TELCO CHURN PREDICTION REPORT

Contents

Introduction

Customer churn presents a significant challenge for telecommunications firms, impacting their revenue, acquisition costs, and market standing. It also affects brand reputation, customer lifetime value, and opportunities for upselling. Therefore, mitigating churn is crucial for financial stability, competitiveness, and maintaining investor trust.

The Problem

Customer churn is a pressing issue for telecommunications companies, as it directly impacts their financial performance, acquisition costs, and market share. Moreover, it tarnishes brand image, diminishes customer value, and hampers potential revenue from additional services. Consequently, addressing churn is imperative for ensuring financial resilience, market relevance, and sustaining investor confidence.

The Goal

This project seeks to build a predictive model capable of forecasting customer churn in a telecommunications company. Through the use of advanced analytics and machine learning techniques, the aim is to identify customers at risk of churning in advance. This proactive approach enables the implementation of targeted retention strategies, ultimately leading to a reduction in churn rates.

The Objectives

i. Extract meaningful insights from the provided dataset.

ii. Identify key factors contributing to customer churn.

iii. Develop a classification model to predict churn probability accurately.

iv. Refine the model to enhance its predictive performance.

v. Deploy the model within a web application for practical implementation and usage.

Data Description

The dataset consists of a sample size of 7043 and 21 features. Most features are categorical, apart from 3 features which are tenure, total charges and monthly charges. The dataset comprises details regarding:

1. Demographic Information: Encompasses gender, age group, and whether customers have partners or dependents.
2. Churn: Indicates whether customers have recently terminated their subscription or service.
3. Services: Includes phone, multiple lines, internet, online backup, device protection, online security, tech support, streaming TV, and Streaming movie features.
4. Customer Account Details: Covers tenure, contract type, payment method, paperless billing, monthly charges, and total charges.

Data Acquisition

Data was downloaded from Kaggle: (https://www.kaggle.com/datasets/blastchar/telco-customer-churn)

Data Preprocessing

Before proceeding with modeling, the data underwent several preprocessing steps, including:

❖ ***Formatting:*** Total Charges and Senior Citizen appeared to be an object data type and integer respectively. Therefore, it was converted to float data type.

❖ ***Handling missing values:*** NaN values were initially absent in the dataset. However, upon formatting, null values emerged in the total charges column, necessitating their removal.

❖ ***Handling Duplicated Values:*** Fortunately, there were no duplicated values observed in the dataset.

Exploratory Data Analysis (EDA)

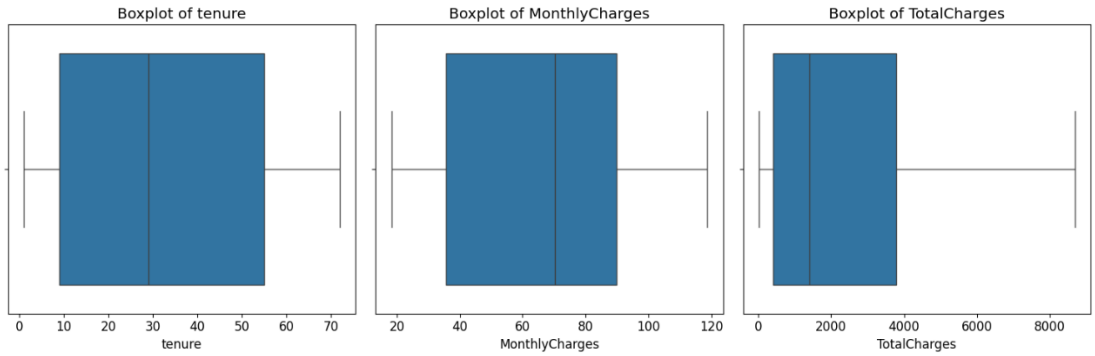
EDA was performed to gain insights into the data and understand the relationships between different features and the target variable. Some key observations from EDA include:

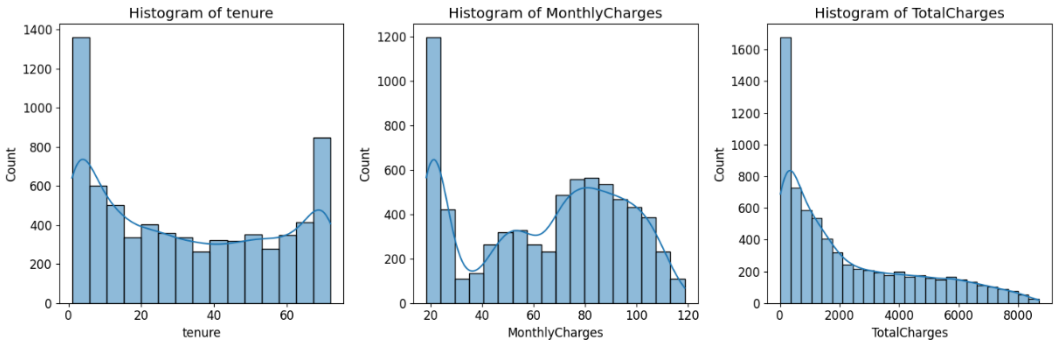
Summary Statistics

- The customer tenure ranges from a minimum of 1 month to a maximum of 72 months, with a mean tenure of 32 months.
- Total charges vary from approximately $19 to $8684, with a mean charge of approximately $2283.
- Monthly charges range from approximately $18 to $119, with a mean monthly charge of approximately $65.

Distribution

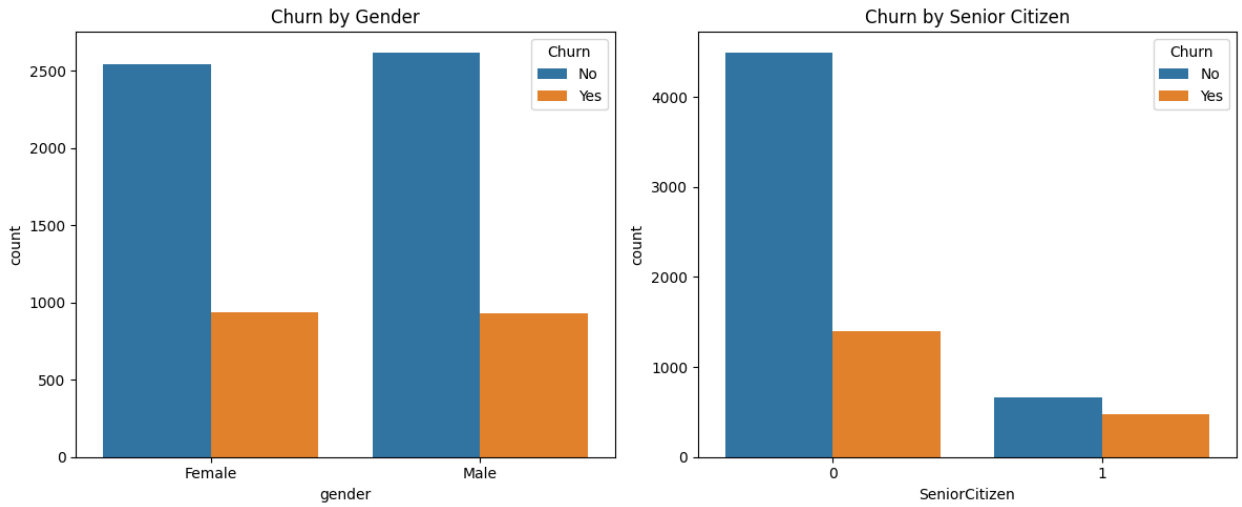
The numerical features show no outliers or extreme values, and most of the variables demonstrate right skewness.



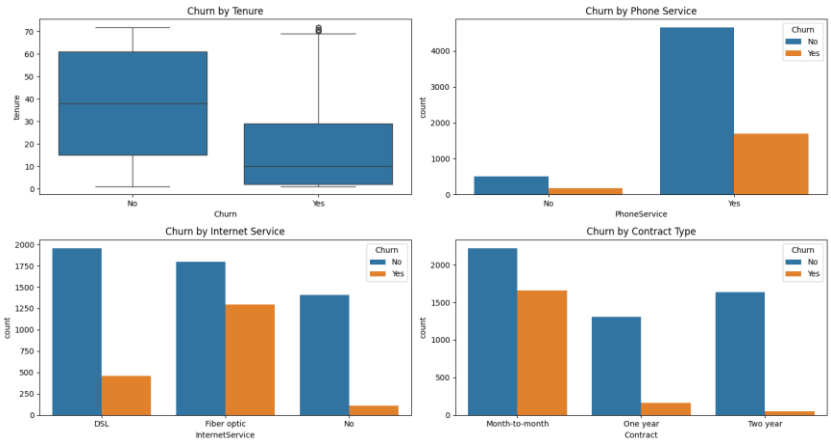


Insights

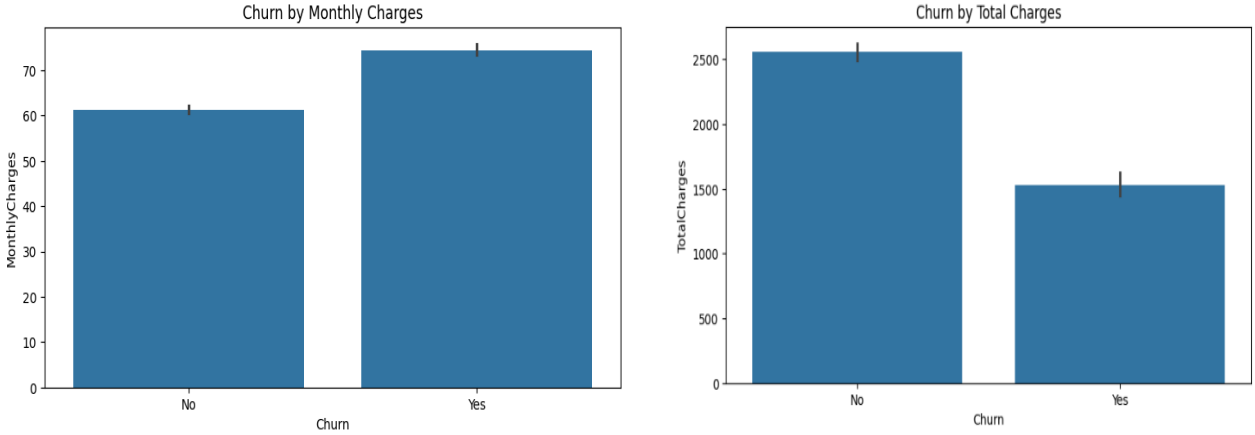
i. Female customers exhibit higher churn rates compared to male customers, and non-senior citizens display higher churn rates compared to senior citizens.



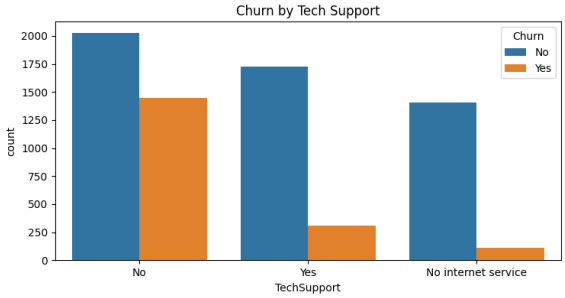
ii. The data indicates that customers who churned ("yes" churn) tend to have longer tenure with the company. Among customers with phone service, there is a higher incidence of churn compared to those without it. Similarly, customers with fiber optic internet service are more likely to churn, while those without any internet service tend to have lower churn rates. Additionally, customers on a month-to-month contract are more prone to churn compared to those with longer-term contracts, such as a 2-year agreement.

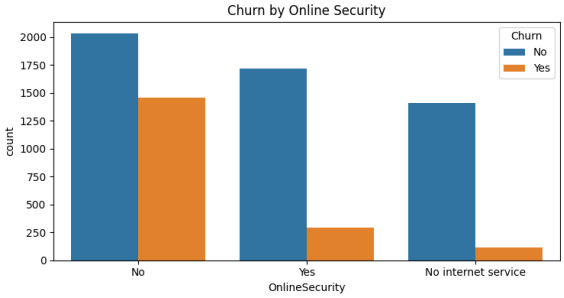


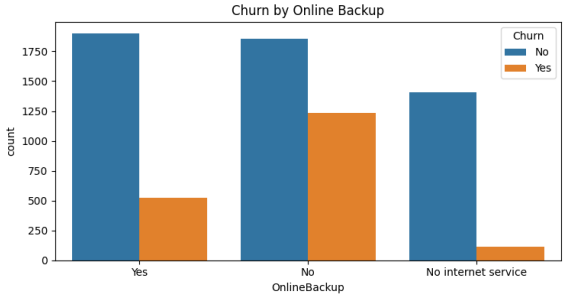
iii. The customers with high monthly charges and have low total charges higher churn rates.



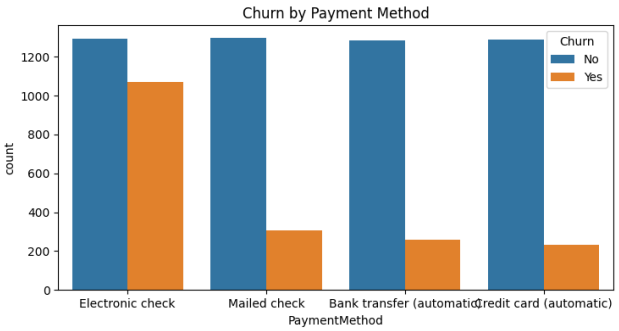
iv. Customers with no tech support, online security and online backup have a high churn rate while those with no internet service have low churn rates.

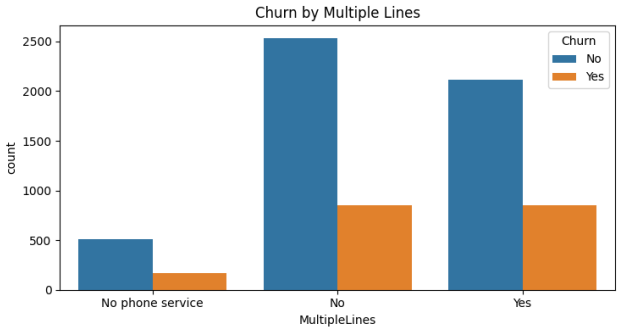


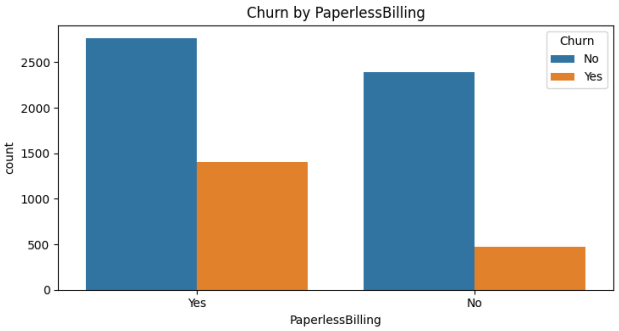




v. Customers using electronic check and paperless billing have a higher rate of churn. In addition, customers without multiple lines churn less.







- The EDA shows females and younger customers churn more. This could be due to different needs, price sensitivity, or marketing effectiveness (gender) and life stages, price awareness, or tech-savviness (age).
- Customers with longer tenures are more likely to churn. This could suggest that they were initially satisfied with the service or product but eventually decided to discontinue it after some time. Customers on month-to-month plans churn more than those locked into longer contracts. This suggests a commitment factor; short-term contracts offer flexibility but less incentive to stay.
- Customers with more services churn more. Phone users might have more options to switch if unhappy. Fiber users pay more and expect high quality, leading to churn if unmet. Those without internet have less need to switch, resulting in lower churn. This suggests churn might be linked to service usage and satisfaction with value for money.
- Customers who don't have tech support, online security, or backup might churn more due to price sensitivity or not seeing the value in these services. They might be comfortable managing without them or be more budget-conscious.
- Customers using electronic checks and paperless billing churn more. This could be due to payment issues or a preference for traditional billing methods. Customers with multiple lines have a higher need for the service and are more likely to churn if it's not met, whereas those without them might have a lower overall need and churn for different reasons.

Feature Engineering

Feature Selection and Extraction

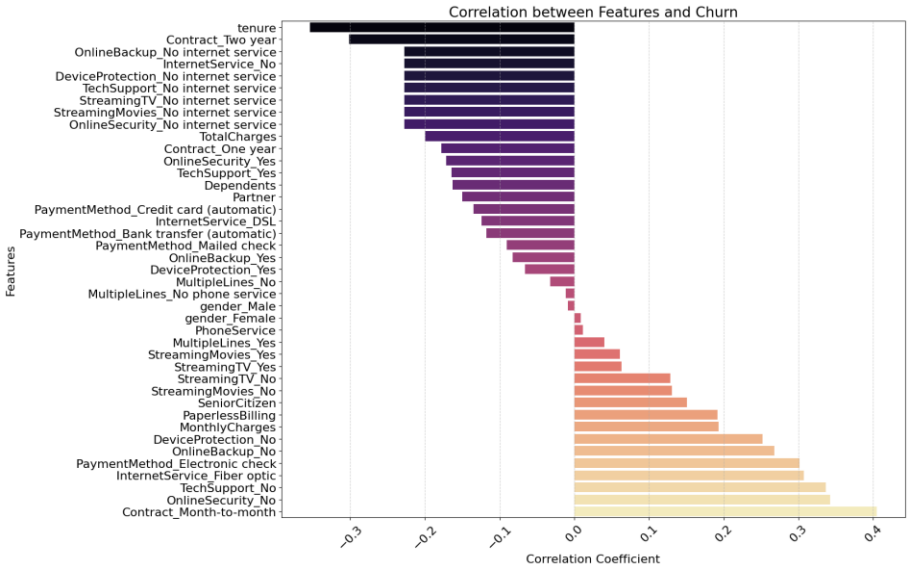
Involves identifying and choosing the most relevant features from the dataset for model training.

Conversion of categorical features to numerical

The features 'Churn', 'Partner', 'Dependents', 'Paperless Billing', and 'Phone Service' were transformed into Boolean data types, where 'Yes' and 'No' were mapped to True and False, respectively. Additionally, dummy

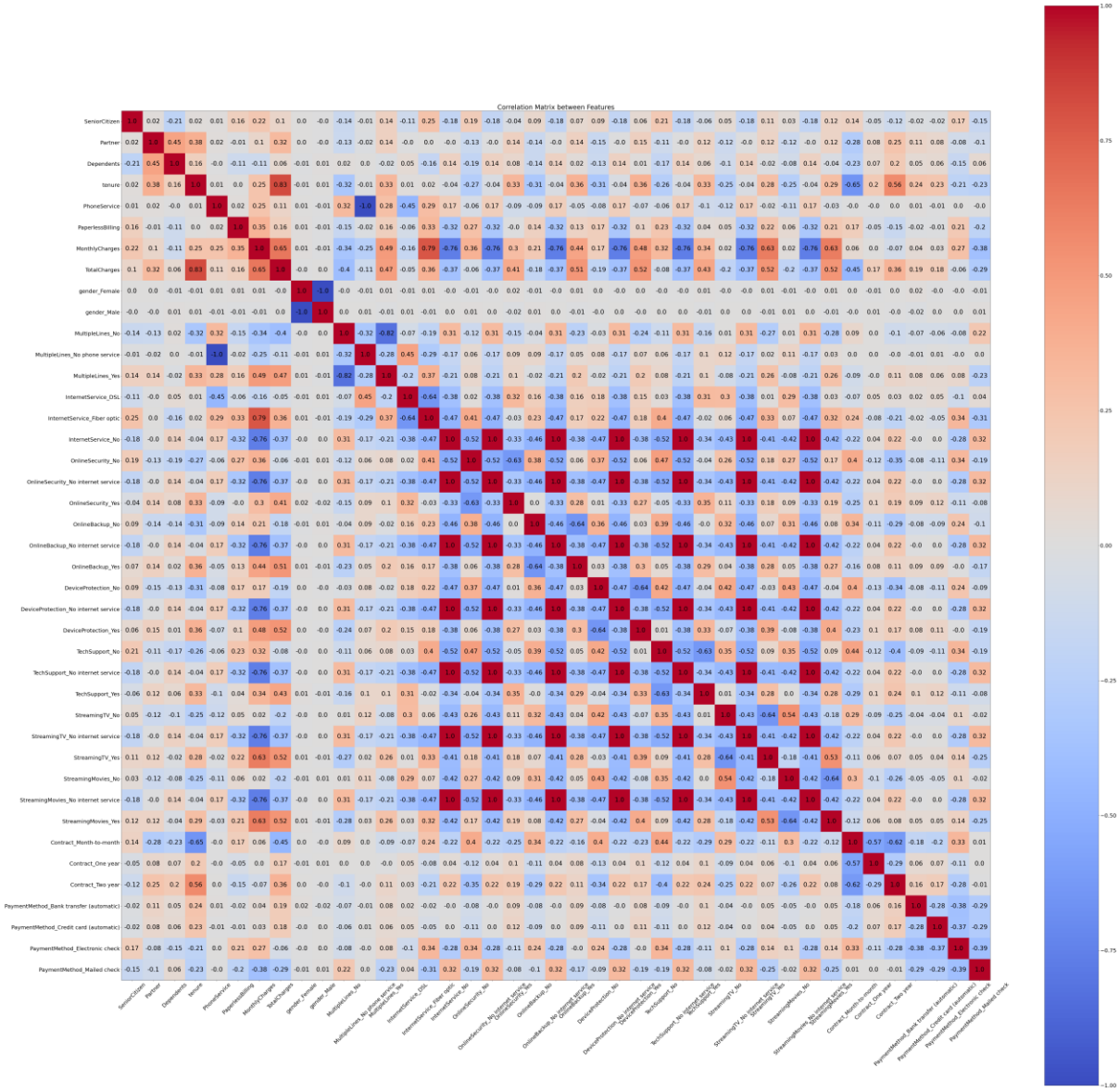
variables were created in the dataset to convert features with Boolean and object data types into numerical representations.

Feature Correlation with the Target Variable



- The feature most positively correlated with "Churn" is "Contract_Month-to-month"
- The feature most negatively correlated with "Churn" is "Tenure".
- The features containing "no internet service" are redundant.

Multi-collinearity



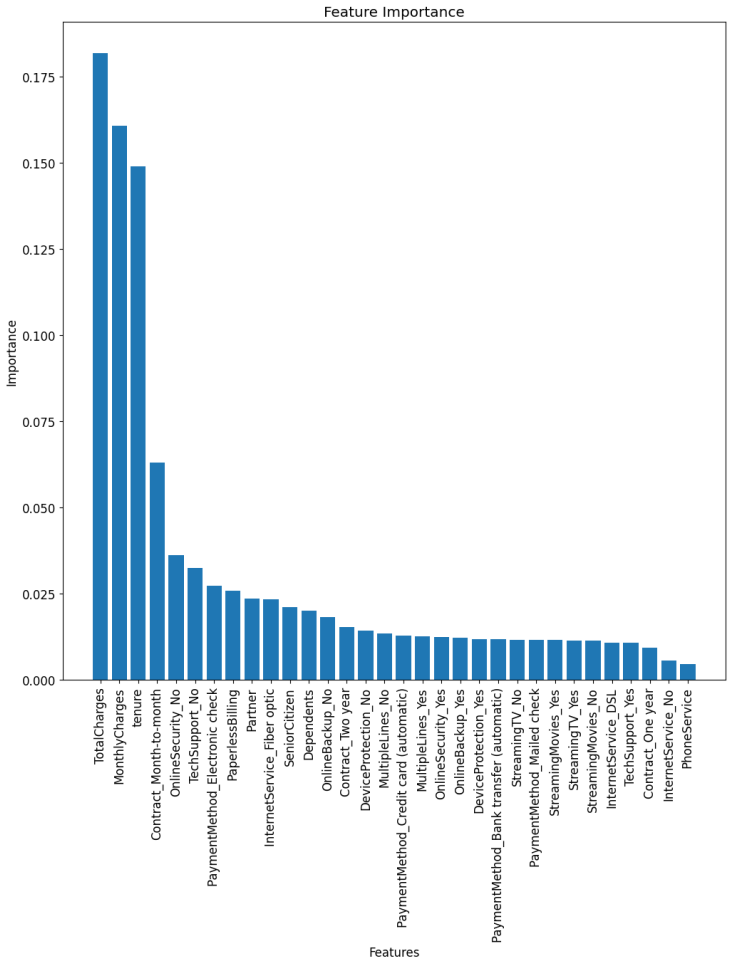
Total Charges is strongly positively correlated with Tenure, while Monthly Charges exhibit a strong positive correlation with Internet Service Fiber Optic. However, tenure, monthly and total charges, and fiber optic are highly significant predictors of churn.

Out of the 7 features containing "no internet service", we will retain only the " No Internet Service" feature as the others are redundant.

Dropping Columns

The features associated with 'no internet service' and those with weak correlations to churn, which are deemed insignificant, were removed from the dataset. Additionally, Customer ID was excluded from the dataset during an earlier stage.

Feature Importance



SelectKBest was also applied to extract top 16 best features

	Feature	Score
7	TotalCharges	629630.8103
3	tenure	16377.32809
6	MonthlyCharges	3653.074681
25	Contract_Month-to-month	516.714004
27	Contract_Two year	486.223101
30	PaymentMethod_Electronic check	424.113152
13	OnlineSecurity_No	414.036636
19	TechSupport_No	404.010811
11	InternetService_Fiber optic	372.082851
12	InternetService_No	285.475152
15	OnlineBackup_No	282.490201
17	DeviceProtection_No	250.189168
26	Contract_One year	176.608724
14	OnlineSecurity_Yes	147.165601
20	TechSupport_Yes	135.439602
0	SeniorCitizen	133.482766

Features of low importance and had low scores were dropped.

Feature Scaling

Splitting the dataset

The dataset was split into training and testing sets. In addition, the training data set is further split into training and validation set.

Balancing the dataset

The training and testing data was balanced using SMOTE, a technique that synthesizes new instances of the minority class to equalize class distribution. Balancing the dataset aids in training the model by preventing bias towards any particular class. This approach ensures that the model does not favor the majority class simply due to its larger representation in the data.

Standardizing the data

The original dataset was standardized to follow a normal distribution, that is, have a mean 0 and a variance of 1.

Model Selection

Several machine learning models were used and evaluated on the dataset. The following models were considered: Logistic Regression, Random Forest Classifier, Ada Boost Classifier, K Neighbors Classifier, Gradient Boosting Classifier and XG Boost Classifier.

Model Training and Evaluation

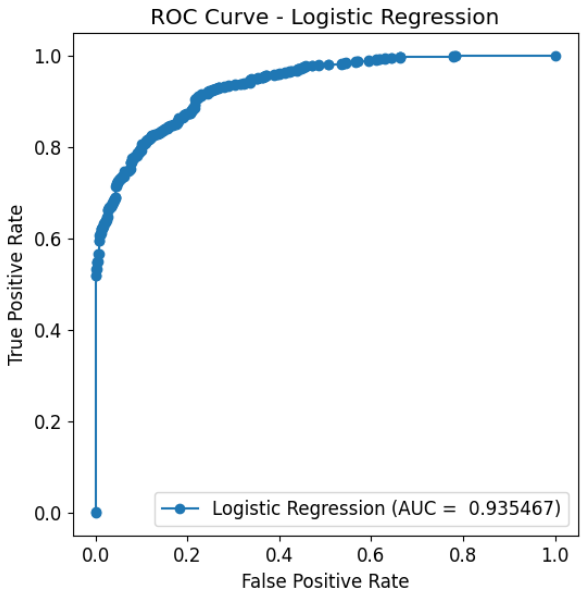
The performance of each model was evaluated using appropriate evaluation metrics. The evaluation metrics used are:

i. Accuracy Score
ii. Classification Report
iii. Confusion Matrix
iv. ROC AUC score,
v. ROC curve

Results

Based on our findings, Logistic Regression demonstrated superior performance compared to other models. It achieved an accuracy score of 0.84 and an AUC-ROC value of 0.935467 using the validation data, indicating its superiority relative to the other models evaluated in the study.

The model was further evaluated using an unseen data(test data), it achieved an accuracy score of 0.79 and an AUC-ROC value of 0.825564.



From the output provided above, it's evident that our model's prediction accuracy isn't perfect, but it demonstrates optimal performance.

In summary, the modeling study showcased the superiority of Logistic Regression over other models regarding predictive accuracy. This is highlighted by the Logistic Regression model's impressive performance, particularly in its ROC-AUC score and ROC curve, indicating heightened precision in predictions. Hence, considering the evaluation metrics provided, the Logistic Regression model emerges as the preferred option for churn prediction. A key takeaway from this analysis underscores the critical role of Total Charges, Monthly Charges, Tenure, and Contract as fundamental factors influencing churn rates for the telecommunication company.