**Customer Churn Prediction: Analysis and Recommendations**

**Abstract**

I conducted research in this project to investigate customer churn. Customer churn represents

a critical challenge for the telecom industry, with significant implications for revenue and

customer retention strategies. My fascinating motto is to predict the root causes and main

reasons of the customer exit (Churn) from one telecom industry and entry to other industries.

Customer churn represents a critical challenge for the telecom industry, with significant

implications for revenue and customer retention strategies. Retaining customers are more cost

effective than gaining new ones, which makes churn prediction a vital tool for strategic planning.

My methods used in this project investigates the factors contributing to customer churn,

leveraging data visualization and machine learning models. Various machine learning algorithms

were applied to predict churn and identify its drivers. After performing thorough

research, it was determined that Random Forest model achieved the highest accuracy of 77.19%.

This overall high percentage demonstrated reliability in predicting churn for this project and

proved to be the most effective. Through these findings and exploring key trends, we were able

to highlight the value as well as the importance of leveraging data-driven approaches to enhance

retention and motivate customer loyalty. The goal here from a business standpoint is to acquire

long term growth along with a competitive advantage. We want to address real-world business

challenges that can arise and suggest the best solutions that benefit the company and its potential

of longevity. This paper discusses the dataset, methods, results, machines, and actionable

business insights to mitigate churn.

**Introduction**

Churn is my Target/Dependent feature for my project work flow. In this project we will

be using a dataset of 7043 entries and 21 features from Kaggle. We will explore the factors that

strongly impact the subscriber’s exit. These findings will be identified & displayed throughout

my entire project journey to highlight key factors using Data Visualization and Machine

Learning Modeling Algorithms. This paper discusses the dataset, methods, results, actionable

business insights and focusing on the key factors influencing churn and develop predictive

models to mitigate it. At the end of this project we will emphasize and share my discovery of the

best suggestions that can prevent customer churn, retain customers and acquire new customers.

**Data and Methods**

**2.1 Dataset Overview**

- Source: Kaggle.

- Structure: 7043 rows and 21 features.

- Handling Missing Values:

- Total Charges: Filled missing values with the mean.

- Removed 11 rows where tenure = 0 to maintain data integrity.

**Critical Features:**

- Numerical: Tenure, Monthly Charges, Total Charges.

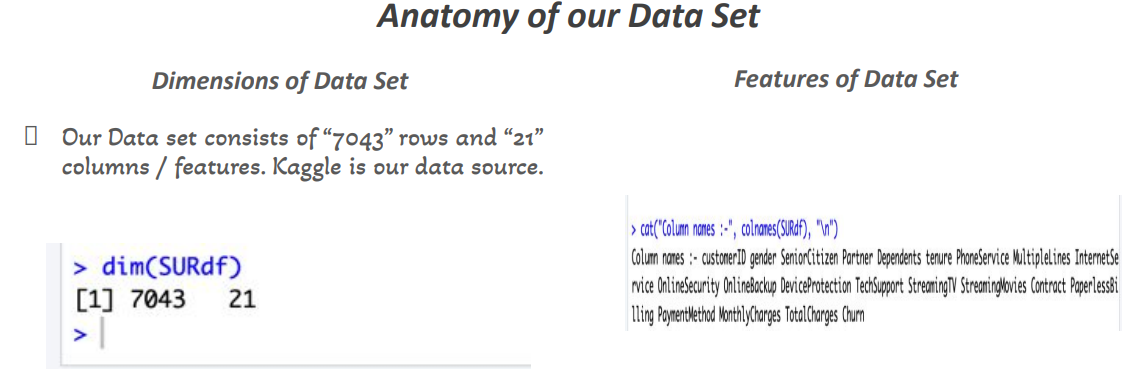
- Categorical: Gender, Internet Service, Contract, Payment Method.

**2.2 Feature Engineering**

- Senior Citizen: Mapped values 0 to "No" and 1 to "Yes" for readability.

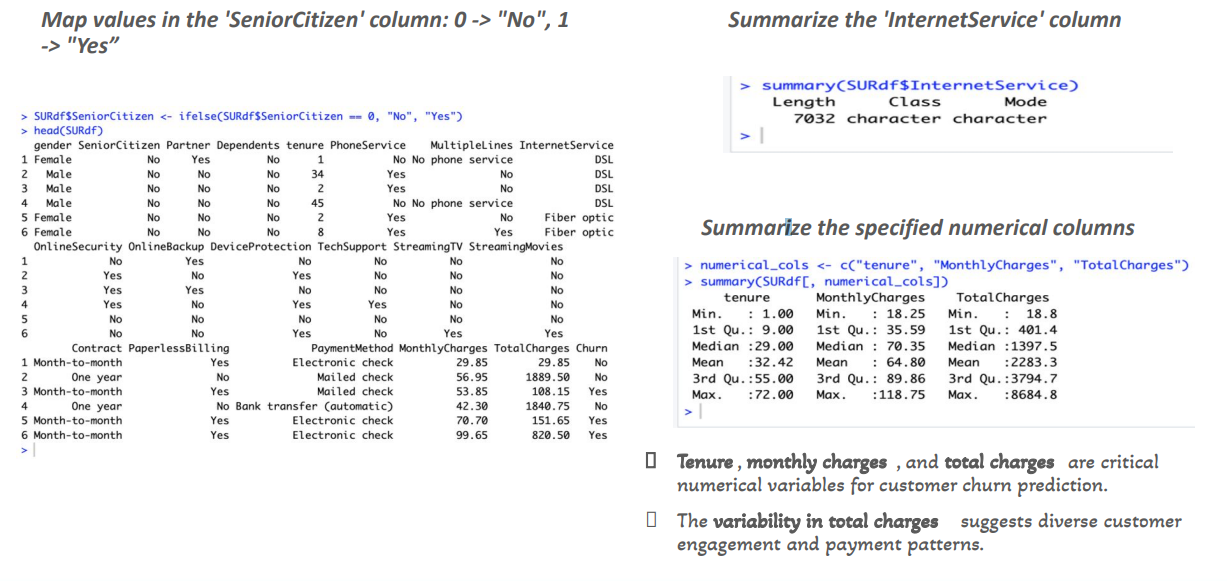
- Normalized numerical features to ensure consistent scaling.

- Encoded categorical variables using label and one-hot encoding methods.









**2.3 Data Visualization**

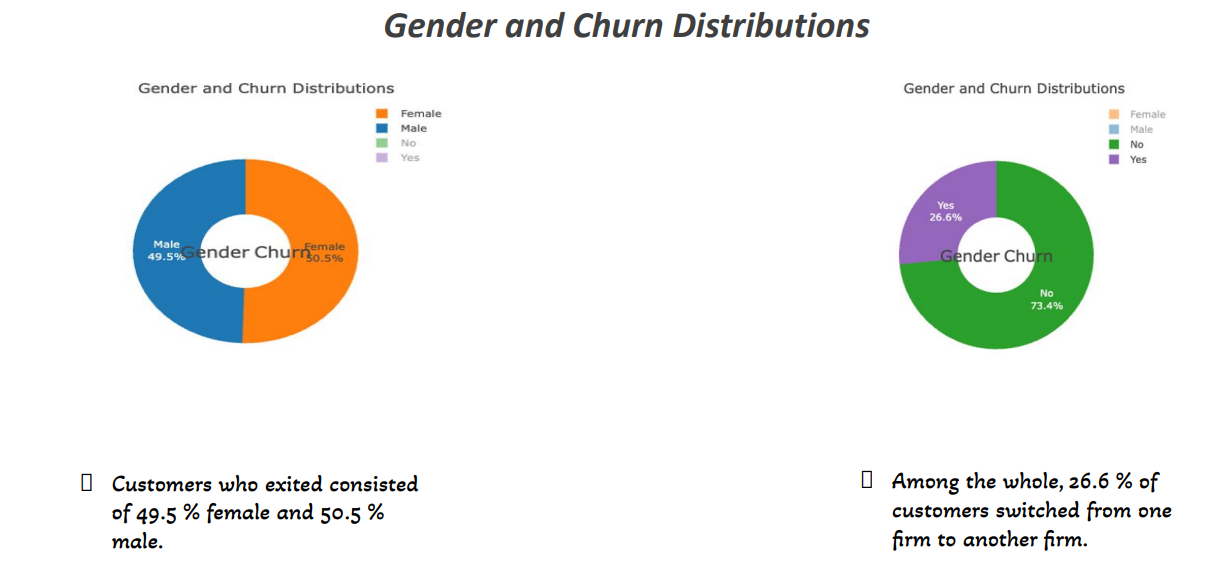
**Key trends and distributions were visualized:**

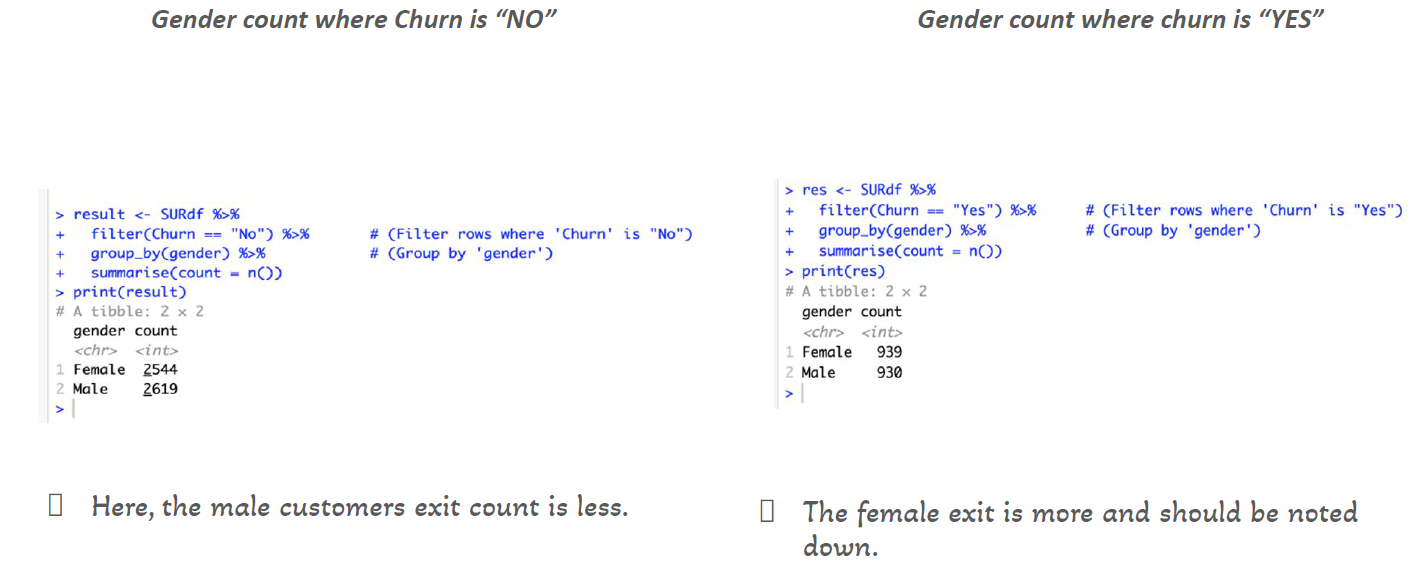
- Gender and Churn: Female customers showed slightly higher churn rates.

- Payment Methods: Electronic check users exhibited the highest churn.

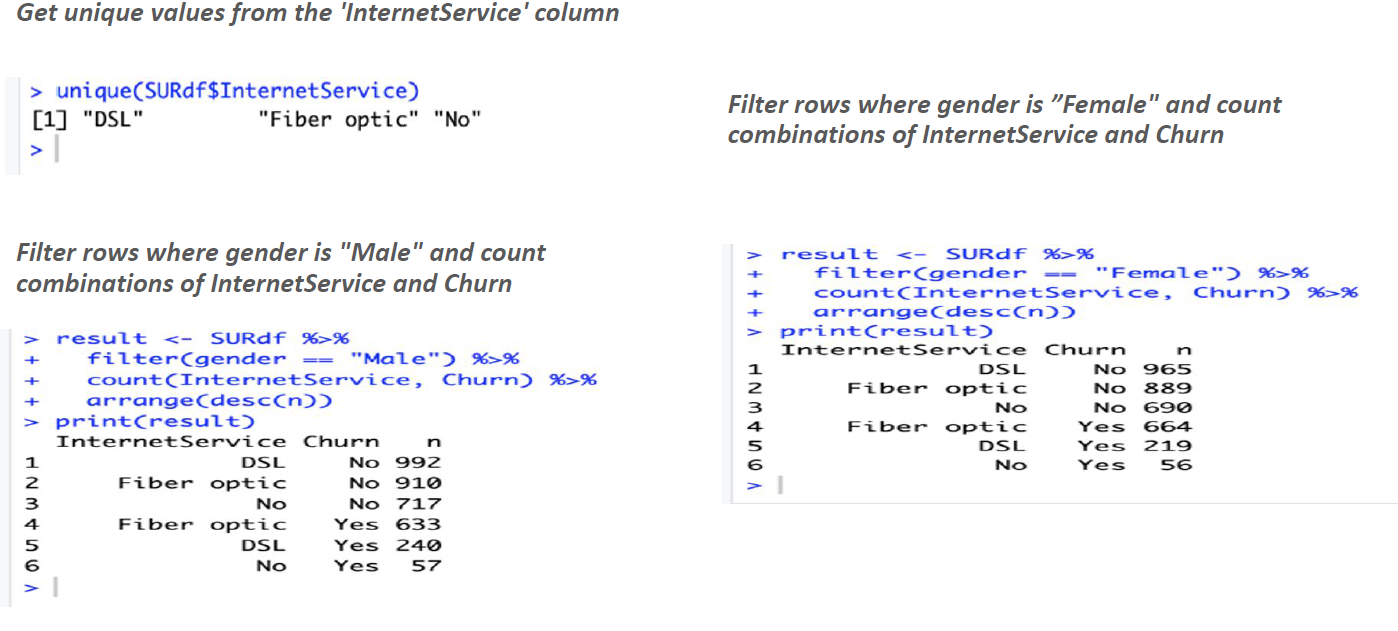
- Contract Types: Month-to-month contracts had the most significant churn rates.

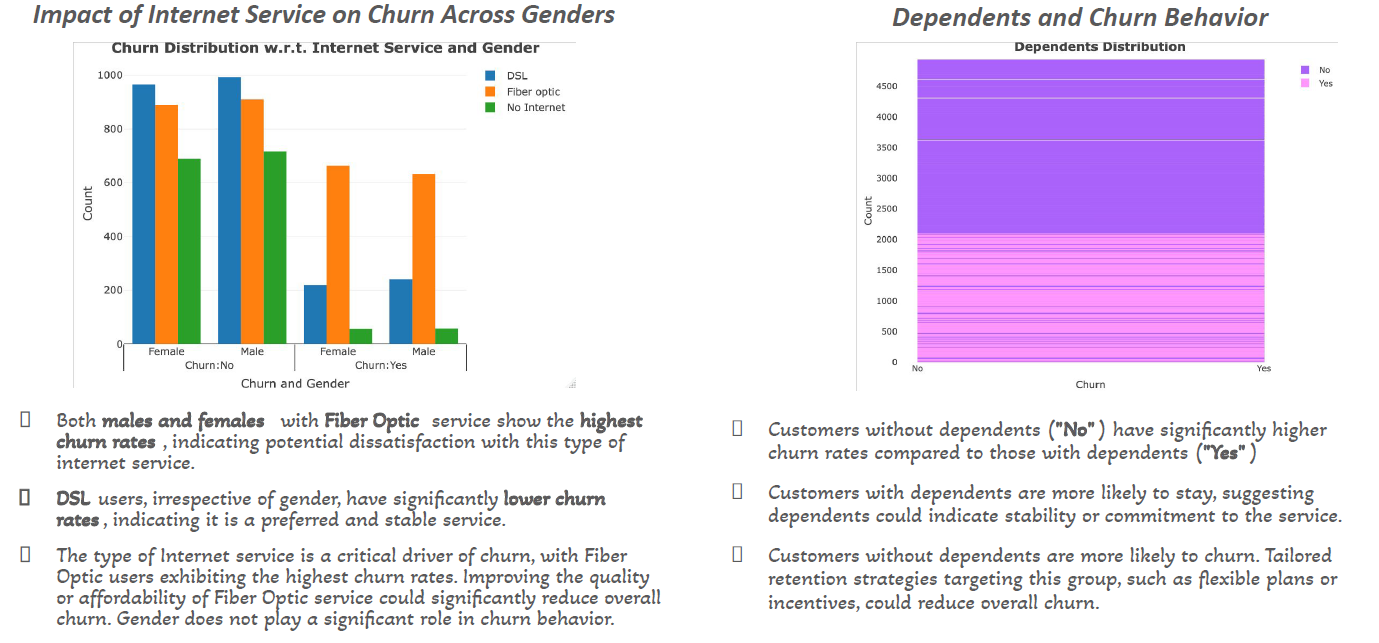
- Internet Services: Fiber Optic users showed higher churn rates compared to DSL users.

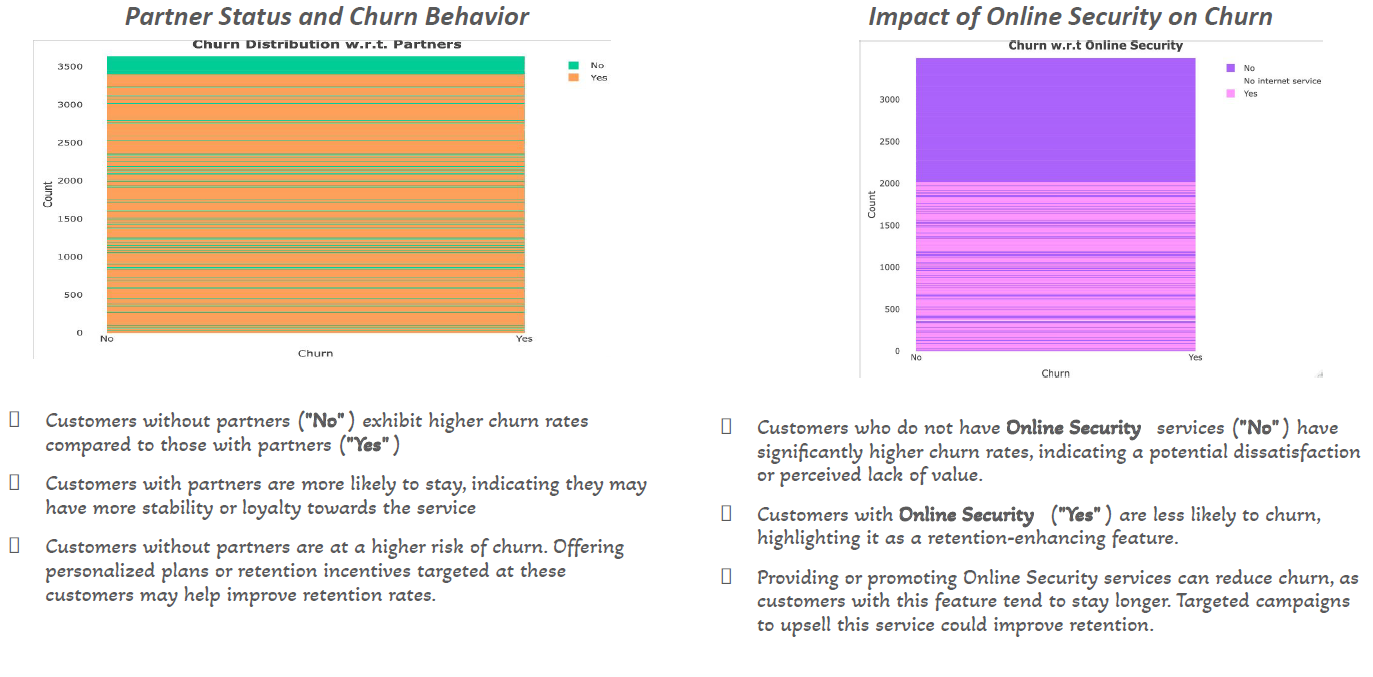


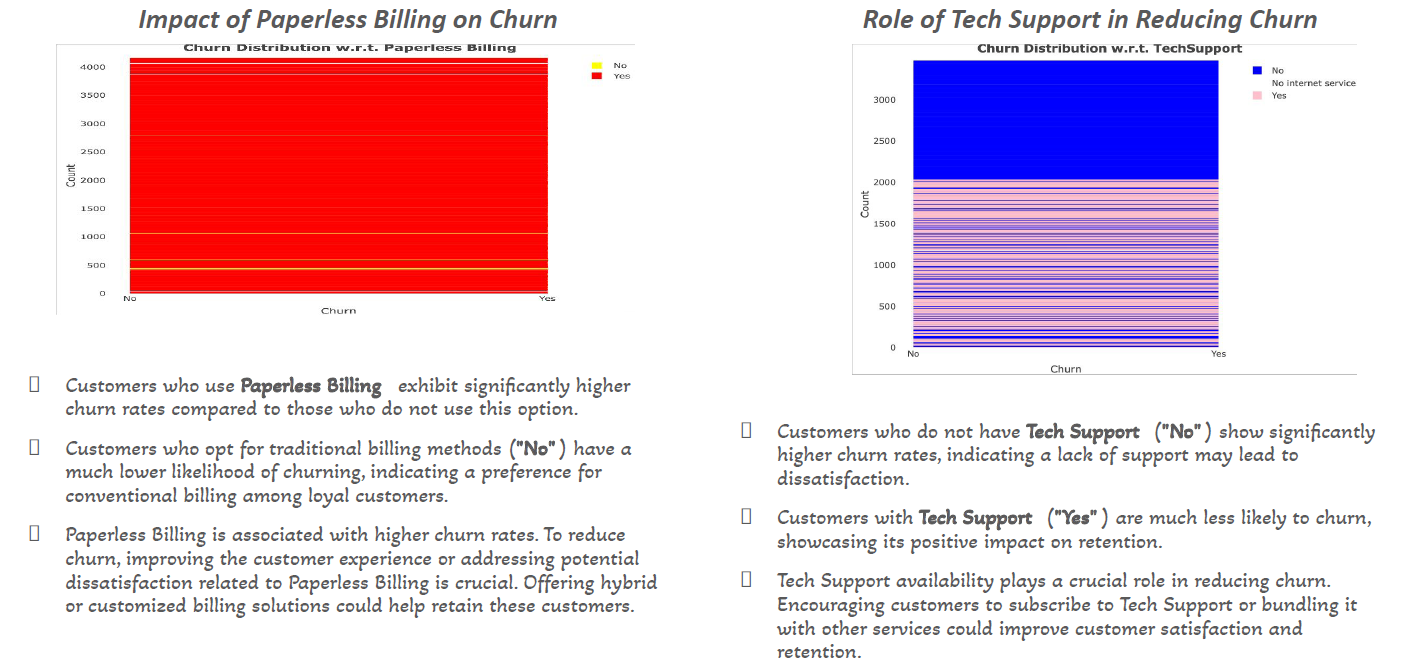


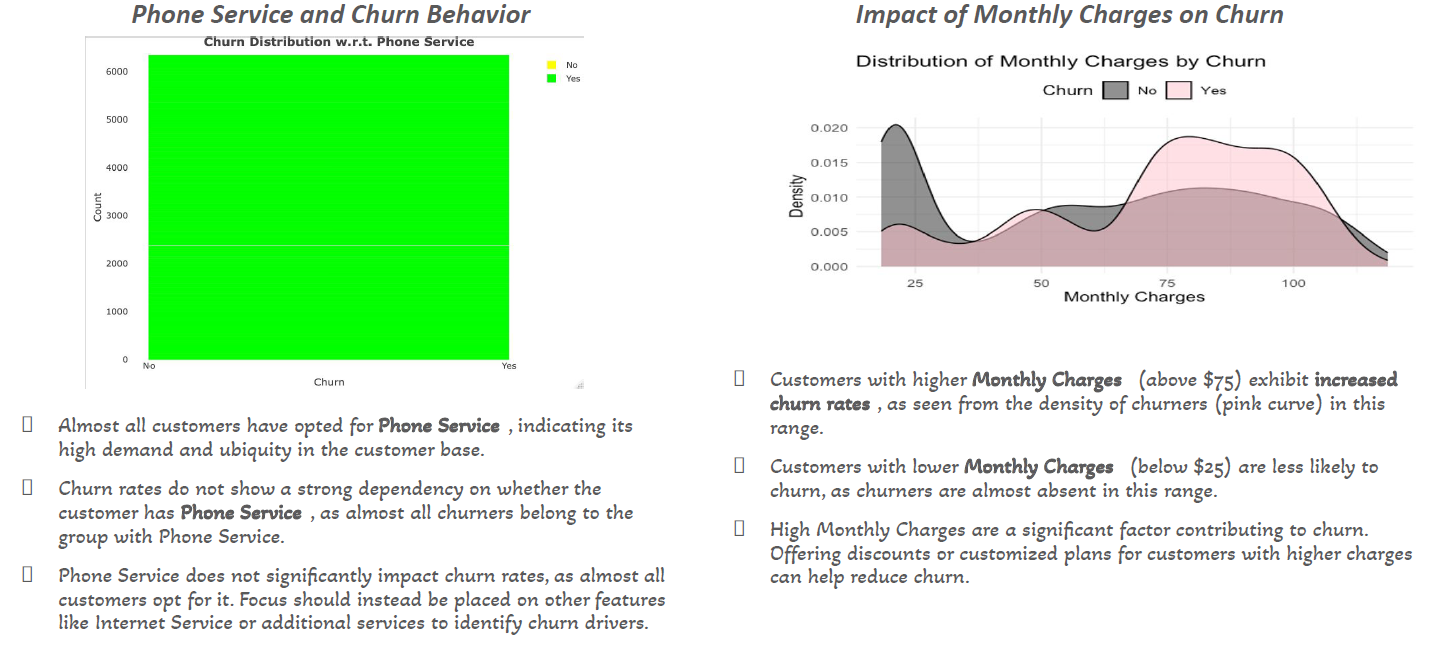


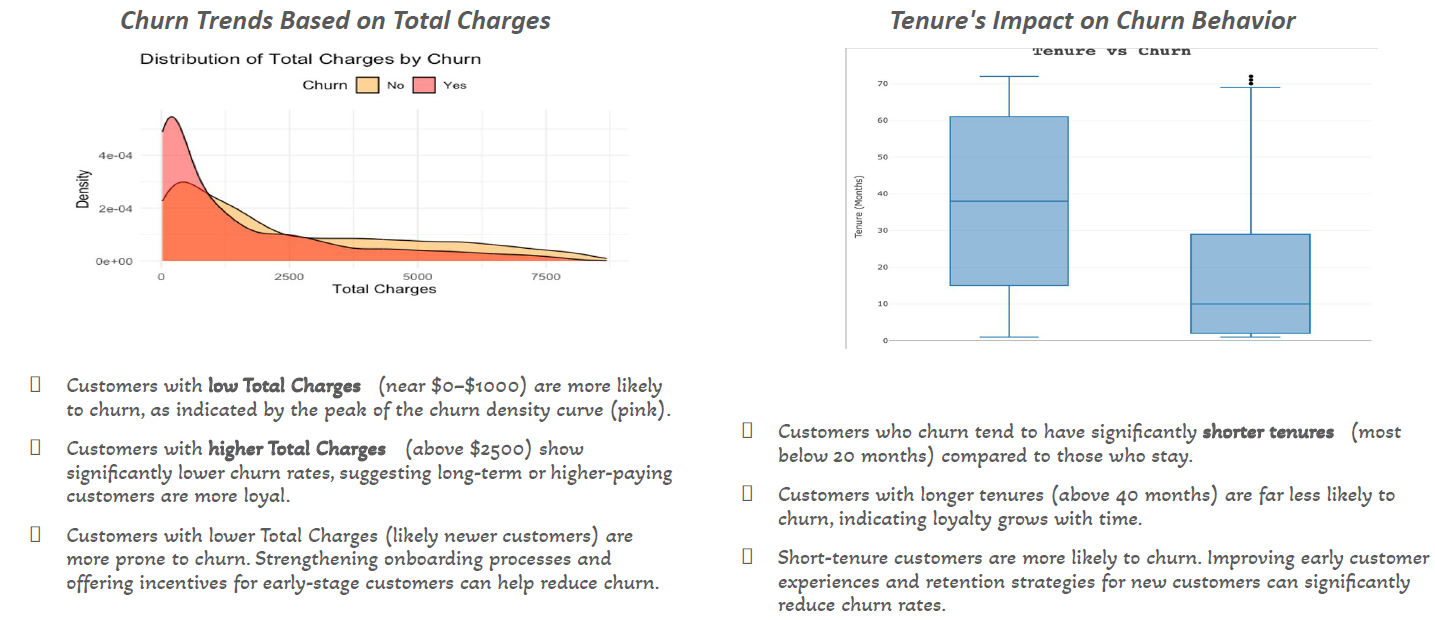


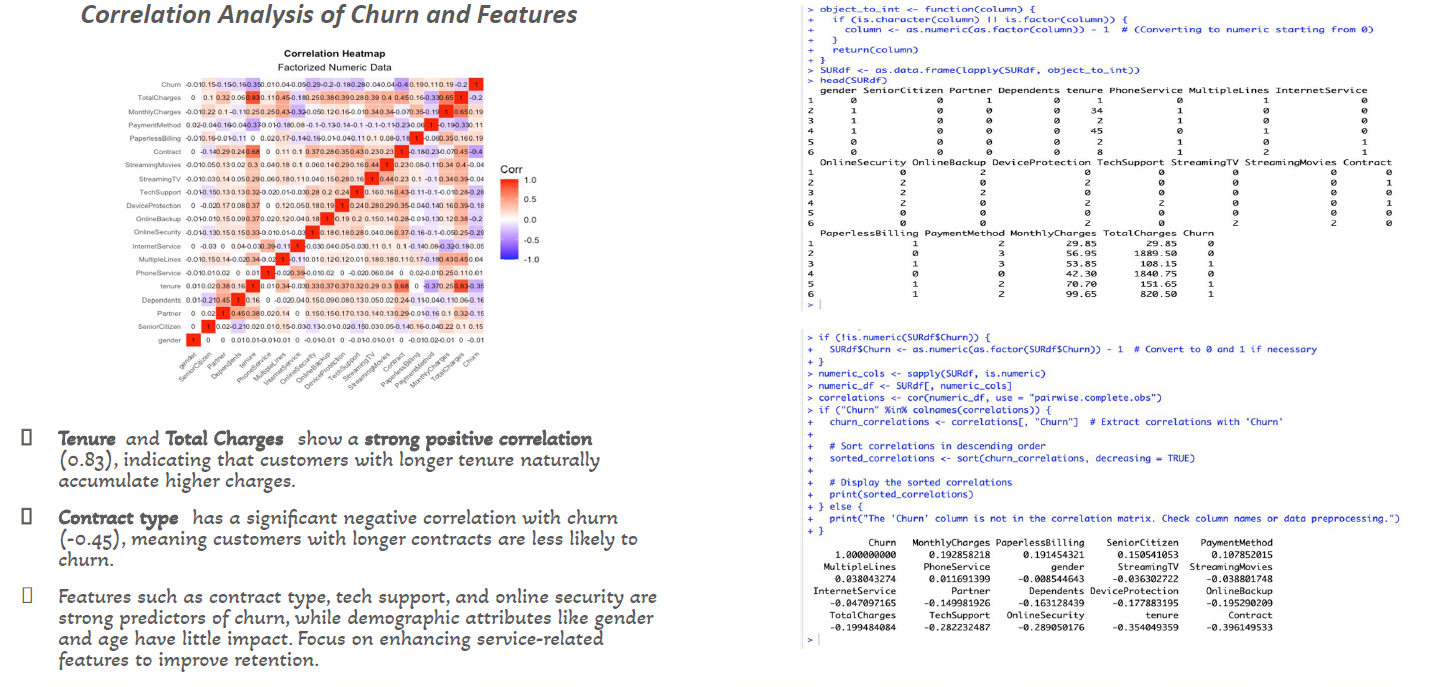


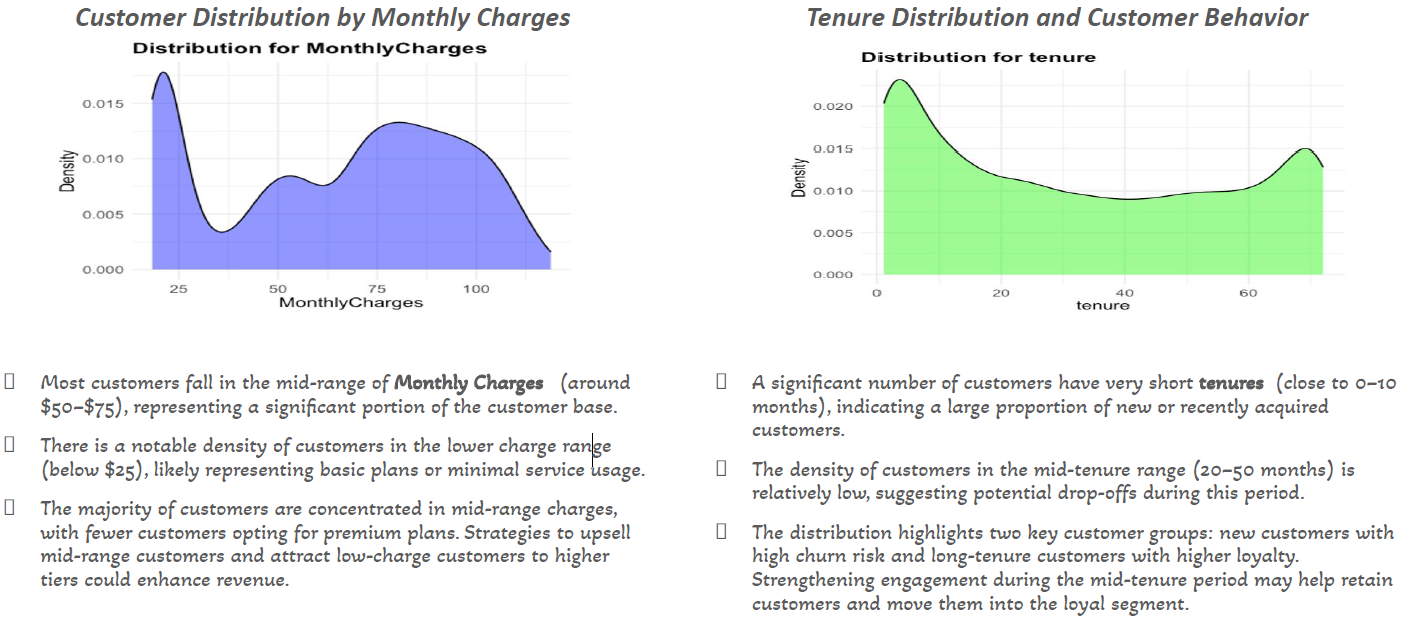


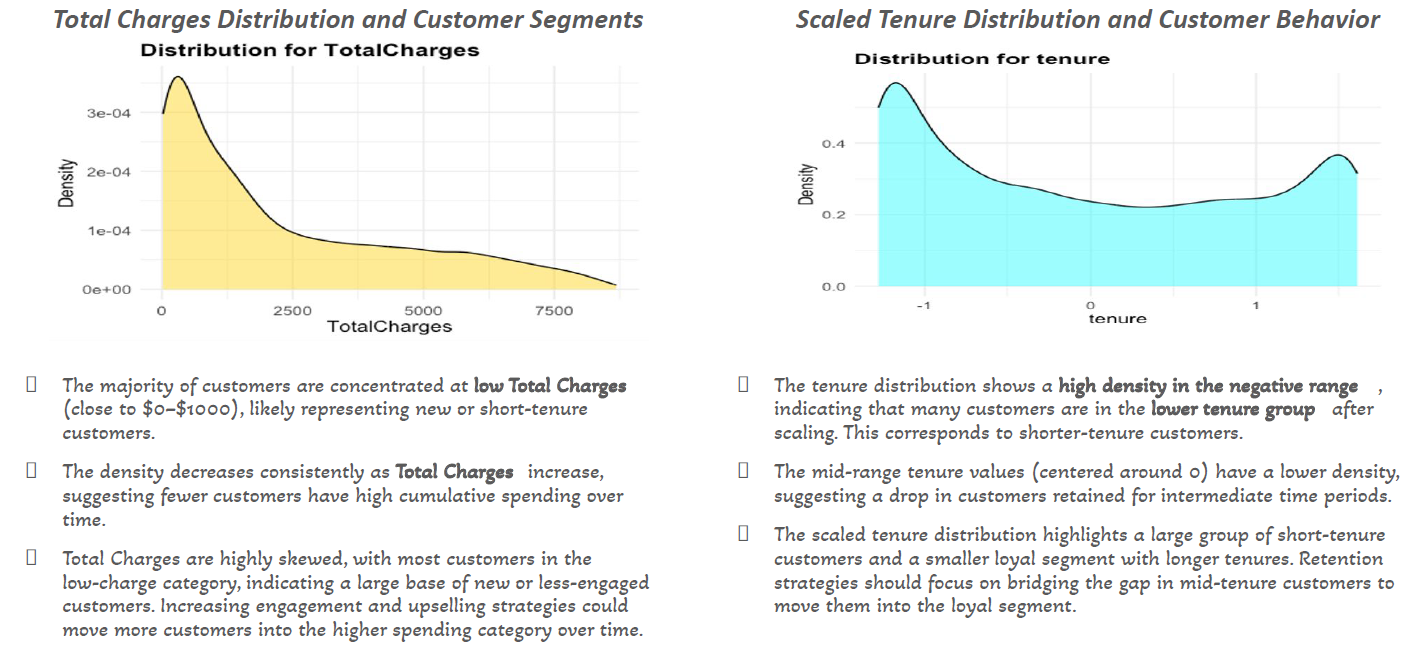


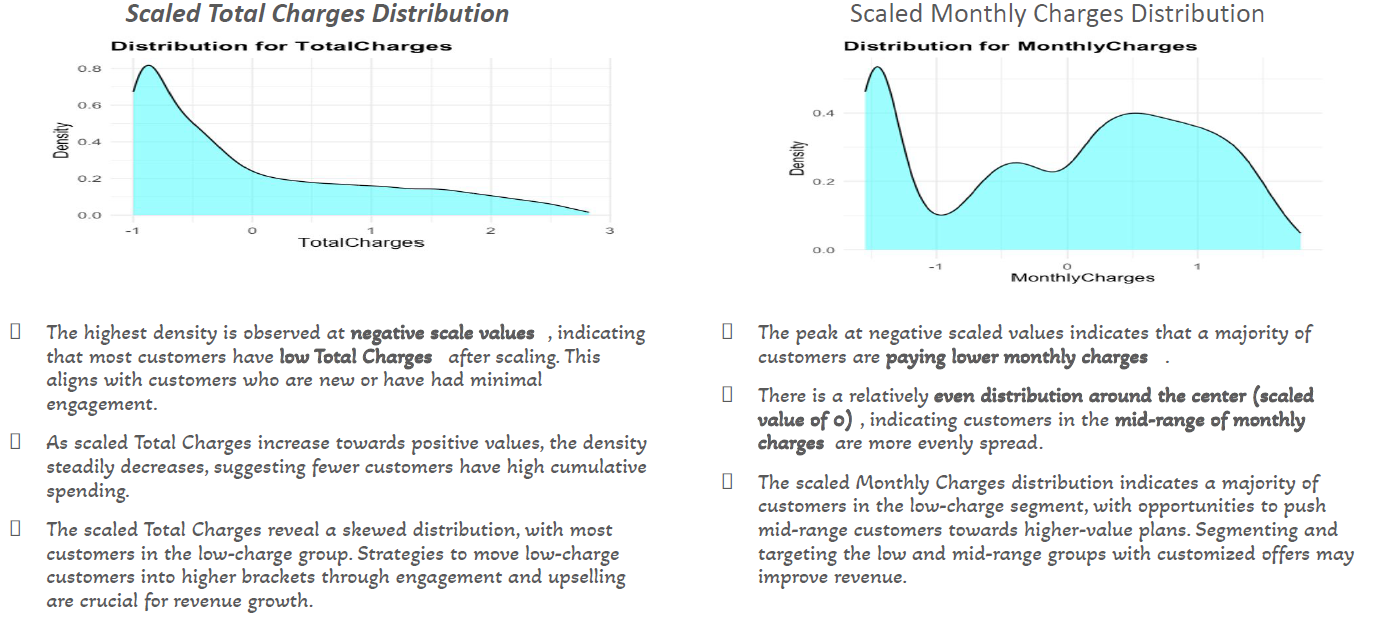








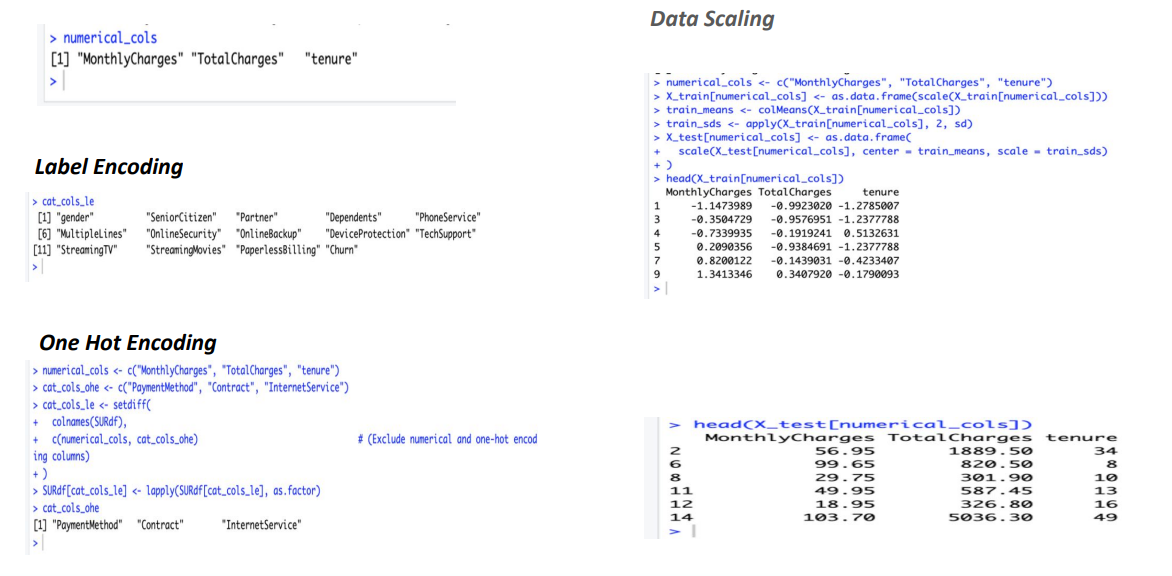




**2.4 Data Preprocessing**

- Normalized numerical features to reduce the impact of scale.

- Balanced class distributions using oversampling techniques.



**Machine Learning Models**

**3.1 Models Tested**

**K-Nearest Neighbors (KNN):**

- Accuracy: 72.83%.

- Struggled with class imbalance, predicting non-churn more effectively than churn.

- Random Forest:

- Best-performing model with 77.19% accuracy.

- Demonstrated strong feature importance and balance in predictions.

- Logistic Regression:

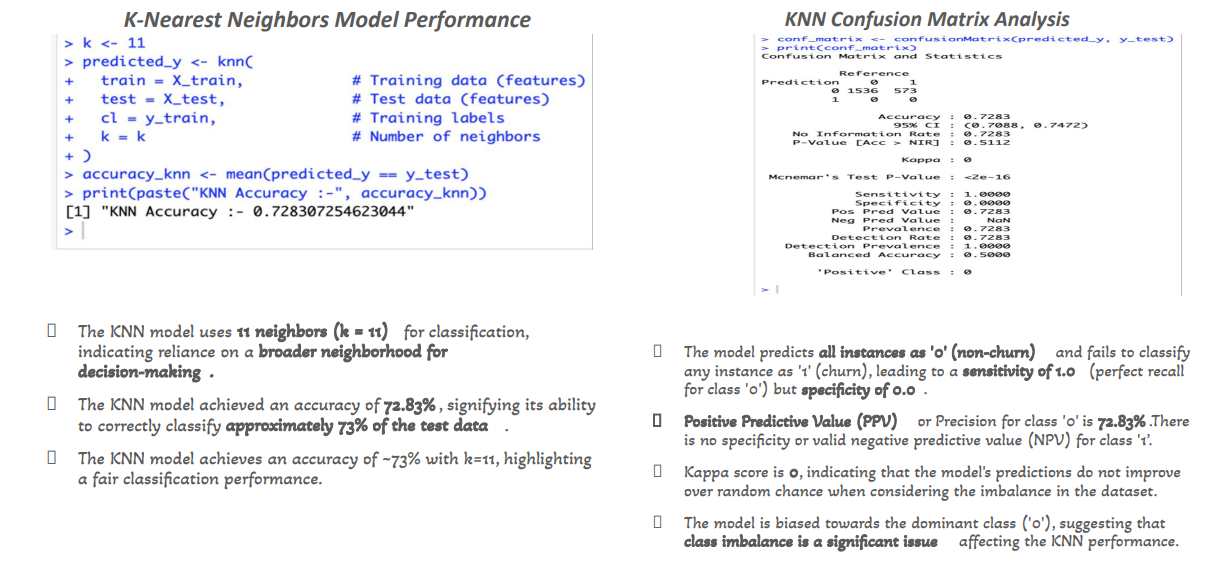
- Accuracy: 69.99%.

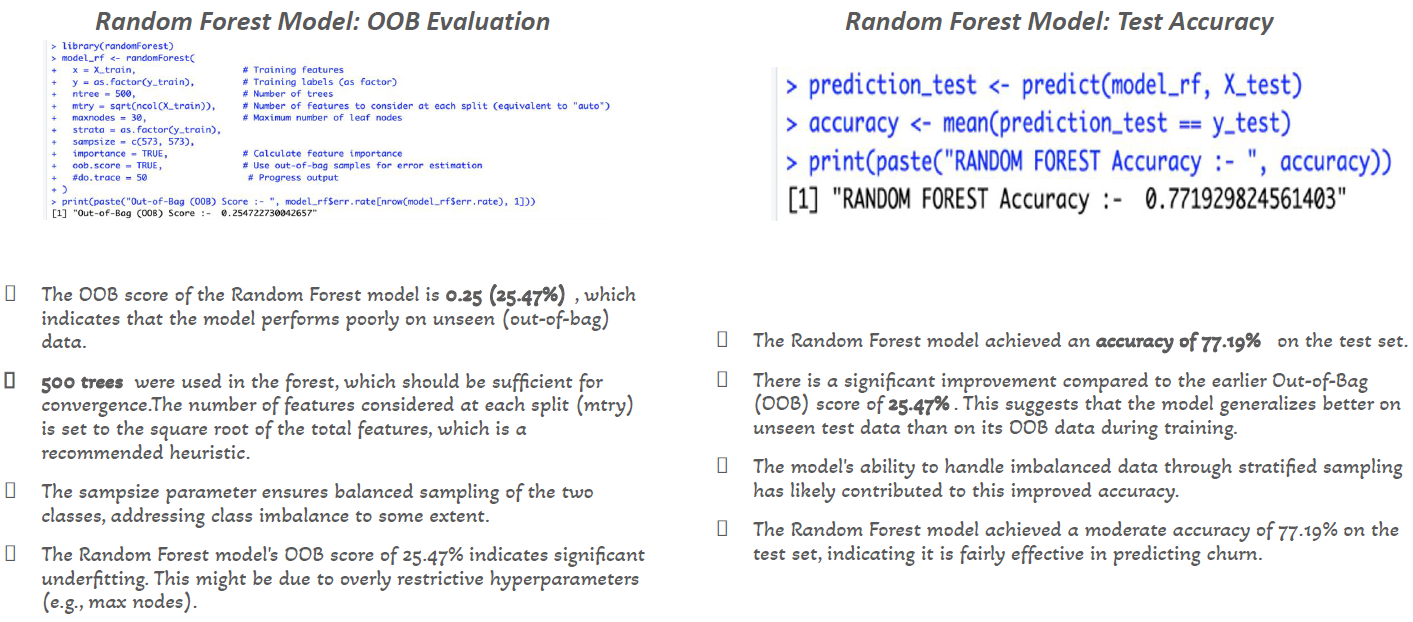
- Balanced dataset using oversampling improved model performance.

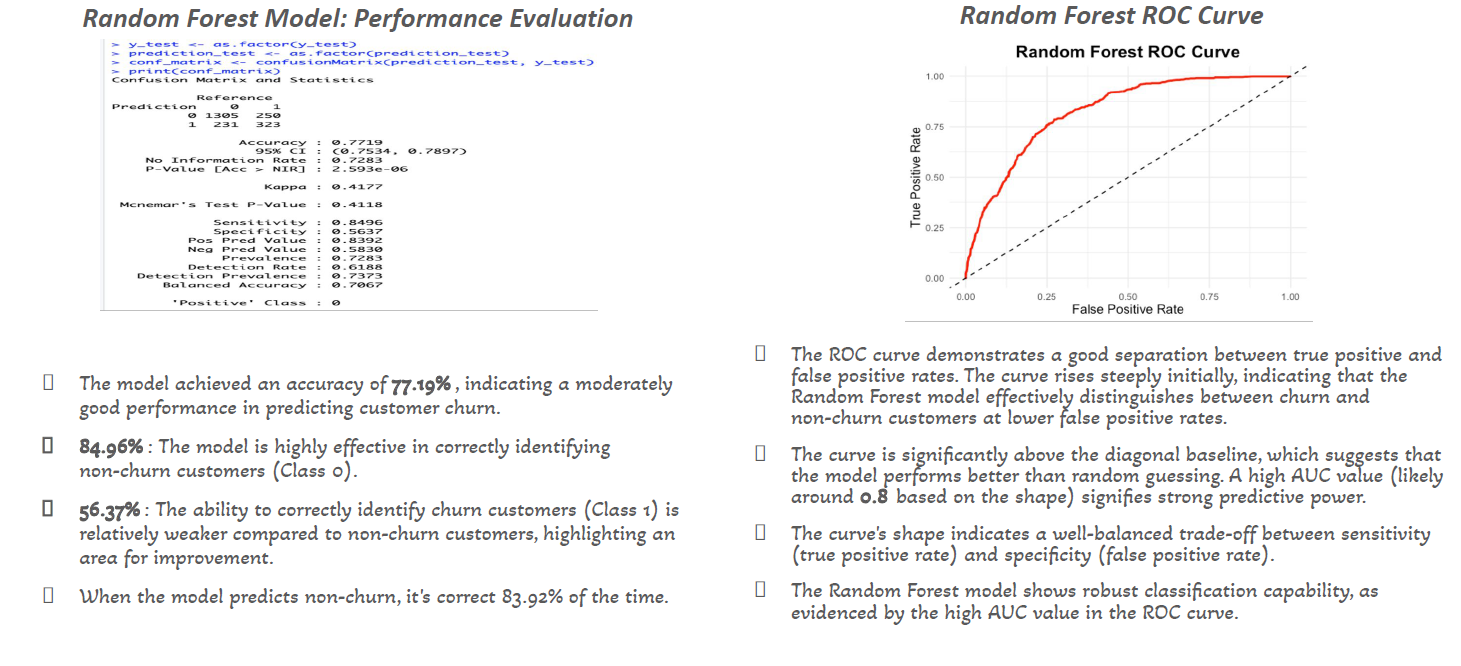
- Other Models:

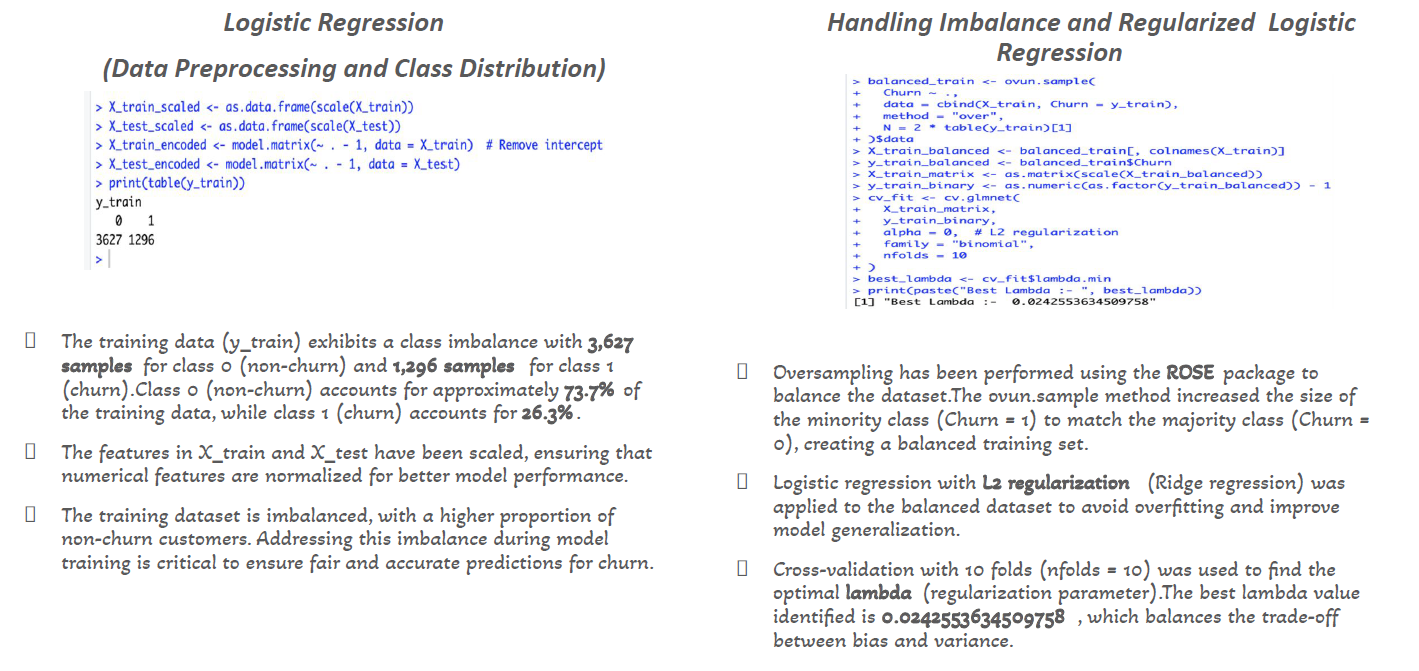
- Decision Trees and AdaBoost provided moderate results.

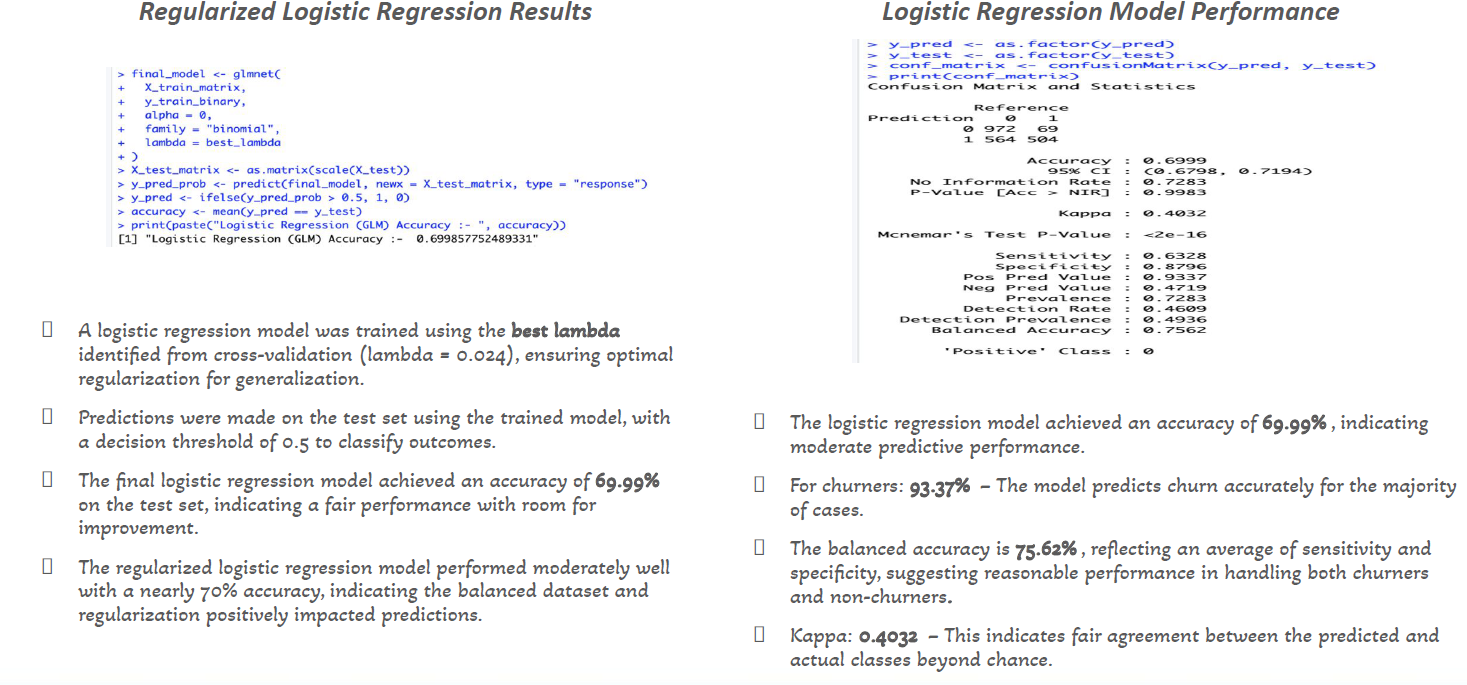
- XG Boost highlighted issues with false negatives.



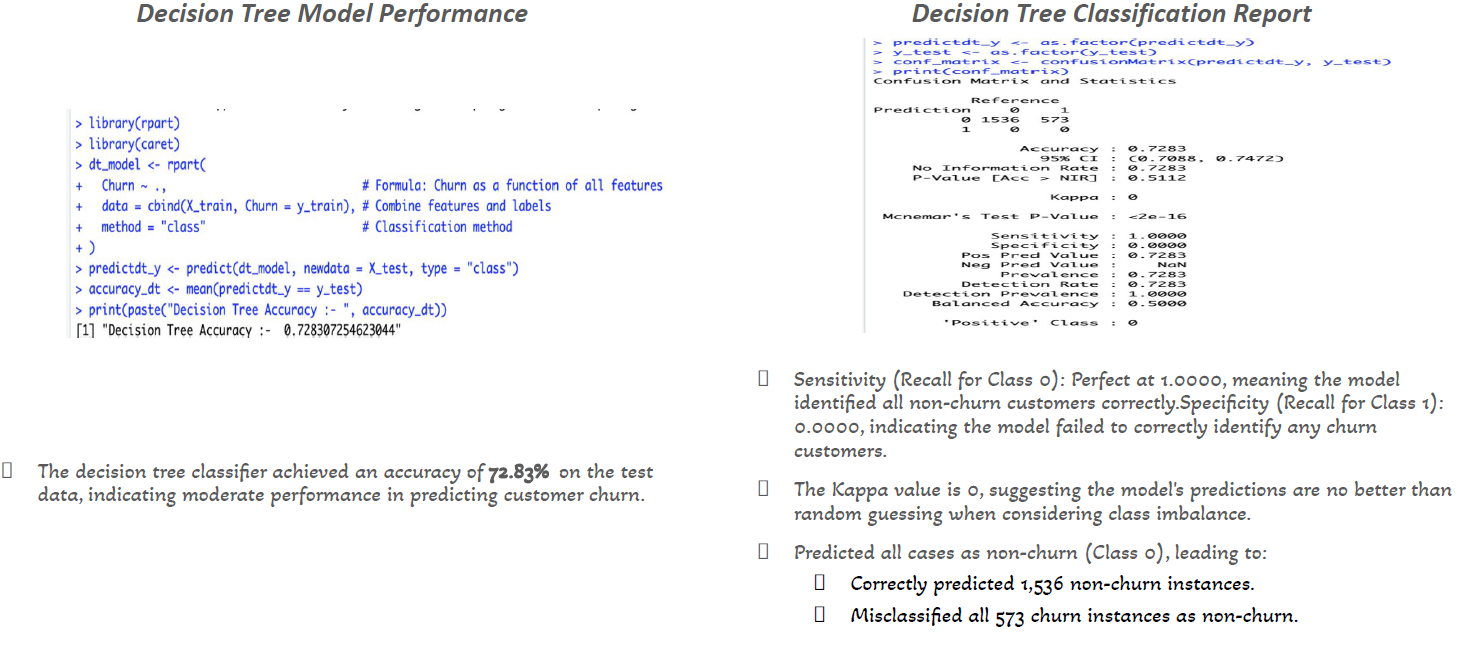


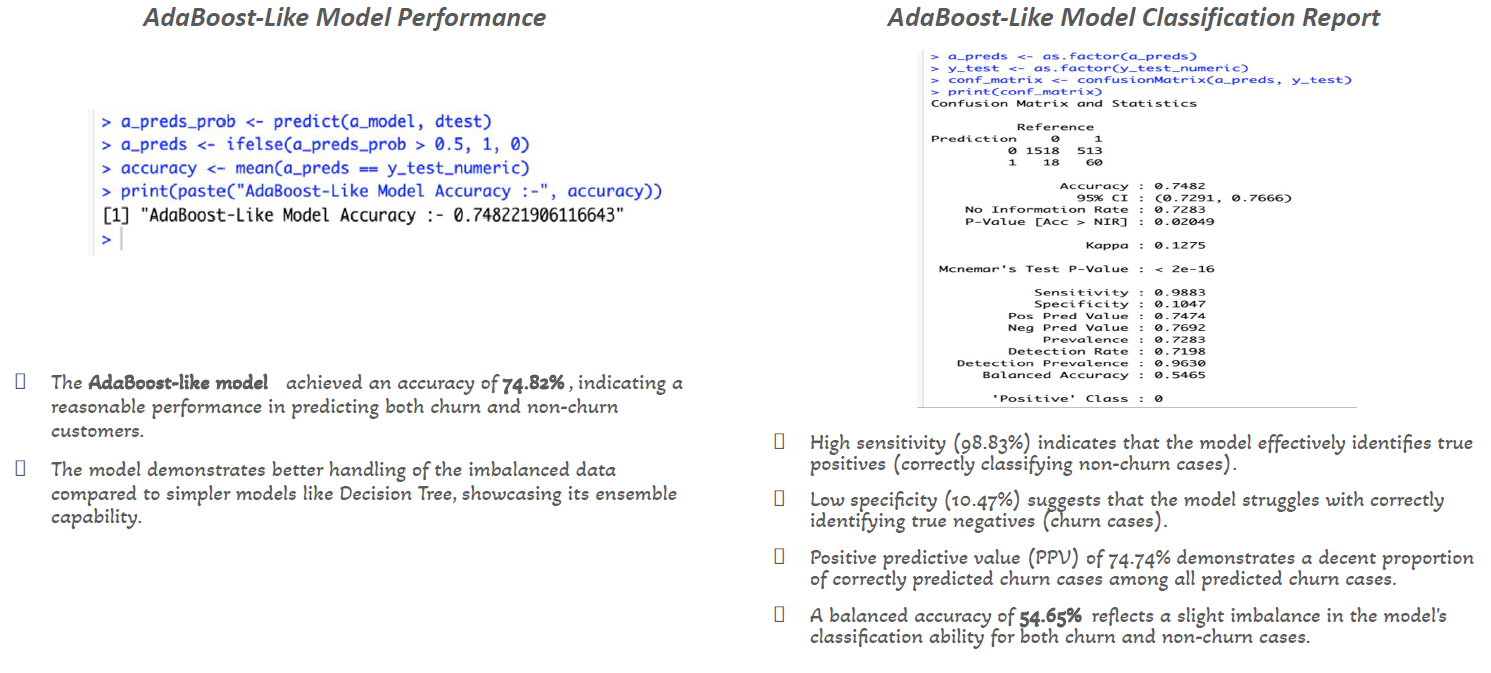












**3.2 Model Evaluation Metrics**

- Accuracy, Precision, Recall, ROC-AUC, and Kappa Score were used to assess model

performance. Random Forest emerged as the most reliable model due to its high accuracy and

balanced prediction capabilities.

**Results**

**4.1 Key Drivers of Churn**

- Contract Types:

- Customers on month-to-month contracts exhibited higher churn rates.

- Longer contracts (one or two years) correlated with greater customer retention.

**Payment Methods:**

- Electronic check users were more likely to churn, suggesting dissatisfaction with this payment

method.

- Automatic payments via credit cards or bank transfers showed lower churn rates.

**Internet Services:**

- Fiber Optic users had the highest churn rates, indicating potential service quality issues.

- DSL users showed lower churn rates, suggesting higher satisfaction.

**Dependents and Partners:**

- Customers without dependents or partners exhibited higher churn rates.

**Service Features:**

- Lack of online security and tech support significantly increased churn.

- Customers with these features were more likely to stay.

**Monthly and Total Charges:**

- High monthly charges (≥ $75) correlated with increased churn.

- Customers with low total charges (≤ $1000) were more prone to churn, indicating risks among new customers.



**4.2 Business Insights**

- Incentivize Long-Term Contracts:

- Offer discounts to encourage customers to shift from month-to-month to annual contracts.

- Improve Payment Methods:

- Promote secure and automatic payment methods to enhance satisfaction.

- Enhance Internet Services:

- Focus on improving Fiber Optic services to reduce dissatisfaction.

- Targeted Retention Programs:

- Develop incentives for customers without dependents or partners.

- Upsell Service Features:

- Bundle online security and tech support with base packages to increase retention.

- Reduce Early Churn:

- Strengthen onboarding processes and provide benefits for new customers.

**Conclusion**

The key drivers of churn in this report included contract types, payment methods, and service

features. Random Forest proved to be the most effective model for predicting churn,

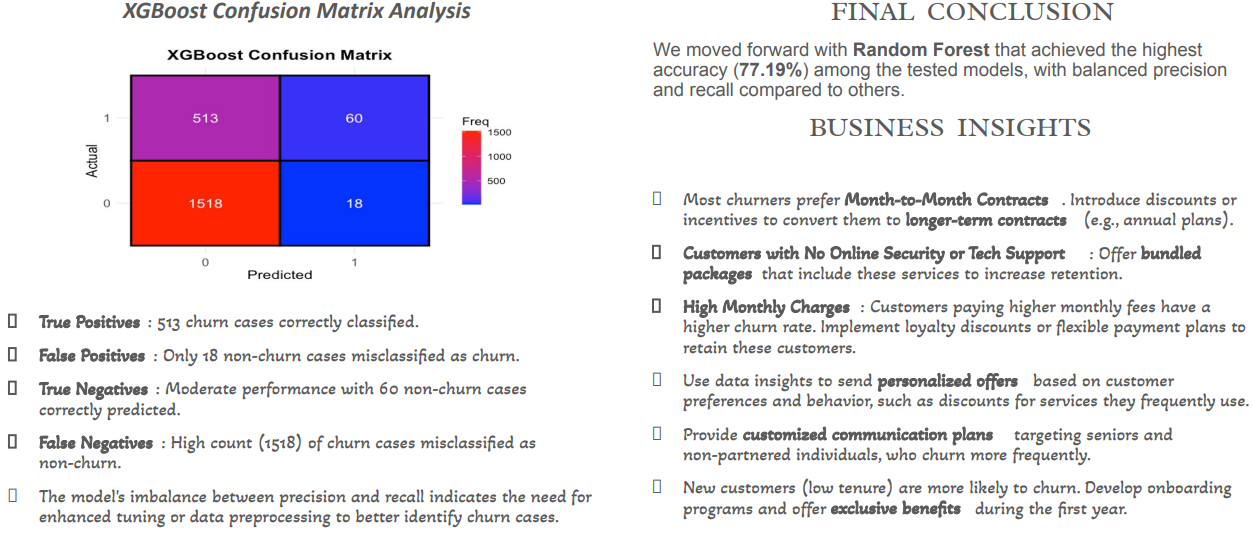
achieving an accuracy of 77.19%. According to my research, businesses should focus on

addressing these factors through targeted strategies, such as incentivizing long-term contracts,

enhancing service quality, and providing personalized retention offers. Implementing these

recommendations identified by my research can significantly reduce churn and improve

customer loyalty in the future moving forward.



**References**

Kaggle. (n.d.). Customer Churn Dataset. Retrieved from [Kaggle](<https://www.kaggle.com/>).

Scikit-learn Documentation. (n.d.). Retrieved from [https://scikit-learn.org](<https://scikit-learn.org>).

Python Data Analysis Library (Pandas). (n.d.). Retrieved from [https://pandas.pydata.org](<https://pandas.pydata.org>).

OpenAI. (2024). *ChatGPT*(Dec 10 version) [Large language model]. <https://chat.openai.com>

*Note: We used this GPT for only generating the theoretical explanations for my own Visualizations.*

*THE END*