## Introduction

Customer churn is a critical issue in the telecom industry, where retaining existing customers is often more cost-effective than acquiring new ones. Customer Churn phenomenon occurs when customers discontinue using a subscription or service, which poses a significant challenge for companies trying to hold onto/retain their clientele and maintain growth. The Telecom Customer Churn dataset from Kaggle provides a compelling and fascinating insight into the dynamics of customer retention for a telecommunications organization in California during Q2 2022. This dataset provides detailed records of customer demographics, account information, and service usage patterns enabling a comprehensive understanding of the factors contributing to churn.

With a total of 38 features and 7,043 customers, and two tables, the other table representing the unique Zip Code and Population, this data set offers an excellent foundation for building a robust churn prediction model.

The primary objective of this project is to perform data exploration, visualization & build a predictive model, using RapidMiner to identify customers who are likely to churn. By accurately predicting churn, we will not only uncover actionable insights into customer behavior but also the analysis will serve as a powerful tool to enhance retention strategies & enhance customer satisfaction making it an indispensable resource for business decision-making in the telecom sector.

## Background

## Inspired by *Akshat Mishra*'s analysis (2020), we reference his use of machine learning models to predict customer churn. In his work, he employed Decision Tree Classifier, achieving the best results with accuracy: 84% due to its ability to handle complex data patterns effectively. Our approach will build on this by exploring similar models while optimizing parameters to improve performance. [Link to the Published Article](https://www.kaggle.com/code/imakshatmishra/customer-churn-prediction)

## Assumptions & Constraints

The analysis of the Telecom Customer Churn dataset published by Maven Analytics is based on several key assumptions to ensure the accuracy and relevance of the findings. Firstly, it is assumed that the dataset is complete, with minimal missing or erroneous values that could impact the reliability of the analysis. And the present missing values are to be handled carefully. Each row is presumed to represent a unique customer with accurately recorded attributes. Secondly, "Customer Status" column is assumed to be correctly labeled, reflecting whether a customer has indeed terminated their service during the observation period or Stayed else Joined. This is an important factor for building predictive models, as mislabeling could lead to wrong conclusions about churn behavior. On the other hand, our data is just OF California city which acts as a major constraint in our analysis.

## Descriptive Analysis & Data Exploration.

Figure1 illustrates the descriptive analysis of all the columns present in our dataset. The dataset includes a balanced mix of numerical attributes (e.g., Tenure, Total revenue, monthly charges, total charges, etc.) and categorical variables (e.g., contract type, internet service type, and payment method), making it well-suited for predictive analytics. We can observe that there are missing values present in some of the columns which we will handle accordingly.

#### Customer Churn Table

* **Customer Status**: Indicates the status of the customer at the end of the quarter. [Chart 1] The pie chart gives a Great description of this column illustrating the fact that 4720 consumers, or the majority, have remained faithful suggests that customer retention is high. Still, 1869 clients have departed, which is a sizable percentage of the total. Furthermore, only 454 new consumers have signed up. The telecom firm must concentrate on decreasing customer attrition, which is our main analysis goal.
* **Gender**: The customer’s gender either male (3555) or female (3488). [Chart.2] Indicates gender-based distribution of customer status representing "Stayed" as the most common category for both sexes, with slightly higher figures for men (2,382) than for women (2,338). There is no gender difference in churn rates, since the "Churned" group shows an equal distribution of 939 females and 930 males & for both sexes, the "Joined" category is the least prevalent. Hence it can be said to be balanced overall, showcasing no individual trends.
* **Age**: The customer’s current age, in years ranging from 19 years to 80 years. [Chart 3] represents the age Distribution of Customers and their status. Looking at the number of customers in each age group suggests a stable customer base across different ages. Comparing the age groups, the higher number of new joiners and churned found in young adults than in old age with a little stable stayed count across young and middle-aged customers.
* **Married:** if the customer is married. Indicating majority of customers are unmarried.
* **Number of Dependents**: Indicates the number of dependents that live with the customer. Minimum being 0 & Maximum 9 dependents. [Chart.4] The graph depicts the relationship between customer status and Number of Dependents indicating that there are very few customers with more than 3 dependents. Customers with 0 dependents have the highest number of Stayed, churned & joined indicating a significant portion of Customer base in single which can be backed up by above descriptive analysis of married attribute as most of the customers are unmarried.
* **Zip Code:** The zip code of the customer’s primary residence. “*This column will be converted to nominal as it doesn't have any integral significance.”* [Chart.5] illustrates the geographical representation of customers in our data based on their latitude, longitude and Zip Code.
* **Number of Referrals:** Indicates the number of times the customer has referred someone to this company ranging from 0 up to 11 referrals made.
* **Tenure in Months**: Indicates the total number of months that the customer has been with the company till today. Minimum is 1 and maximum of 72 months. The box plot [Chart 6] visualizes the tenure distribution (in months) across different customer statuses. It can be observed that Customers with longer tenures are more likely to stay, suggesting that loyalty builds over time. High churn rates in the early months indicate potential. onboarding or service-related issues that need to be addressed to improve retention new customers often have a tenure of less than two months.
* **Phone Service, Multiple Lines:** Indicates if the customer subscribed to home phone service & subscribes to multiple telephone lines with the company respectively (Yes, No). Chart.7 shows most customers with phone service stayed with the service (green bar), with over 4,000 customers & a huge portions churned (red bar). Similarly with no Phone service very few stayed as most of the customers opted for this service.

*“Observing the data, the nulls are present for Multiple Lines only when the phone service value is NO. 682 missing values observed which will later be converted to “No”. As if there is no phone service the nulls in multiple lines can be replaced with “No””.*

* **Avg Monthly Long-Distance Charges:** Indicates the customer’s average long-distance charges. 49.99 is the maximum charge and 1.101 is the minimum.

*“Observing the data, the nulls are present for this column only when the phone service value is NO. 682 missing values, which will be converted to 0 as the customer is not subscribed to home phone service the Avg Monthly Long-Distance Charges will be 0”.*

* **Internet Service, Internet Type:** Indicates if the customer subscribes to Internet service with the company (Yes, No). If yes then **Internet Type** indicates type of internet connection (DSL, Fiber Optic, Cable).
* **Contract:** Indicates the customer’s current contract type with Month-to-Month opted by 3610 customers Chart.8 represents Churn rate across different contract type which can be interpreted as Customers with yearly and two-yearly contract type are likely to be stayed and monthly contract type customers mostly churn as compared to other categories.
* **Payment Method**: Indicates how the customer pays their bill: Bank Withdrawal being the most common method for 3909 customers and then Credit Card, Mailed Check. [Chart.9] The bar chart depicts customer churn based on the payment method. The Credit Card method has a strong retention rate, suggesting it’s a preferred and convenient choice for customers. Despite being popular, the Bank Withdrawal method sees a notable number of churned customers, indicating potential dissatisfaction or issues with this option. Mailed checks are the least used method.

*“Observing the data, the nulls are present for* ***Internet Type, Avg Monthly GB Download, Online Security, Online Backup, Device Protection Plan, Premium Tech Support, Streaming TV, Streaming Movies, Streaming Music, Unlimited Data*** *columns only when the I****nternet service*** *value is NO. Hence 1526 missing values in* ***Internet type****, will be converted to “No Internet” as the customer is not subscribed internet service, the Internet types will be No Internet. And 1526 missing values in all other columns will be converted to “NO” as the customer is not subscribed internet service & these all columns are connected to internet service like streaming movies is done with internet so on and so forth”.*

* **Monthly Charge**: Indicates the customer’s current total monthly charge for all their services from the company. The scatter plot [Chart 10] shows the relationship between monthly charges, tenure (in months), and customer status. Most long-tenure customers (above 30 months) tend to stay regardless of their monthly charges. Churn is more prevalent among customers with higher monthly charges (above $70) and shorter tenures (under 20 months). The "Joined" category is concentrated near the beginning of the tenure axis, which makes sense as they are new customers.
* **Total Charges**: Indicates the customer’s total charges. The scatter plot [Chart 11] shows the relationship between Total charges, tenure (in months), and customer status Customers who have stayed tend to cluster toward higher tenure (above 40 months) and higher total charges (above $2000). Churned customers are concentrated in the low tenure range (0–20 months) and low to moderate total charges (under $2000). **Low Tenure = High Churn and vice versa.**
* **Total Revenue**: Indicates the company's total revenue from this customer (Total Charges - Total Refunds + Total Extra Data Charges + Total Lond Distance Charges). [Chart.12] The box plot illustrates the distribution of total revenue generated by customers, categorized by their status. Stayed Customers generate the highest and most variable revenue, indicating that loyal customers contribute significantly to the company’s revenue. Although churned customers generate less revenue on average than stayed customers, they still contribute a substantial amount.
* **Churn Category & Churn Reason:** A high-level category for the customer’s reason for churning, which is asked when they leave the company: Attitude, Competitor, Dissatisfaction, Other, Price (directly related to Churn Reason) being a customer’s specific reason for leaving the company. [Chart.13 ] represents the Prominent reason and categories which lead the customer to churn.

*“Observing the data we can say that the missing values in* ***Churned Category*** *and* ***Churn Reason*** *are only present for customer that are churned and hence the 5174 missing values here can be replaced with* ***Not Churned*** *indicating that all the missing values in these columns are customers who are not churned (stayed + joined)”*

#### Zip Code Population Table

* **Zip Code & Population**: A current population estimate for the entire Zip Code area.

## Data Exploration and Key insights

* Higher Churn in Middle Age whereas younger individuals are more likely to explore and switch services, contributing to higher join and churn rates on the other hand older customers might have more loyalty or fewer alternatives, resulting in fewer changes in status.
* Customers without dependents might have more flexibility to switch services, leading to a higher churn rate. The churn rate decreases as the number of dependents increases. & Customers with 1-3 dependents may seek stability, leading to higher retention.
* Loyal customers are likely to have longer tenure and higher engagement, leading to higher revenue signifying high revenue from these, Moderate Revenue from Churned Customers. & Low Revenue from New Joiners.
* Both Credit Card and Bank Withdrawal are widely used, but Bank Withdrawal has a disproportionately high churn rate, suggesting room for improvement in customer satisfaction or service quality for this payment method.
* Tenure appears to be a critical factor for customer retention. The longer a customer stays, the more likely they are to continue staying. Customers with tenure under 20 months are at a higher risk of churning.
* Higher total charges often correlate with longer tenure and staying customers, suggesting these customers are more engaged with the service.
* Phone service plays a critical role in customer retention, as most customers with phone service stayed.

## Model Iteration

We evaluated three machine learning algorithms for predicting customer churn: Decision Tree, Random Forest, and Gradient Boosted trees. These models were chosen based on their effectiveness in classification tasks and their prevalence in churn prediction studies.

1. **Model Chosen:**

**Decision Tree**: We Choose with a Decision Tree classifier due to its simplicity and interpretability and comparing the ROC of all these models we can easily interpret that decision is the best though there is little to no difference. The initial model achieved an accuracy of 83.99%. with maximal depth of 7 and size split of 20. And cross validation was set to 8.

1. **Iterations and Adjustments**
2. Looking at the accuracy we can change the iterations for increasing the accuracy even by 1 % that would be considered as a prominent improvement. For the second Iteration we took the maximal depth as 9 and increased the size split as 35 and similarly increased the minimal leaf size to 18.

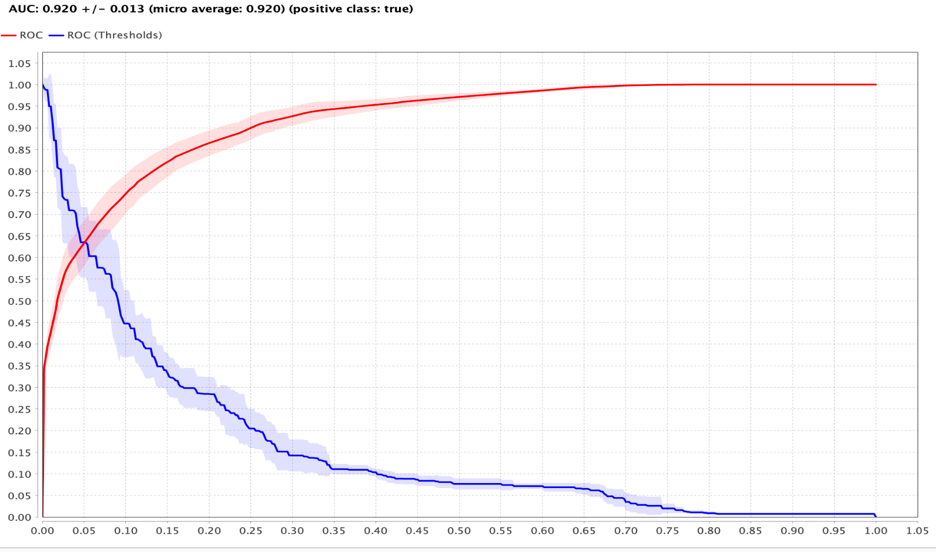
By limiting depth, ensuring larger splits, and requiring a minimum number of examples in leaves, the tree becomes simpler and focuses on broader trends in the data. And increasing the outer cross validation to 10. And now we will be keeping it constant for the rest of the iterations.  
Working with these parameters we got accuracy of about 84.60 % hence we are going in the right direction with the increasing values of parameters.

With one more similar increasing parameter value iteration we got the accuracy of 85.75 %

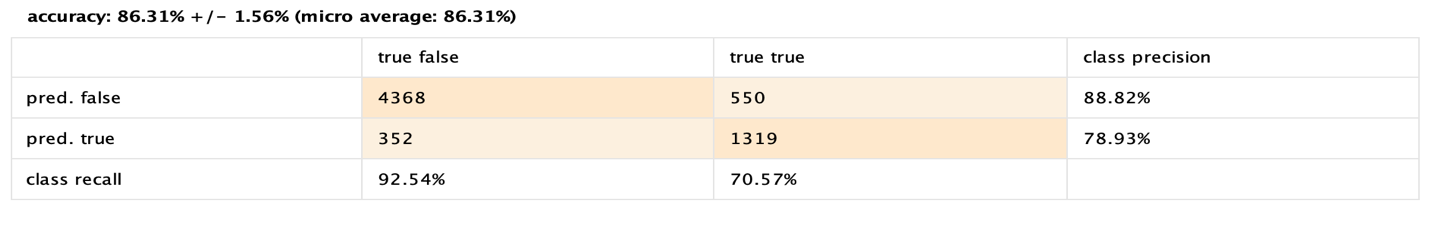
1. As our accuracy increased by increasing the parameters we will be doing the same to increase the accuracy to some more extent. Maximal depth 15, minimal leaf size 20 and minimal size of split 45. The accuracy kept on increasing to about 86.31% indicating a significant increase and specifying that these parameters play a vital role in the model.
2. For the third iteration maximal depth 20, minimal leaf size 20 and minimal size of split 50 but it didn't increase the accuracy of our model hence indicating the model reached its optimal capacity. Keeping it as our final model iteration.

## Model Evaluation & Analysis

After comparing the ROC of all three models, we selected Decision tree as the final one with the last iteration parameters, which shows the highest performance of the model.



The above chart specifies the ROC & AUC. The AUC value displayed at the top-left of the graph (0.920) quantifies the overall performance of the model. A perfect model has a performance of 1.0 indicating 0.920 an excellent model based on performance, meaning it is excellent at distinguishing between positive and negative classes. The curve closely approaches the top-left corner, indicating strong performance. We can also depict that increasing the TPR (true positives) may also increase the FPR (false positives), depending on the threshold. This graph allows you to choose a threshold that balances the tradeoff based on your specific application.

This confusion matrix evaluates a model predicting whether customers have churned (left the service) or stayed (remained loyal). Here, "True" corresponds to churned customers, and "False" corresponds to customers who stayed. The model achieved an accuracy of 86.31% ± 1.56%. Among the predictions, 550 churned customers were mistakenly identified as stayed (False Negatives), while 1319 churned customers were correctly identified as churned (True Positives). On the other hand, 4368 stayed customers were correctly predicted to have stayed (True Negatives), while 352 stayed customers were wrongly classified as churned (False Positives). The model's precision for predicting churn (True) is 78.93%, meaning when it predicts a customer churned, it is correct about 79% of the time. Its recall for churn is 70.57%, meaning it captures about 71% of the actual churned customers. Similarly, for customers who stayed (False), the model has a high precision of 88.82% and a recall of 92.54%, indicating strong performance in identifying non-churners

## Available Literature Analysis

Compairing the analysis published with ours here are some key points:

**Accuracy:** 86.31% ± 1.56%

**Precision (Churned Customers):** 78.93%

**Recall (Churned Customers):** 70.57%

**True Positives (Churned Correctly Identified):** 1,319

**False Positives (Stayed Incorrectly Identified as Churned):** 352

**False Negatives (Churned Incorrectly Identified as Stayed):** 550

**True Negatives (Stayed Correctly Identified):** 4,368

**Kaggle Model (Customer Churn Prediction):**

**Accuracy:** Approximately 84%

**Precision (Churned Customers):** Approximately 86%

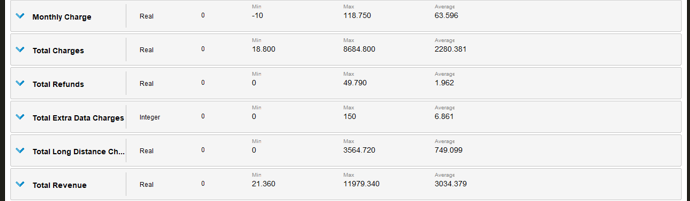
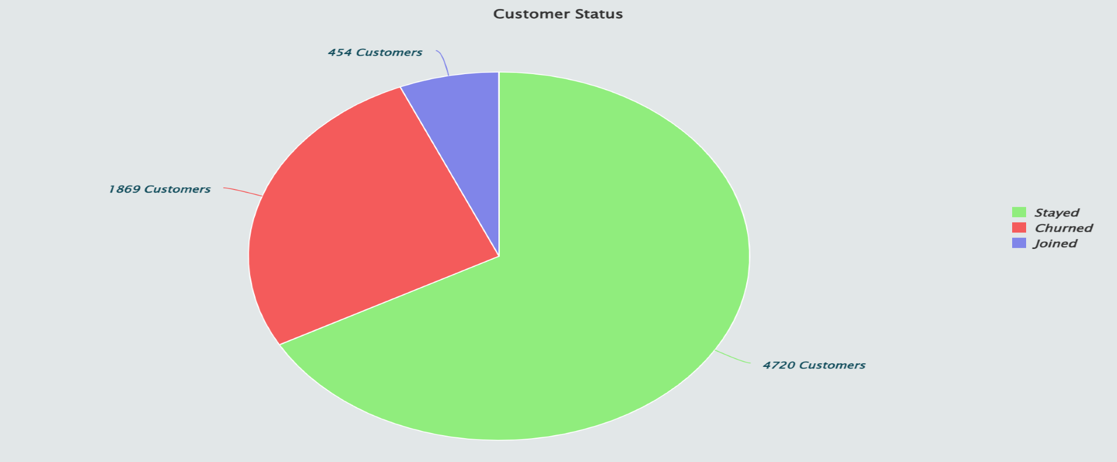
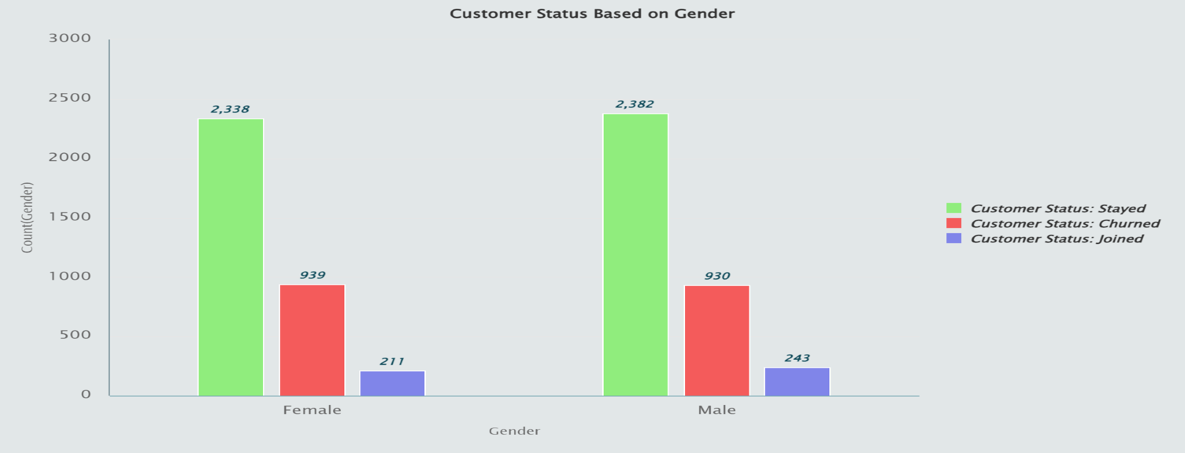
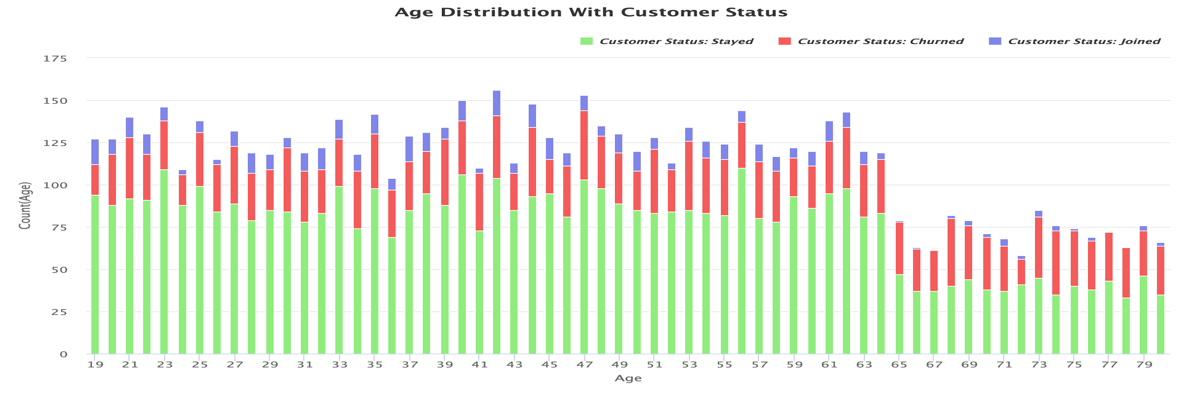
**Recall (Churned Customers):** Approximately 92%

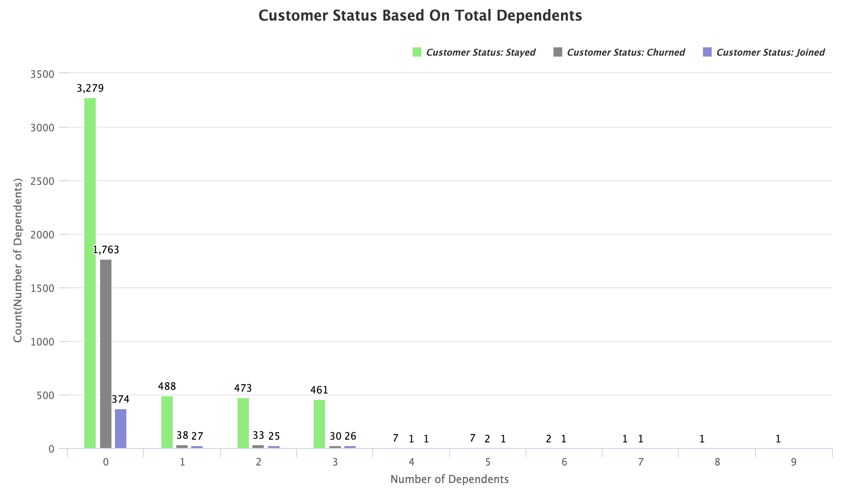
**Accuracy:** The model generated by us exhibits higher accuracy (86.31%) compared to the Kaggle model's approximate 84%, indicating a greater overall correctness in predictions.

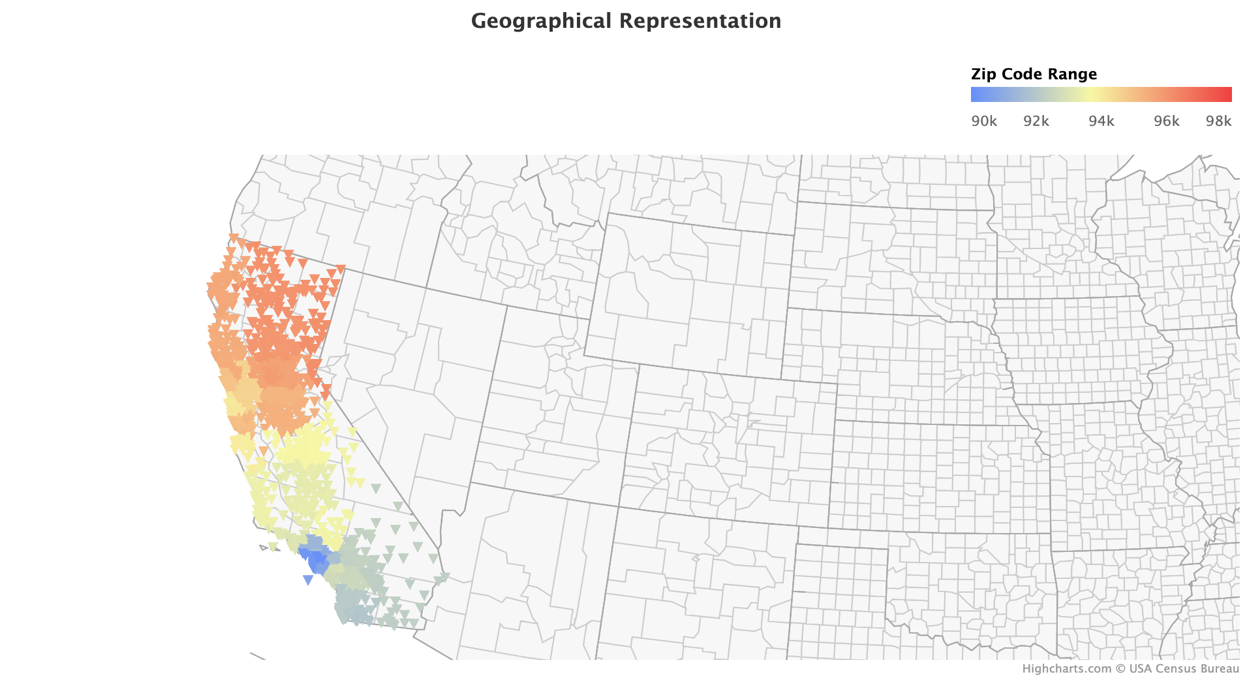
**Precision:** With a precision of 78.93%, our model is more reliable in correctly identifying churned customers than the Kaggle model, which has a precision of around 70%.

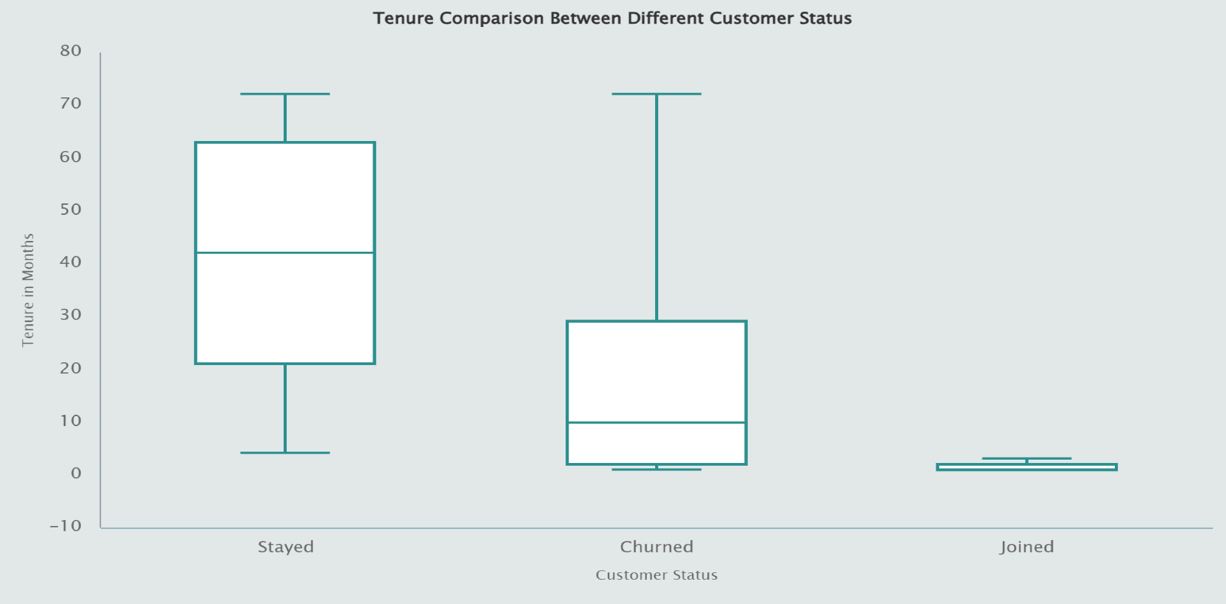
**Recall:** On the other hand, kaggle model also outperforms in recall (92% vs. 70.57%), suggesting it is more effective in capturing actual churned customers.

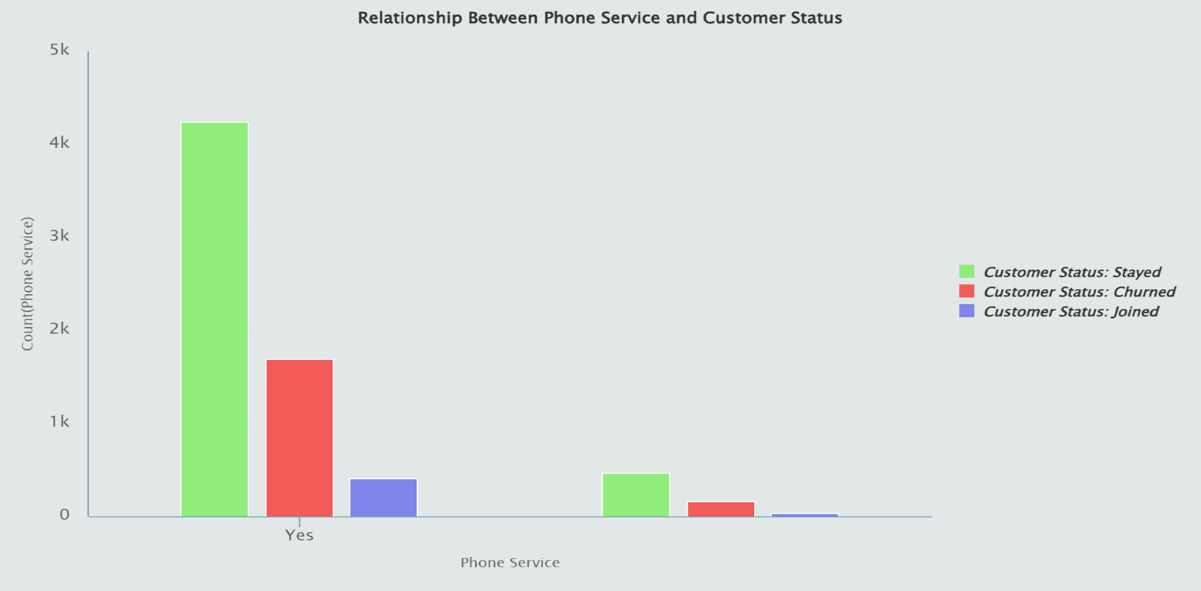
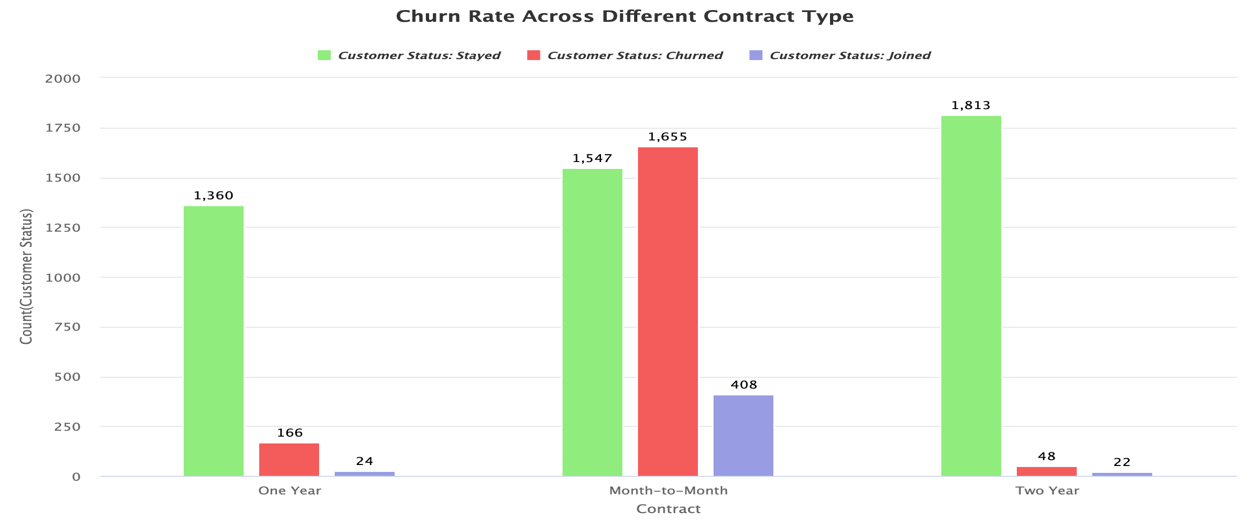
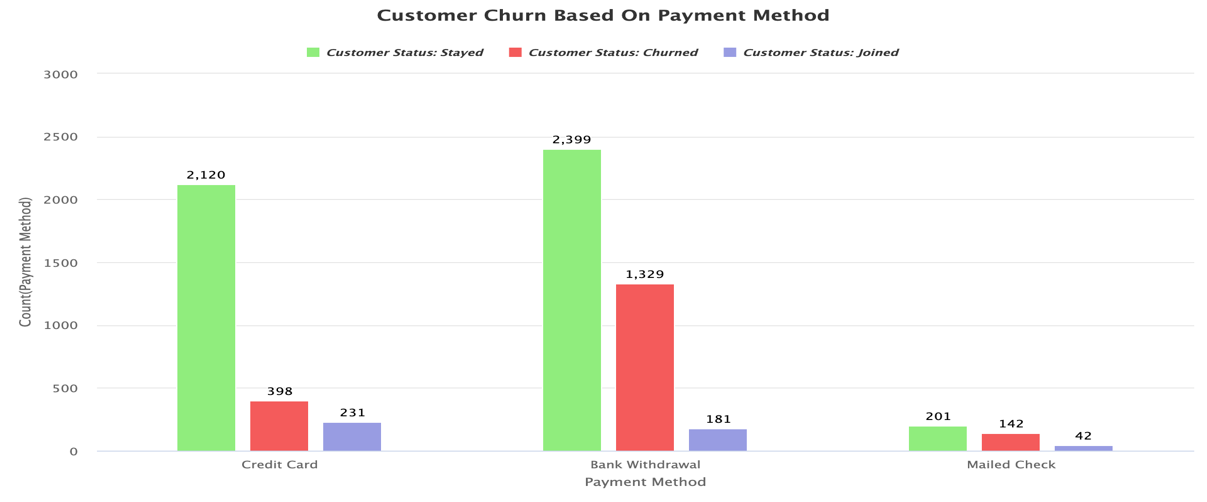
## Appendix - [Link to the Published Article](https://www.kaggle.com/code/imakshatmishra/customer-churn-prediction)

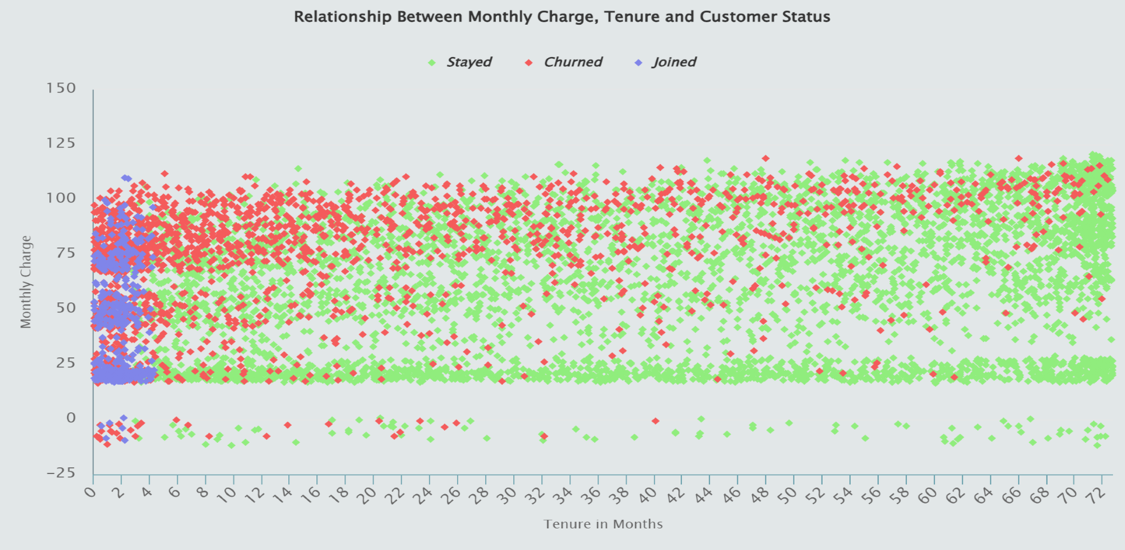
 Figure 1 [Chart 1] [Chart 2][Chart 3]

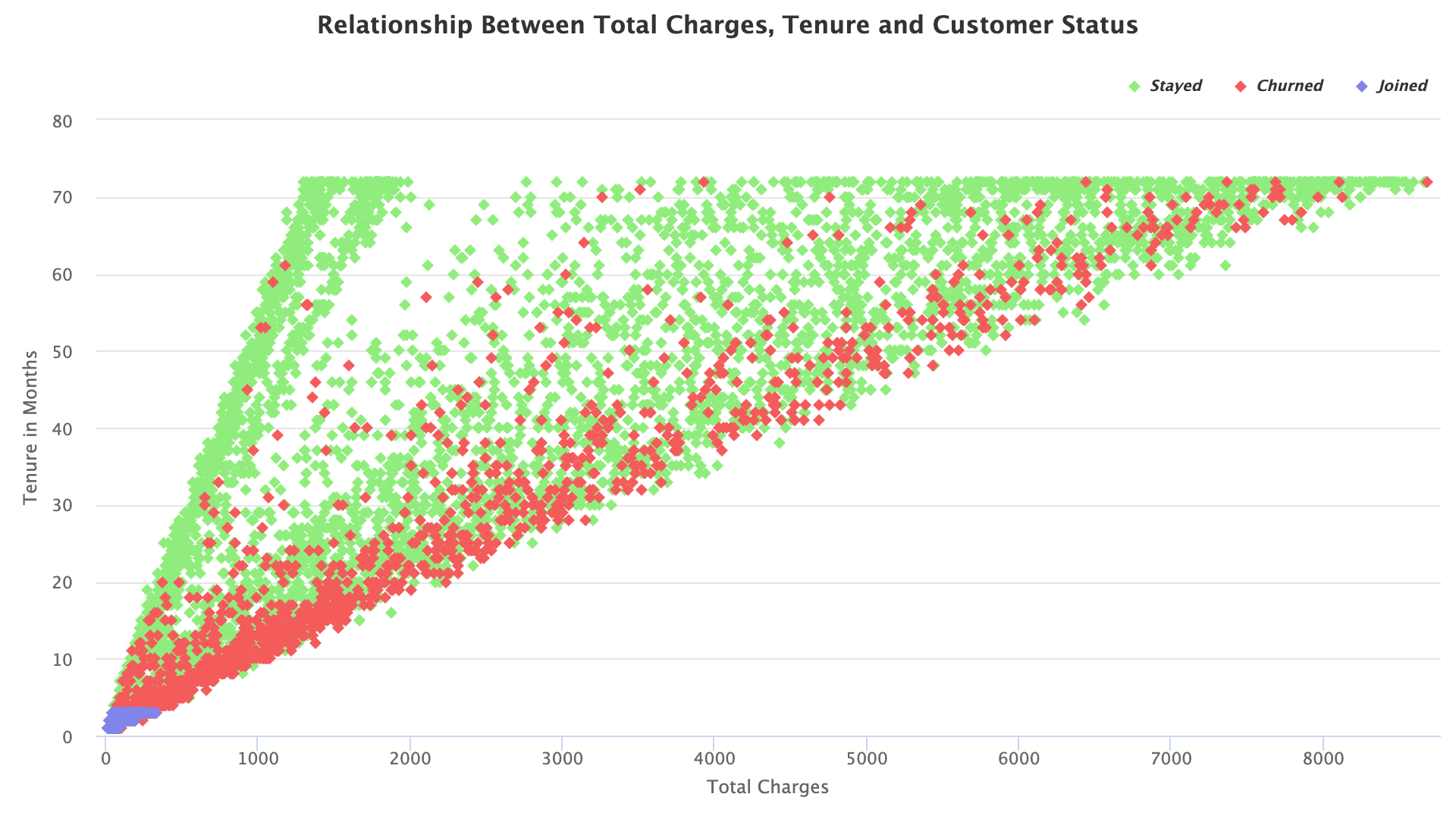
[Chart 4]

 [Chart 5]

 [Chart 6]

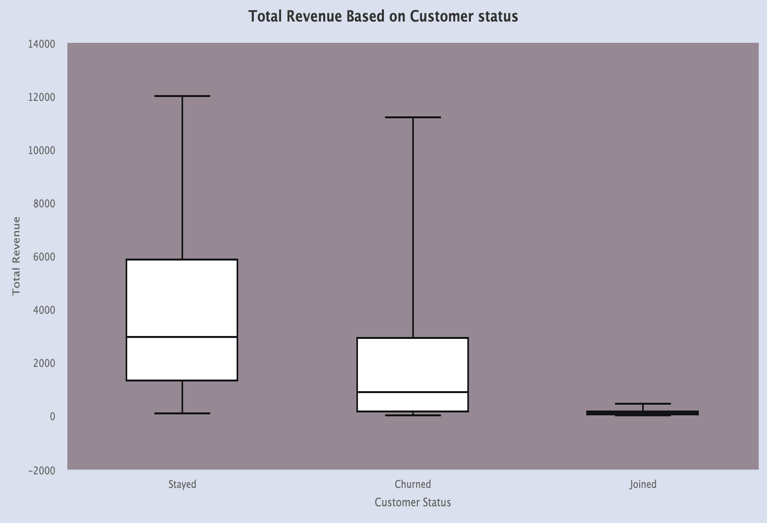
 [Chart 7][Chart 8][Chart 9]

 [Chart 10]

 A close-up of a black background

Description automatically generated

[Chart 11] [Chart 13]

 [Chart 12