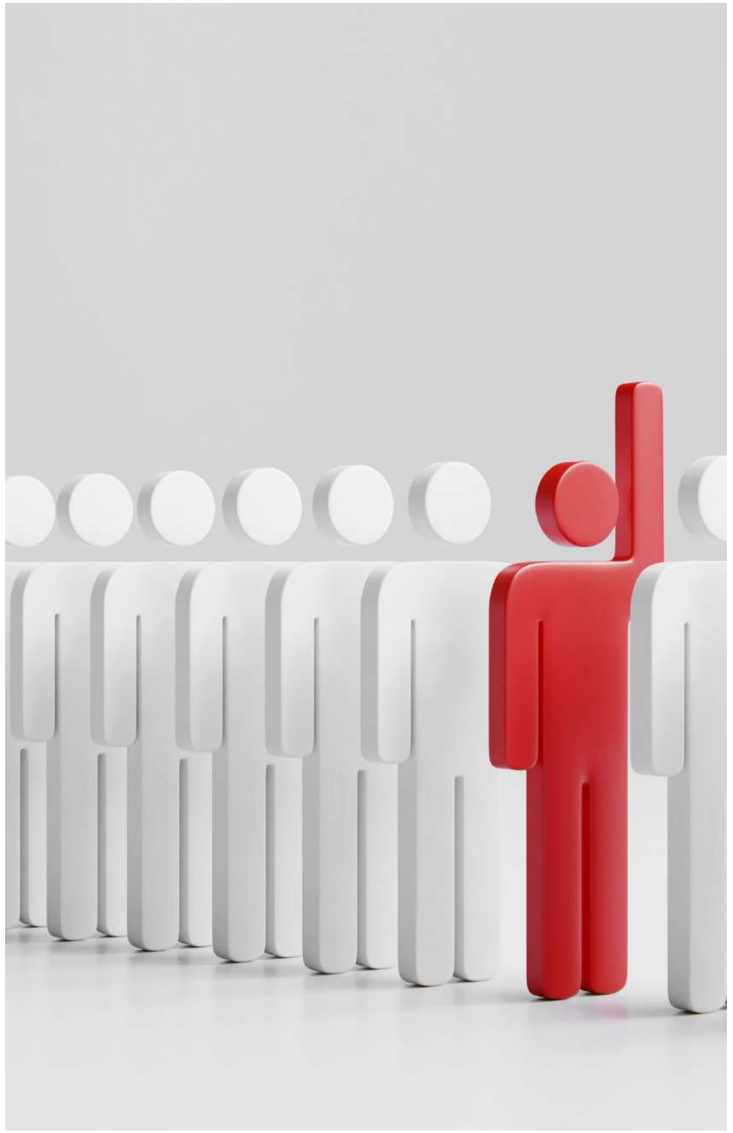


Predicting Customer Churn in an Iranian Telecom Company

A Project by Konstantinos Soufleros

Introduction

In the telecommunications industry, customer churn, or the rate at which customers leave a service, is a crucial metric that impacts the company's revenue and growth. Understanding the factors that lead to customer churn and predicting potential churners can help the company take proactive measures to retain valuable customers.

Objective

The main objective of this analysis is to build a predictive model using logistic regression to identify customers who are likely to churn based on historical data. By predicting potential churners, the telecom company can implement targeted retention strategies, offer personalized promotions, or address customer concerns, ultimately reducing churn rates and enhancing customer loyalty.

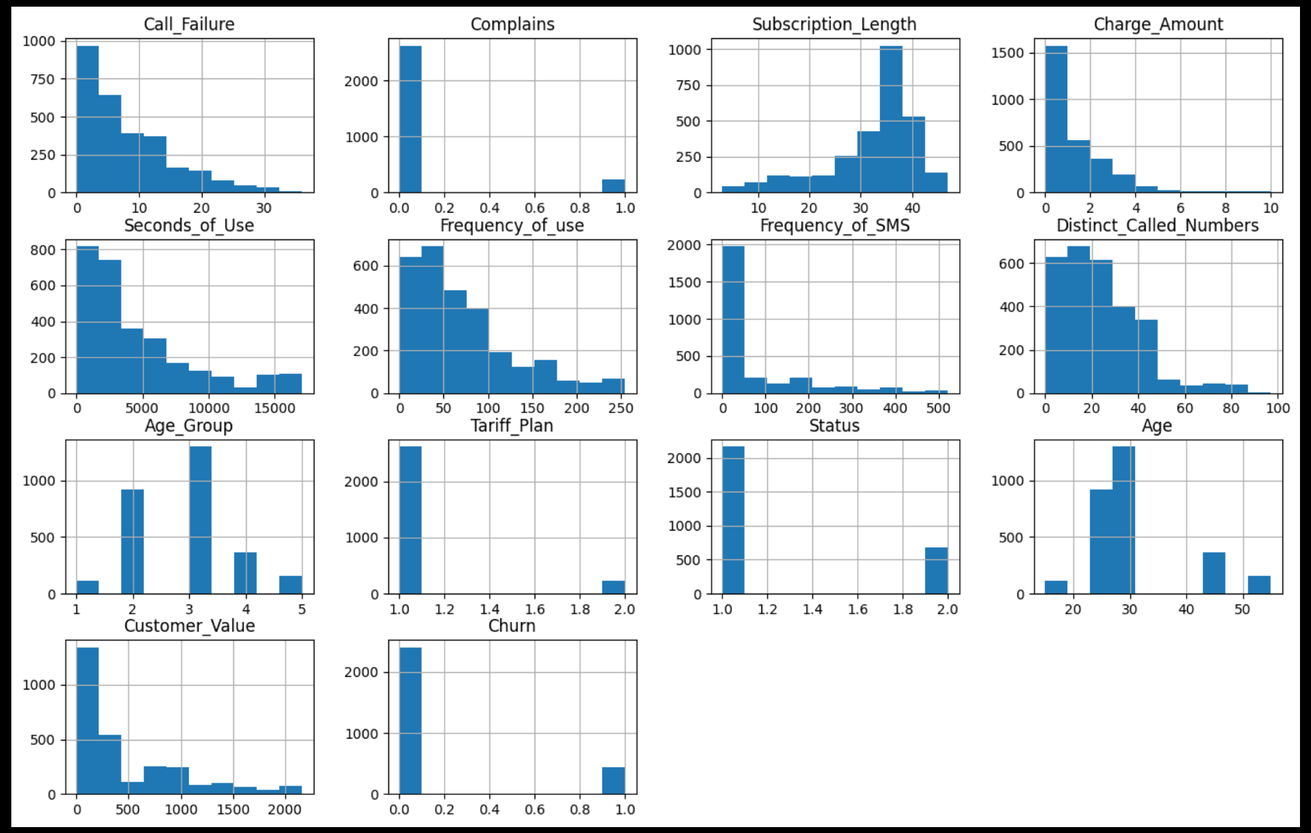
"Can we build an accurate predictive model based on historical data using logistic regression to identify customers who are likely to churn in an Iranian telecom company?"

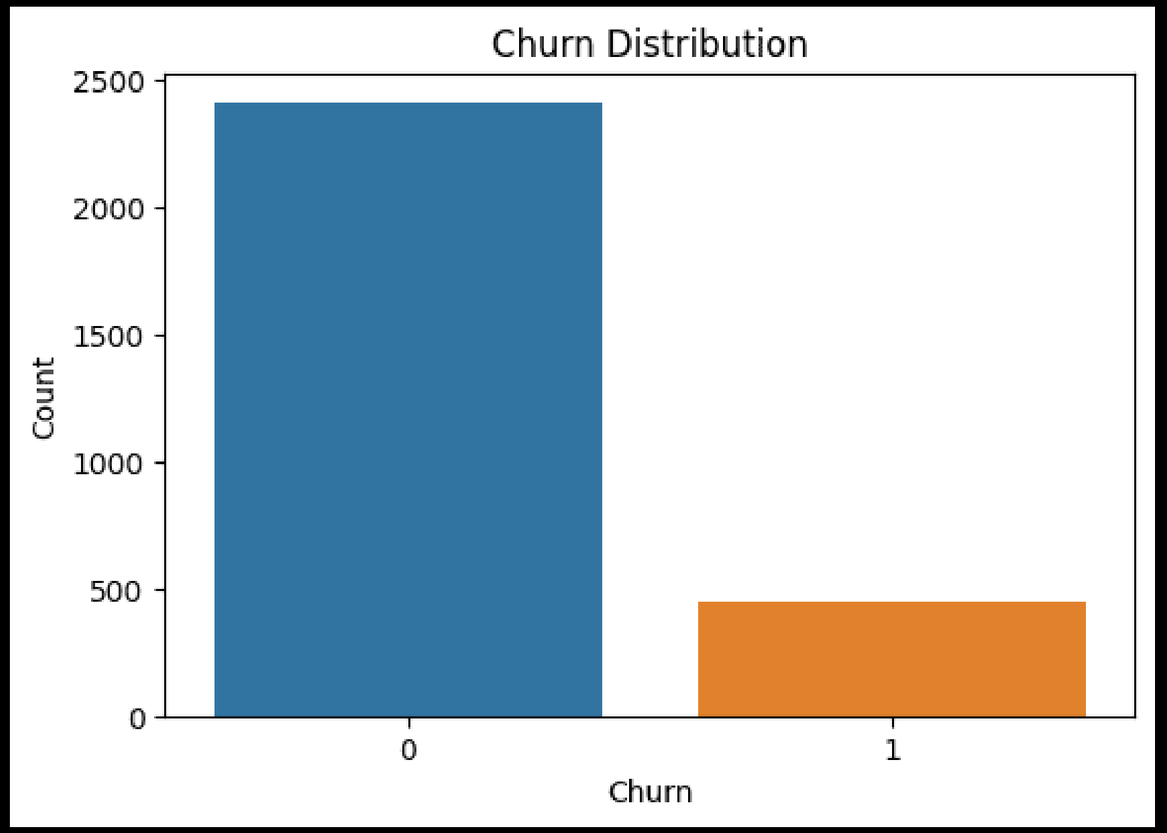
Project Question

Methodology

A systematic methodology was used to predict customer churn. This involved data exploration and preprocessing, feature analysis, and selection of a logistic regression model for prediction.
The model's performance was evaluated using various metrics and interpreted through feature coefficients and SHAP values. Sensitivity analysis was performed to assess model robustness.
The approach utilized data visualization, statistical analysis, and machine learning techniques to offer insights and recommendations for optimizing customer retention strategies and business outcomes.

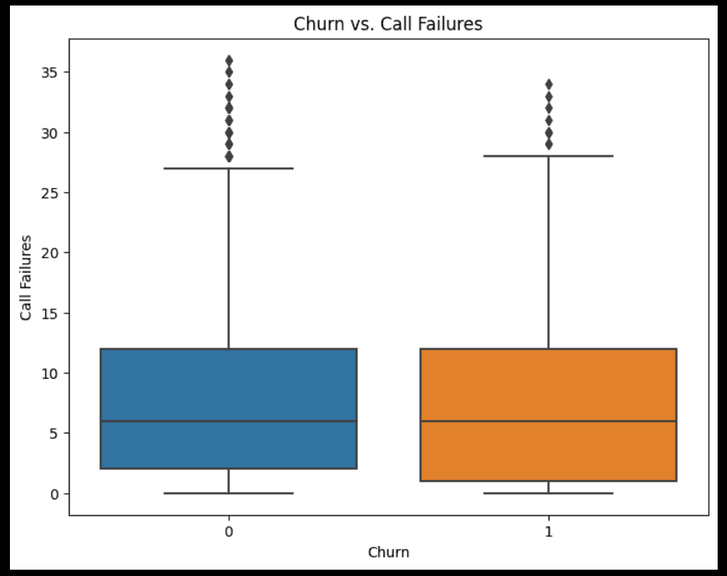
Exploratory Data Analysis

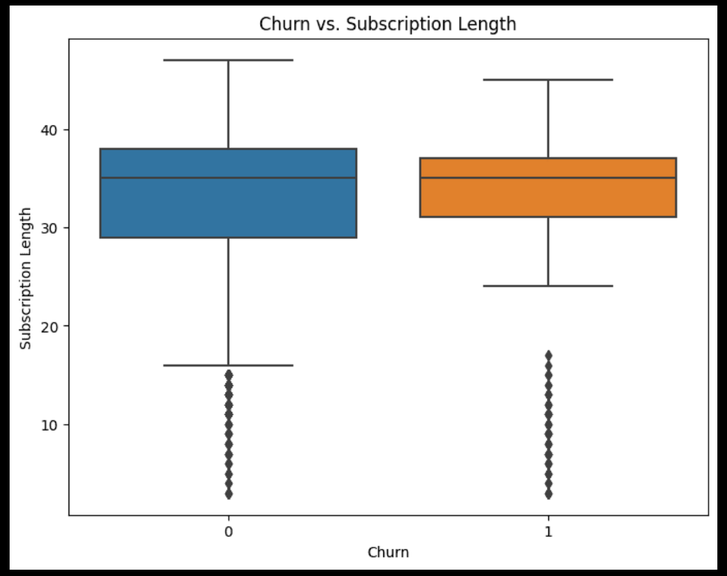


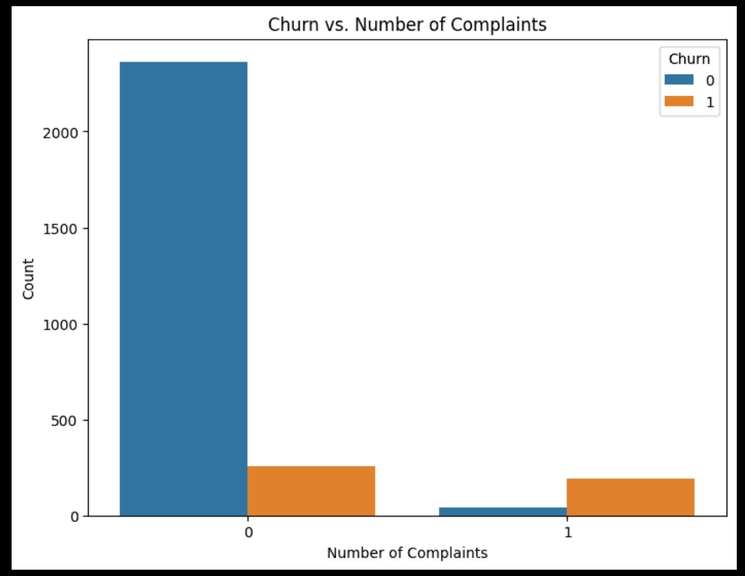


Distributions of the independent variables and the target variable "Churn". There is imbalance between churned(1) and non-churned(0) customers.

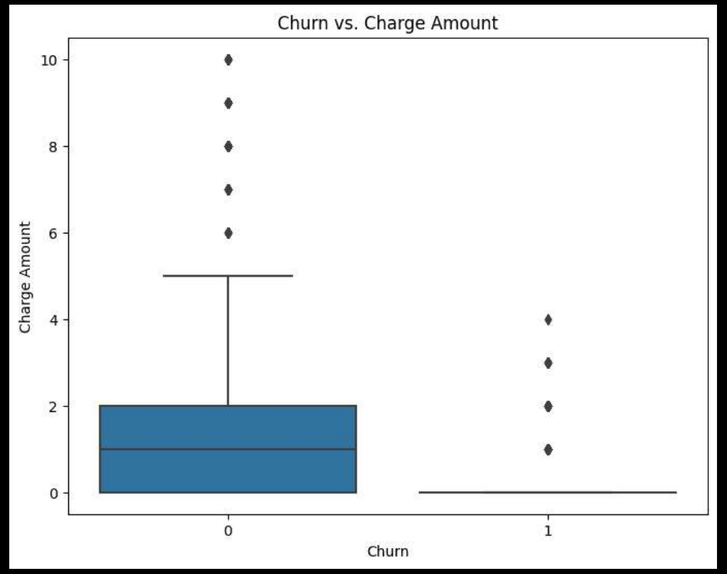
Exploratory Data Analysis

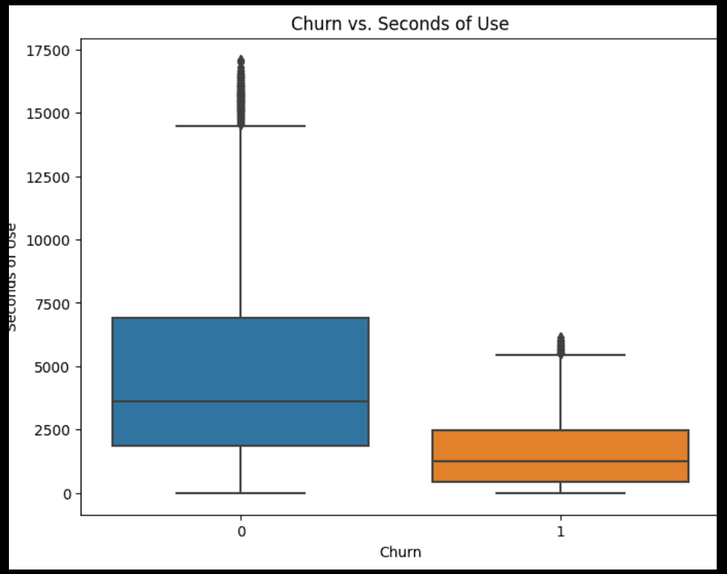


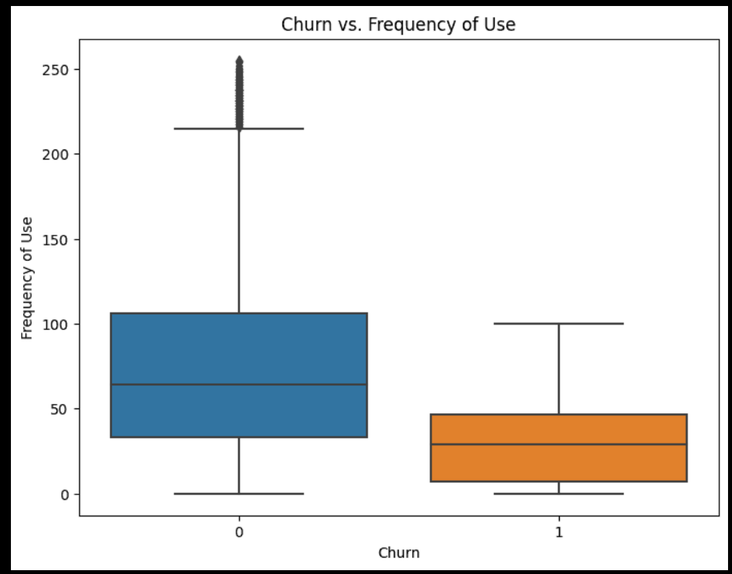




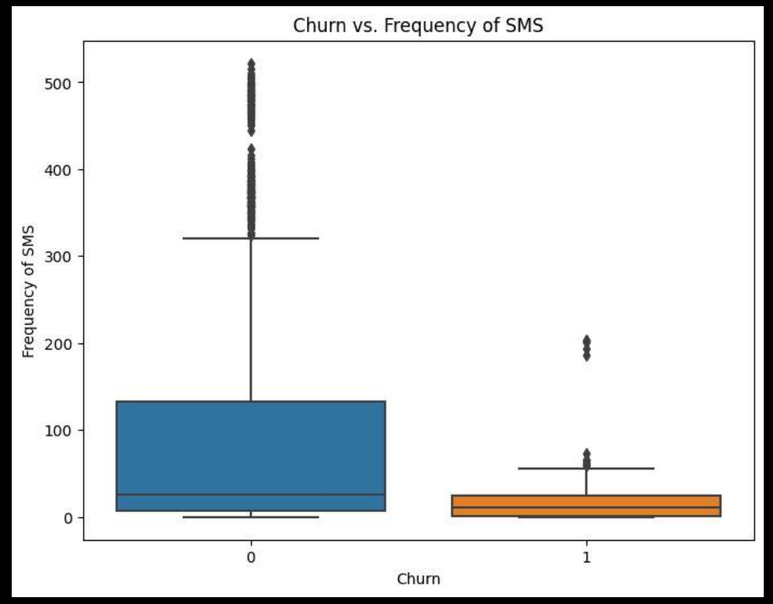
Churn vs. Rest

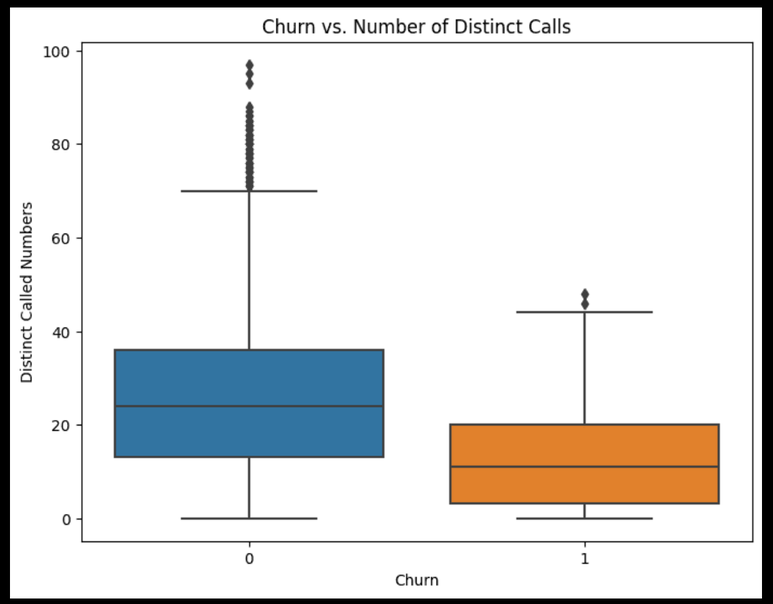


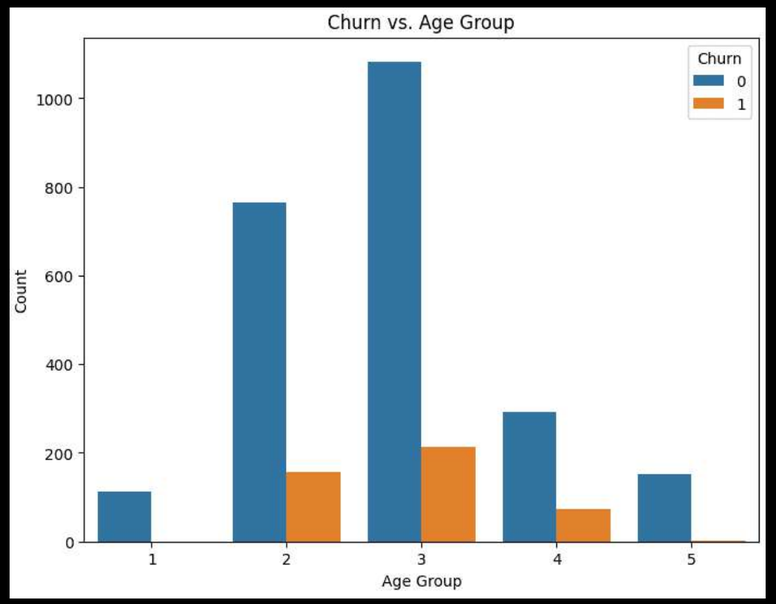


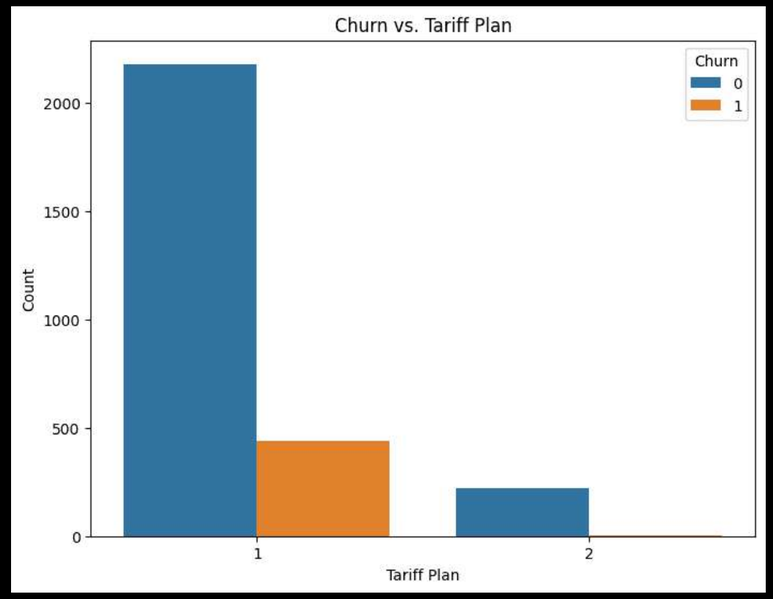


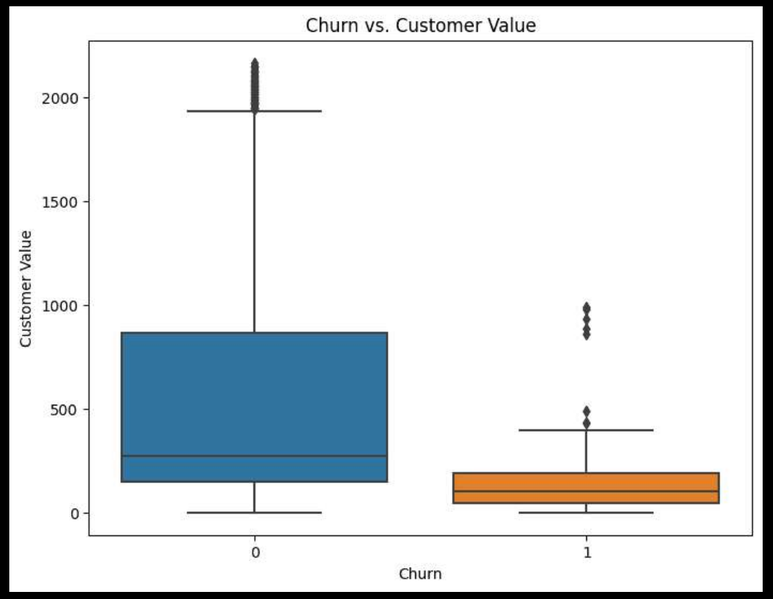
Exploratory Data Analysis

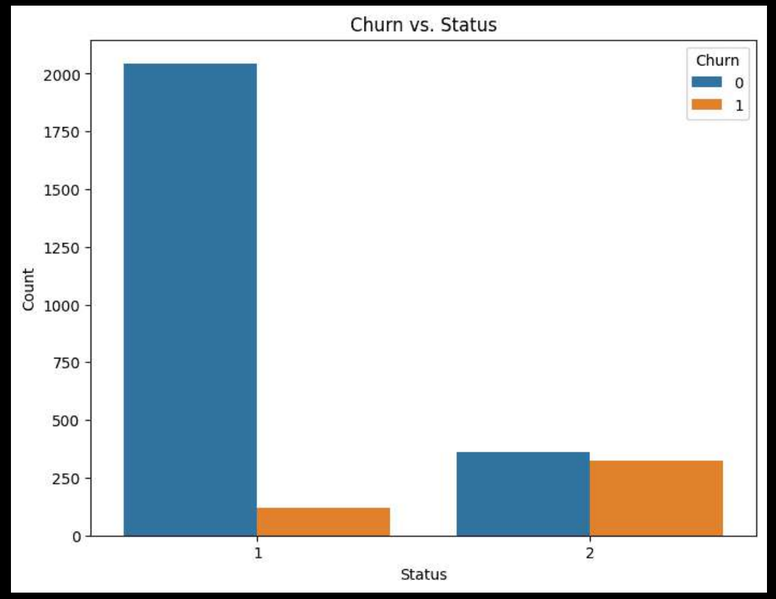








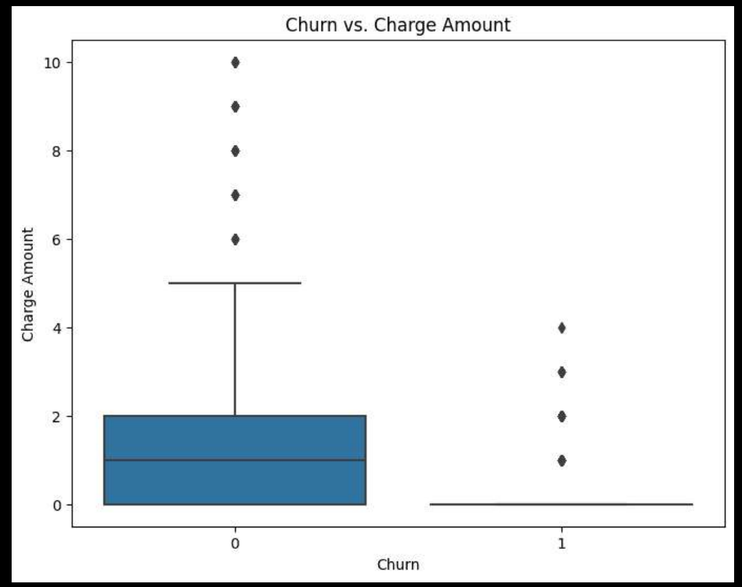


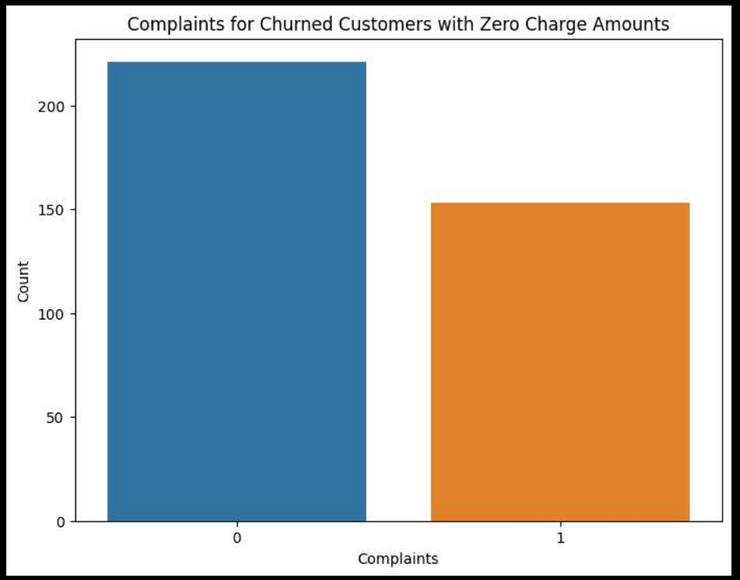


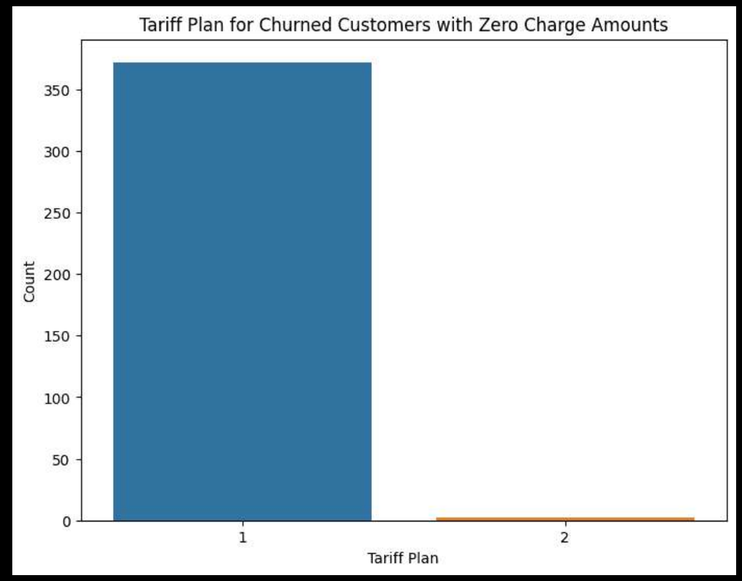
Exploratory Data Analysis

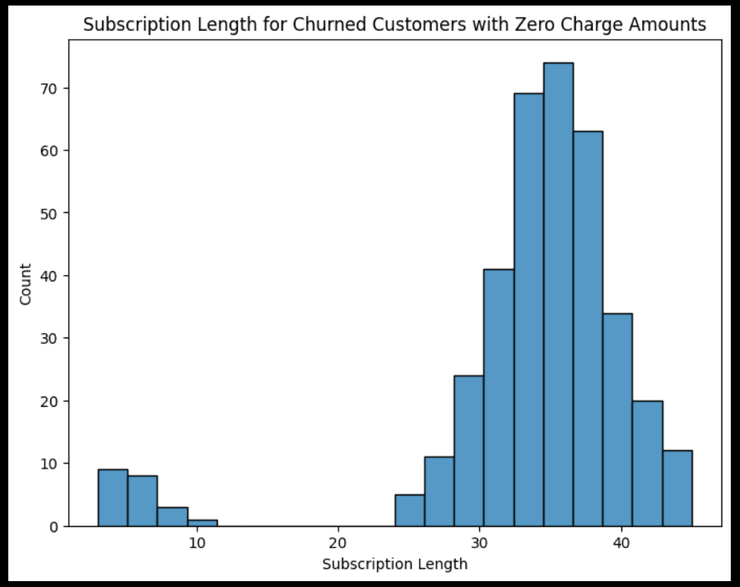
The presence of a straight line at zero in the box plot for churned customers suggests that a significant portion of churned customers have a charge amount of zero.

- Customers on the 'Pay as you go' plan might be more likely to churn without incurring charges.
- Significant portion of churned customers with zero charges have filed complaints, which could have resulted in charge amounts being adjusted or waived.
- The most common subscription lengths are around 34 to 38 months. There is no clear pattern of early terminations among churned customers with zero charges.
- The average seconds of use, frequency of use, and frequency of SMS are moderate. However, some customers have very low or zero usage in these categories, which might have contributed to zero charge amounts.

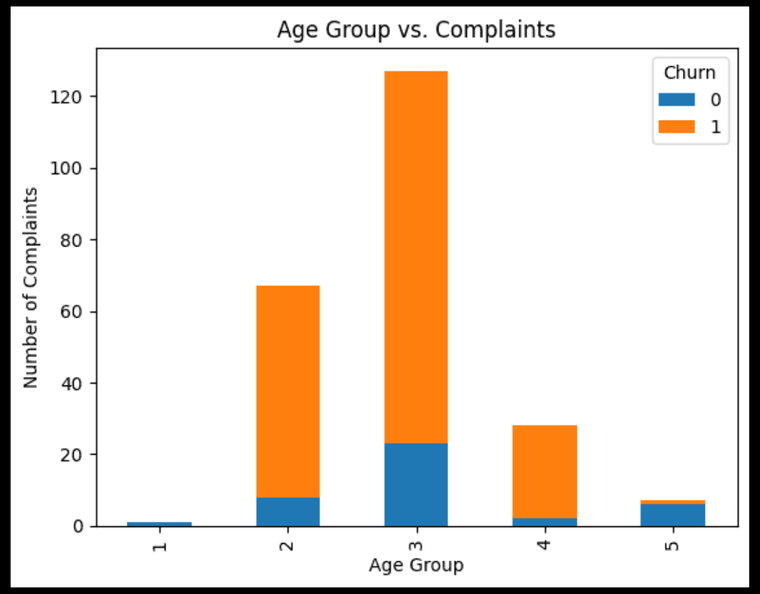


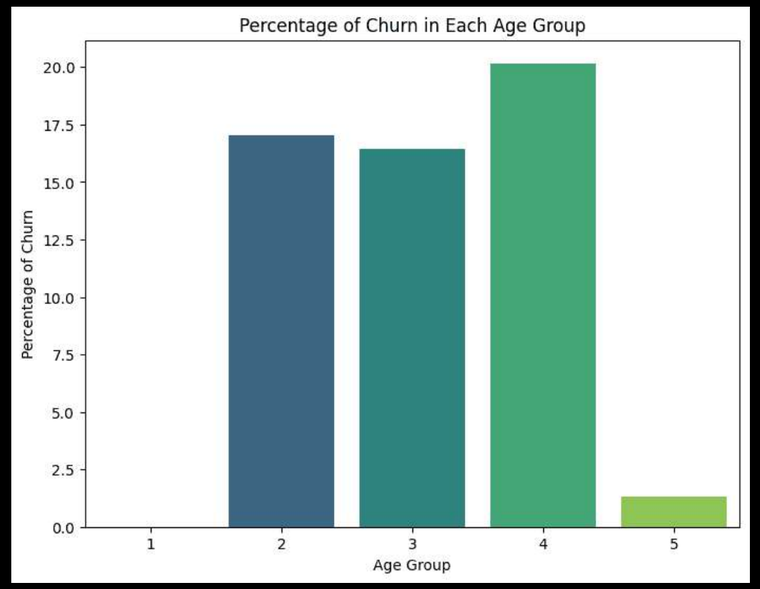






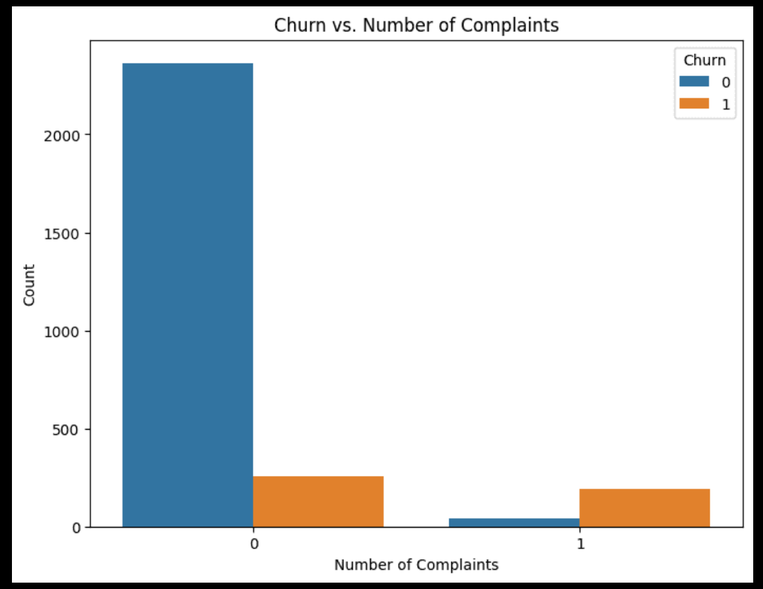
Exploratory Data Analysis

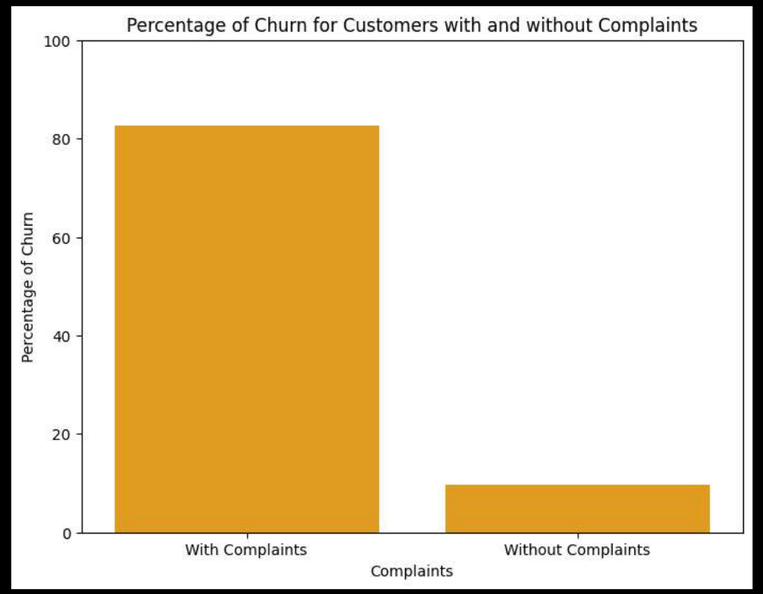




- Age Group 1: NaN - This means that there are no churned customers in Age Group 1 based on the available data.
- Age Group 2: 17.05% - Approximately 17.05% of customers in Age Group 2 have churned.
- Age Group 3: 16.44% - Approximately 16.44% of customers in Age Group 3 have churned.
- Age Group 4: 20.16% - Approximately 20.16% of customers in Age Group 4 have churned.
- Age Group 5: 1.30% - Approximately 1.30% of customers in Age Group 5 have churned.

Exploratory Data Analysis





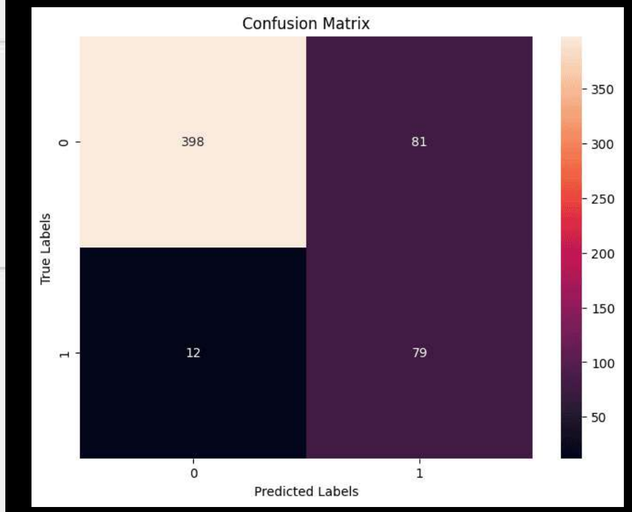
The percentages of churn for customers with and without complaints are quite different, with 82.61% of customers with complaints and only 9.77% of customers without complaints churning.

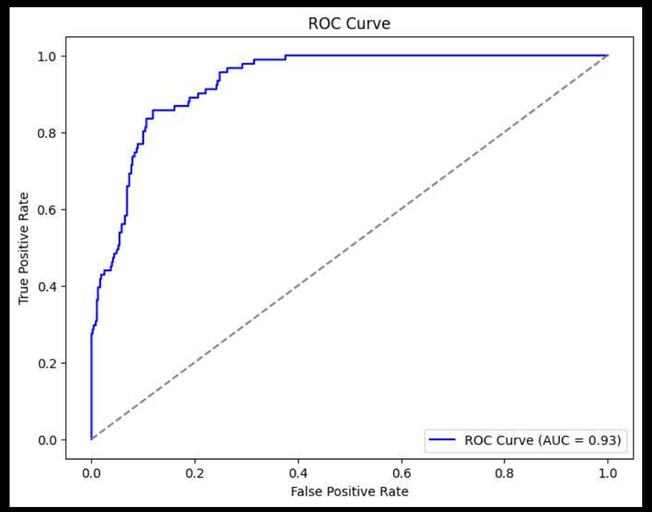
Class Imbalance Mitigation

We have a class imbalance issue in our dataset, where the number of non-churned customers is significantly higher than the number of churned customers. To address this issue, we used SMOTE, a technique that generates synthetic samples for the minority class by interpolating new data points between existing minority class samples. By doing so, SMOTE effectively increases the number of minority class samples, creating a more balanced dataset for training the model.

Model Building: Logistic Regression

- Accuracy: 0.8368 (83.68% of predictions are correct)
- Precision: 0.4938 (Out of the predicted positive churns, 49.38% are correct)
- Recall: 0.8681 (Out of all the actual positive churns, 86.81% are correctly predicted)
- F1 Score: 0.6295 (Harmonic mean of precision and recall)



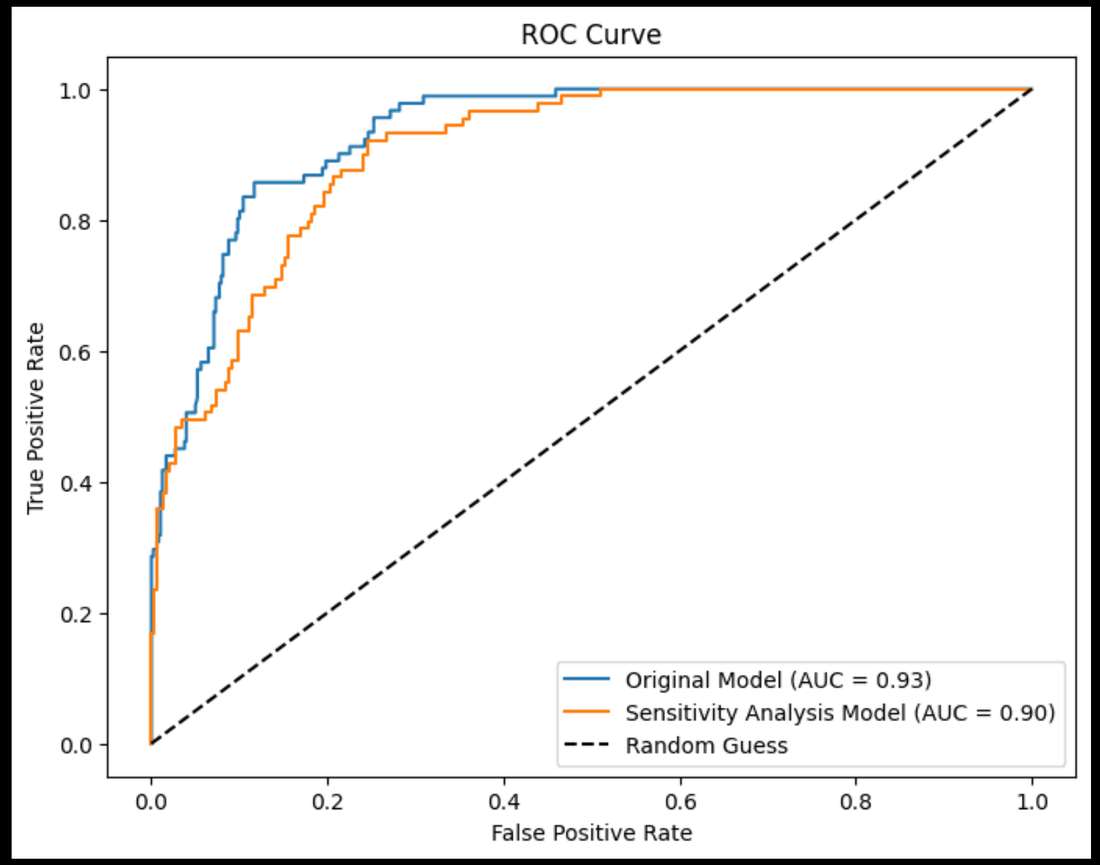


Sensitivity Analysis

The sensitivity analysis focused on the removal of outliers from the dataset. Outliers are data points that deviate significantly from the norm and can have a disproportionate impact on model performance. By conducting sensitivity analysis through outlier removal, we aimed to evaluate the model's response to variations in extreme data points.

- Accuracy: 0.7818 or 78.18%
- Precision: 0.5157 or 51.57%
- Recall: 0.9213 or 92.13%
- F1 Score: 0.66

Model Performance and Robustness



The evaluation metrics for the logistic regression model on the cleaned data are as follows:

Accuracy: 0.7818

Precision: 0.5157

Recall: 0.9213

F1 Score: 0.6612

Original Model (without outlier removal):

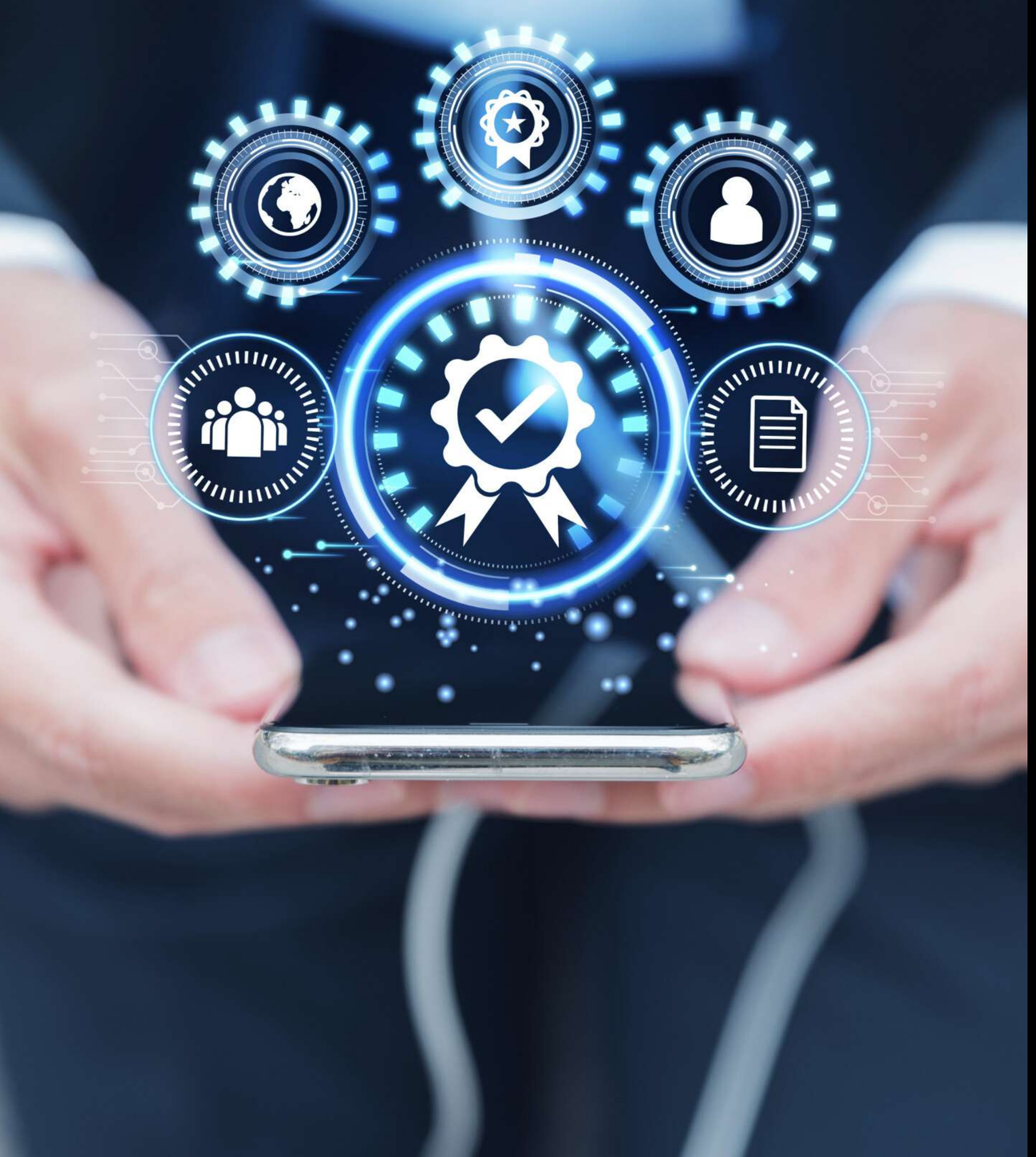
Accuracy: 0.8368

Precision: 0.4938

Recall: 0.8681

F1 Score: 0.6295

Based on the analysis and metrics obtained, we can say that the model is reasonably robust and performs well in predicting churn for the telecom customer dataset.



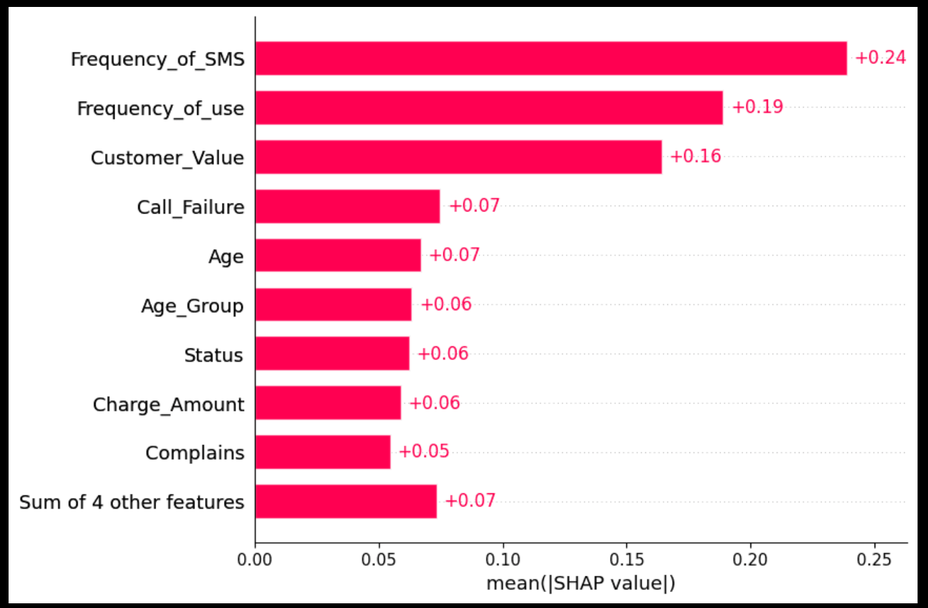
Feature Importance and Interpretation

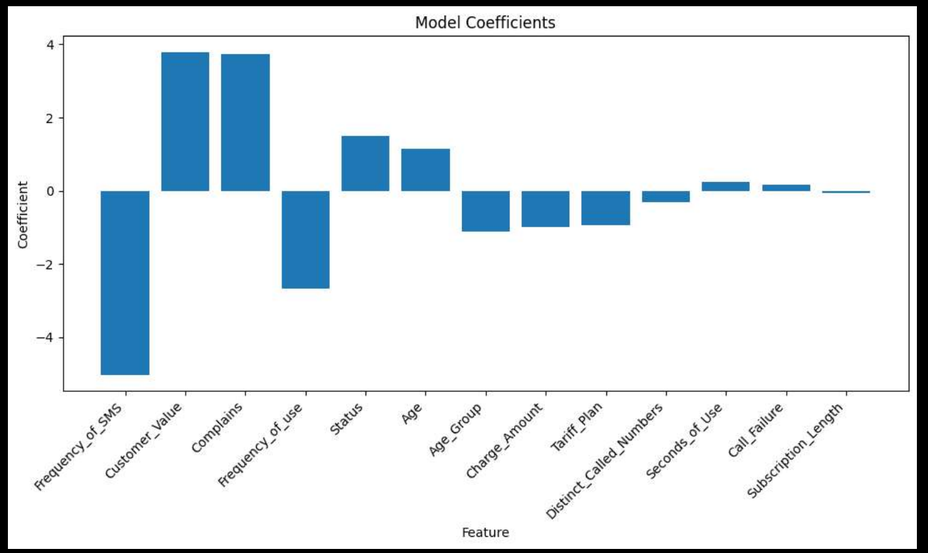
- **Frequency of use and Frequency of SMS:** Customers who make more frequent calls and send more text messages are more likely to churn. This may indicate that highly active customers may find better deals or services elsewhere, leading to churn.
- **Customer Value:** Higher customer value is associated with a reduced likelihood of churn. Customers with higher lifetime value to the company may be more satisfied with the services and benefits they receive, leading to higher retention rates.
- **Call Failure:** Customers experiencing call failures are more likely to churn. Call failures may indicate poor service quality, perceived as unfavorable by customers.
- **Age/Age Group:** Younger customers (lower age group) have a higher probability of churn compared to older customers. This may suggest the need for targeted retention strategies for different age groups.

Feature Importance Comparison

The rankings obtained from SHAP values and feature coefficients may not always match due to the inherent differences in the two methods.
SHAP values provide a more comprehensive and nuanced understanding of the impact of each feature on the model's predictions, taking into account feature interactions and dependencies.
On the other hand, feature coefficients from logistic regression represent the direction and magnitude of the linear relationship between each feature and the target variable.

- **Top 5 SHAP values:** Frequency of SMS, Frequency of use, Customer Value, Call Failure, Age
- **Top 5 Feature coefficients:** Frequency of SMS Customer Value, Complaints, Frequency of use, Status





Key Findings

Our analysis revealed several key findings. Firstly, customers with longer tenure and higher monthly charges were found to have lower churn rates. Secondly, poor network quality and customer service issues were identified as major drivers of churn. Lastly, targeted retention strategies based on customer segmentation showed promising results in reducing churn.

Recommendations

- **Implement targeted retention strategies:** Focus on customers with high 'Frequency_of_SMS', as they are more likely to churn. Offering personalized promotions or incentives to these customers can enhance loyalty.
- **Enhance customer value:** Higher 'Customer_Value' was associated with reduced churn rates. The company should focus on strategies to increase customer value through service enhancements or loyalty programs.
- **Improve Customer Experience and Quality of Service**: Although not represented directly in the model, factors contributing to 'Frequency_of_Use' and 'Distinct_Called_Numbers' may impact churn. Continuously monitor and improve the quality of service, network coverage, and overall customer experience to encourage customer loyalty.
- **Analyze Age Group-Specific Retention Strategies**: The 'Age_Group' feature exhibits varying effects on churn prediction across different age groups. Customize retention strategies based on the preferences and needs of different age groups. For instance, focus on tech-savvy offerings for younger customers and personalized services for older customers.
- **Address customer complaints promptly**: Customers who lodged complaints (Complains=1) showed a higher likelihood of churn. Improving complaint resolution processes and customer support can help retain these customers.

Future Steps and Further Research

- Feature engineering: Explore creating new features or transformations to capture potential non-linear relationships between features and churn.
- Model tuning: Optimize hyperparameters to fine-tune the model and potentially achieve better results.
- Ensemble methods: Explore ensemble techniques like Random Forest or Gradient Boosting to leverage the strength of multiple models.

Conclusion

In conclusion, our predictive analysis of customer churn patterns in an Iranian telecom company has provided valuable insights for the industry. By understanding the factors influencing churn and implementing appropriate strategies, telecom companies can mitigate churn and enhance customer loyalty, ultimately leading to improved business performance.

Thanks!

Do you have any questions?

soufleros.kostas@gmail.com

https://www.linkedin.com/in/konstantinos-soufleros/

https://github.com/kostas696