusion matrix below shows the distribution of errors across the churned and stayed categories.

- Unbalanced Data: 84.6%
- Down-sampled Data: 79.2%
- Up-sampled Data: 81.5%

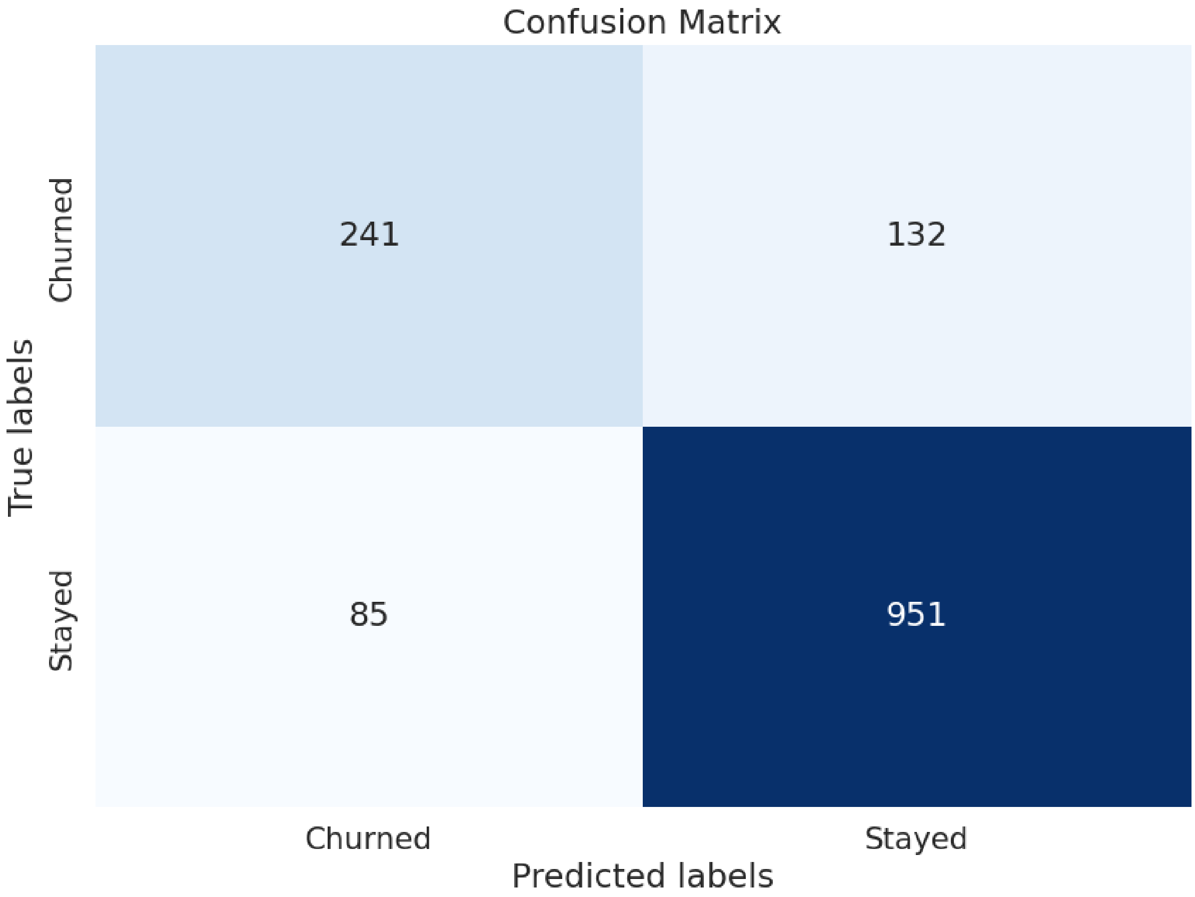


Support Vector Machine

Support Vector Machine (SVM) is a versatile and powerful classification technique used in customer churn analysis to find the hyperplane that best separates the data points of different classes. By maximizing the margin between the closest points of different classes (support vectors), SVM ensures a robust classification model. This method is particularly adept at handling high-dimensional spaces and performing well even when the number of dimensions exceeds the number of samples. Its flexibility, granted by different kernel functions, allows for tackling linear and nonlinear relationships alike, making SVM a formidable tool for predicting customer churn with high accuracy.

The results of the Support Vector Machine models showed the best results were from the unbalanced dataset at 84.6%. The confusion matrix below shows the distribution of errors across the churned and stayed categories.

- Unbalanced Data: 84.6%
- Down-sampled Data: 77.1%
- Up-sampled Data: 76.3%



Models Conclusion

The models used to predict customer churn included Logistic Regression, Random Forest Classifier, and Support Vector Machine. Not surprisingly Random Forest and Support Vector Machine performed the best with a test accuracy of 84.6%. While they shared the same test accuracy, Support Vector Machine was able to predict the customers who churned more accurately than Random Forest. It was able to correctly predict 241 of the 373 customers in the test dataset who churned in Q2 2022.

While the best accuracy scores were achieved with the unbalanced dataset, using the balanced dataset allowed the models to predict more of the churned customers correctly, but it also predicted more of the customers that stayed as churned. Using this type of model for the business would allow them to intervene on more customers who would churn, but it would also incur more costs on interventions for customers who were not going to churn. Depending on the needs of the business, the best model may be the most accurate one or the one that correctly predicts churned customers most frequently.

Conclusion

In the pursuit of understanding and mitigating customer churn within the telecommunications sector, this study embarked on a comprehensive journey through exploratory data analysis and predictive modeling. By dissecting the intricate layers of customer data, we illuminated the multifaceted nature of churn, revealing the pivotal role played by service experiences, pricing strategies, and customer engagement in influencing customer retention decisions.

Our exploratory data analysis meticulously sifted through demographic, service, and financial variables, unearthing critical insights into the behaviors and preferences of customers. Notably, the analysis debunked the significance of demographic factors such as age and gender in predicting churn, instead highlighting the paramount importance of service quality, pricing fairness, and the depth of customer engagement with the service provider.

The predictive models, encompassing Logistic Regression, Random Forest Classification, and Support Vector Machine, served as our analytical compass, guiding us to identify potential churners with commendable accuracy. The models' prowess in deciphering the complex tapestry of churn determinants further emphasized the necessity for telecom companies to adopt nuanced, data-driven strategies to preclude customer departure.

Conclusively, this study not only furnishes telecom companies with a blueprint for crafting targeted interventions aimed at curtailing churn but also underscores the indispensable value of leveraging big data analytics in sculpting customer retention strategies. By adopting a holistic approach that marries competitive pricing with exemplary service delivery and by nurturing long-term customer relationships, telecom providers can significantly diminish churn rates, thereby fostering a stable and loyal customer base.

In an era where customer loyalty is perpetually under siege, the insights gleaned from this analysis beckon telecom companies to reassess their operational, marketing, and customer service paradigms. Through the strategic application of predictive analytics, telecom providers are equipped to anticipate and address the evolving needs and concerns of their customers, ensuring a symbiotic relationship characterized by mutual growth and satisfaction. Ultimately, this study serves as a pragmatic guide for telecom companies striving to thrive in a competitive and dynamic market landscape.