Report on SyriaTel Customer Churn

1.Executive Summary

This report examines the factors influencing customer churn for SyriaTel, a leading telecommunications provider. Using data analysis and machine learning models, the project identifies key predictors of churn, evaluates model performance, and suggests actionable strategies to reduce customer attrition. The study found that features such as customer service calls, tenure, and monthly charges significantly impact churn. The Decision Tree model was optimized and demonstrated the highest accuracy among tested algorithms.

2. Introduction

Project Overview

This project focuses on predicting customer churn for SyriaTel, a leading telecommunications company.Using historical customer data from the bigml_59c28831336c6604c800002a.csv dataset we aim to uncover patterns and behaviors associated with churn. Through data analysis and predictive modelling, we will identify key indicators of churn and build a machine learning classifier to predict whether a customer is likely to churn. With these insights, SyriaTel can take proactive steps to address potential customer dissatisfaction and enhance retention strategies

Background

Customer churn is a critical challenge for telecommunications companies. Retaining customers is more cost-effective than acquiring new ones, making churn prediction vital for profitability and growth.

3.Business Understanding

In the telecommunications industry, customer churn poses a significant challenge for companies like SyriaTel. The objective is to develop a model that predicts whether a customer will soon terminate their services with SyriaTel. This binary classification task aims to uncover patterns in customer behavior and demographic data that may indicate a propensity to churn. The ultimate goal is to aid SyriaTel in reducing the financial impact of customer churn by implementing proactive retention strategies.

Objectives

1. Identify the key factors driving customer churn.
2. Develop predictive models to forecast churn with high accuracy.
3. Provide actionable recommendations to mitigate churn.

Scope

This project focuses on SyriaTel's customer dataset, analyzing variables such as tenure, payment methods, and customer service interactions to predict churn.

Business overview

SyriaTel is one of the leading telecommunications companies, providing a range of services such as mobile telephony, broadband internet, and data services. The company caters to a large and diverse customer base, ranging from individuals to businesses. Given the competitive and dynamic nature of the telecommunications market, maintaining a loyal customer base is critical for SyriaTel's continued growth and profitability.

In the telecommunications sector, customer retention is as important as customer acquisition. However, customer churn (the loss of subscribers) is a common challenge that every telecom company faces. Understanding the causes of churn and developing effective strategies to minimize it is essential for enhancing customer satisfaction, reducing marketing costs, and improving overall business performance.

Problem Statement

SyriaTel faces the challenge of retaining its customer base amidst a competitive telecommunications landscape. Customer churn not only leads to revenue loss but also affects the company's reputation and market position.

The business problem

SyriaTel wants to identify customers likely to churn and understand the factors driving their decision.

4. Metrics of success¶

Churn Rate: Measures the percentage of customers who leave SyriaTel.

Retention Rate: The percentage of customers retained over a period. Higher retention means better customer satisfaction.

Accuracy of Churn Model: Measures how well the churn prediction model identifies at-risk customers.

Precision & Recall: Measures how accurately the model identifies churners and avoids false predictions.

AUC-ROC Score: Indicates how well the model distinguishes between churners and non-churners. Higher AUC means better model performance.

Customer Satisfaction : A score that reflects how satisfied customers are, influencing retention.

By focusing on these metrics, SyriaTel can gauge the effectiveness of its churn management strategies and improve customer retention.

5. Data Understanding

The Churn in Telecom’s dataset from Kaggle contains information about customer activity and whether or not they canceled their subscription with the Telecom firm. The goal of this dataset is to develop predictive models that can help the telecom business reduce the amount of money lost due to customers who don’t stick around for very long.

The dataset contains 3333 entries and 21 columns, including information about the state, account length, area code, phone number, international plan, voice mail plan, number of voice mail messages, total day minutes, total day calls, total day charge, total evening minutes, total evening calls, total evening charge, total night minutes, total night calls, total night charge, total international minutes, total international calls, total international charge, customer service calls and churn.

In this phase of the project, we will focus on getting familiar with the data and identifying any potential data quality issues. We will also perform some initial exploratory data analysis to discover first insights into the data.

Summary of features in the dataset:

state : Different states of the customers

account length: number of days a customer's account has been active

area code : location of the customer

phone number : customer's phone number

international plan : whether the customer uses the international plan or not

voice mail plan : whether the customer has subscribed to vmail plan or not

number vmail messages : if customer has a vmail plan, how many vmail messages do they get

total day minutes : total number of call minutes used during the day

total day calls : total number of calls made during the day

total day charge : total charge on day calls

total eve minutes : total number of call minutes used in the evening

total eve calls : total calls made in the evening

total eve charge : total charge on evening calls

total night minutes: Total number of call minutes used at night

total night calls : Total number of night calls

total night charge : Total charge on night calls

total intl minutes : total international minutes used

total intl calls : total number of international calls made

total intl charge : total charge on international calls

3. Methodology

Data Collection

- The dataset consists of customer demographics, account information, and service usage details.
- Key variables include tenure, monthly charges, total charges, and churn status.

Data Preparation

Import all the necessary libraries

Load the dataset using pandas library

Cleaned missing values and outliers.

Encoded categorical variables.

Standardized numerical features for consistency.

Modeling Approach

1. **Exploratory Data Analysis (EDA):** Identified trends and relationships.
2. **Machine Learning Models:** Tested Decision Tree, Random Forest, and Logistic Regression.
3. **Hyperparameter Tuning:** Used GridSearchCV for optimization.

---

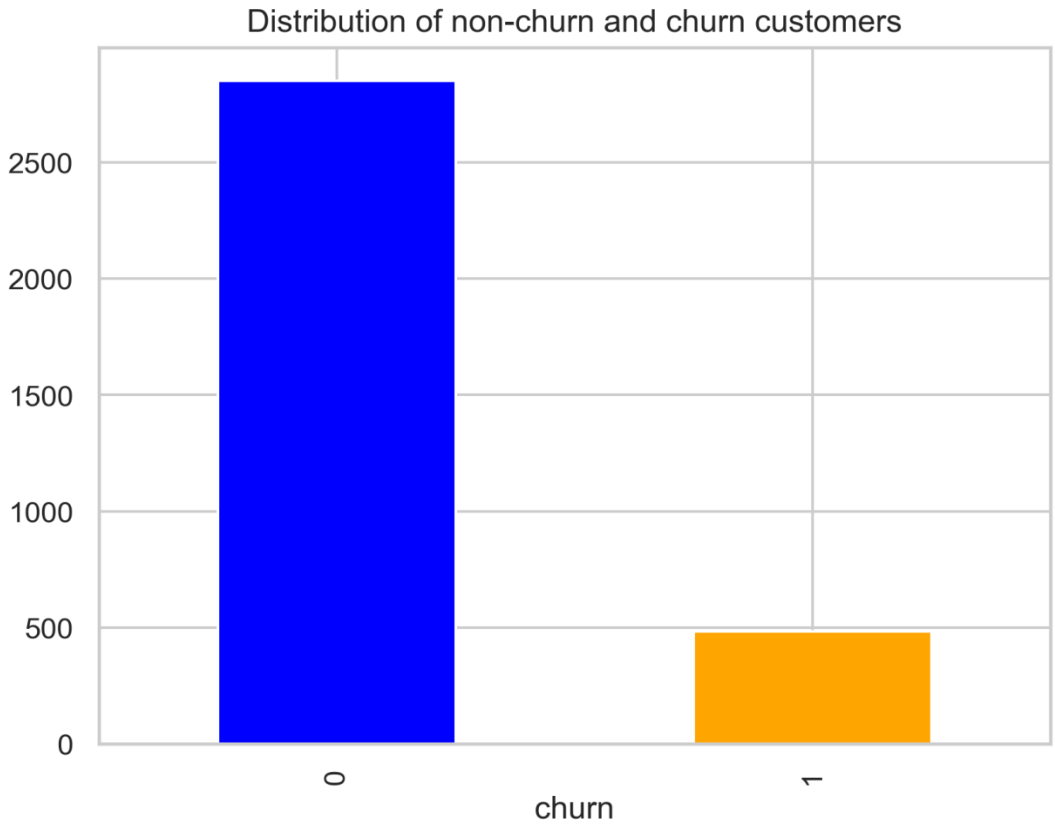
1. **Exploratory Data Analysis (EDA)**

We are going to conduct a comprehensive exploration of the data through univariate, bivariate, and multivariate analysis.

Univariate Analysis

Univariate analysis involves examining a single variable at a time to understand its distribution, central tendency, and variability.

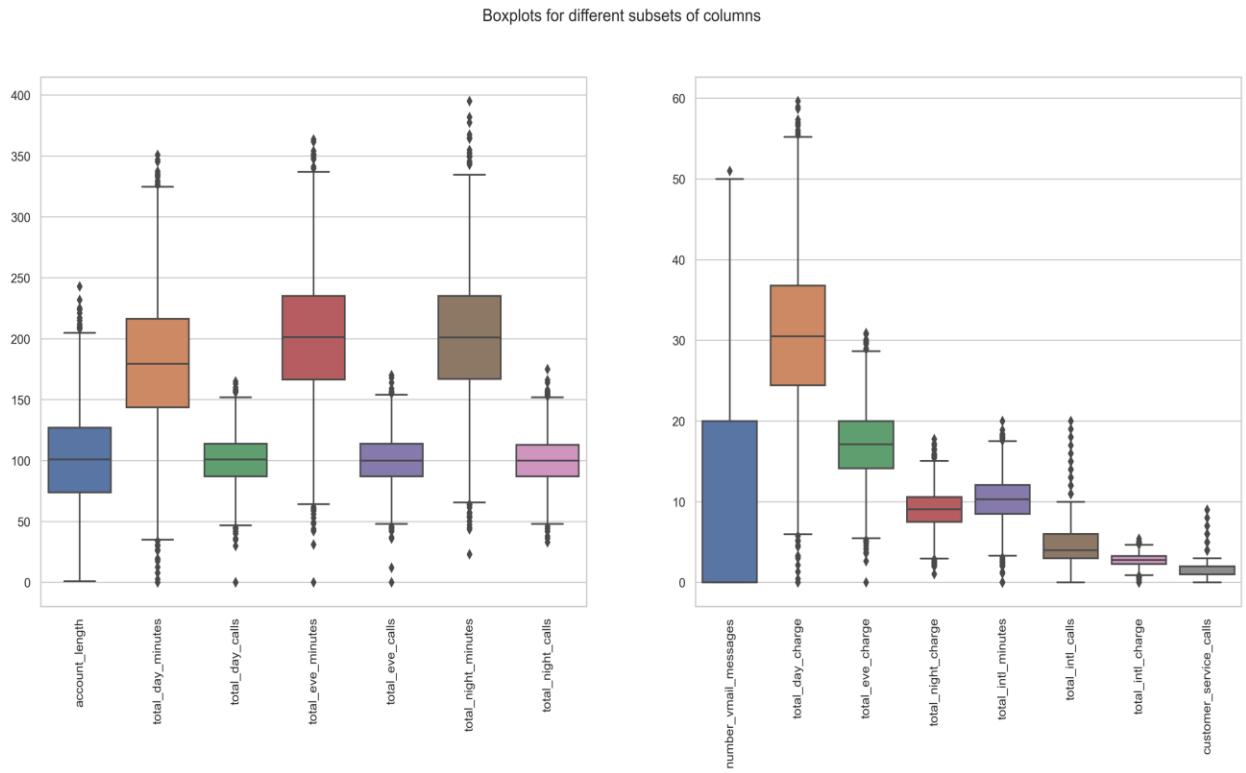
There are 2850 false values, which indicates the number of clients who did not churn. There are also 483 true values, showing the number of clients who left the the company.



In the graph above, it shows that there are more unchurned customers than churned ones,483, who terminated the contract.This shows a data imbalance.

Outliers can significantly impact the performance of machine learning models, which will impacts the feature engineering process.

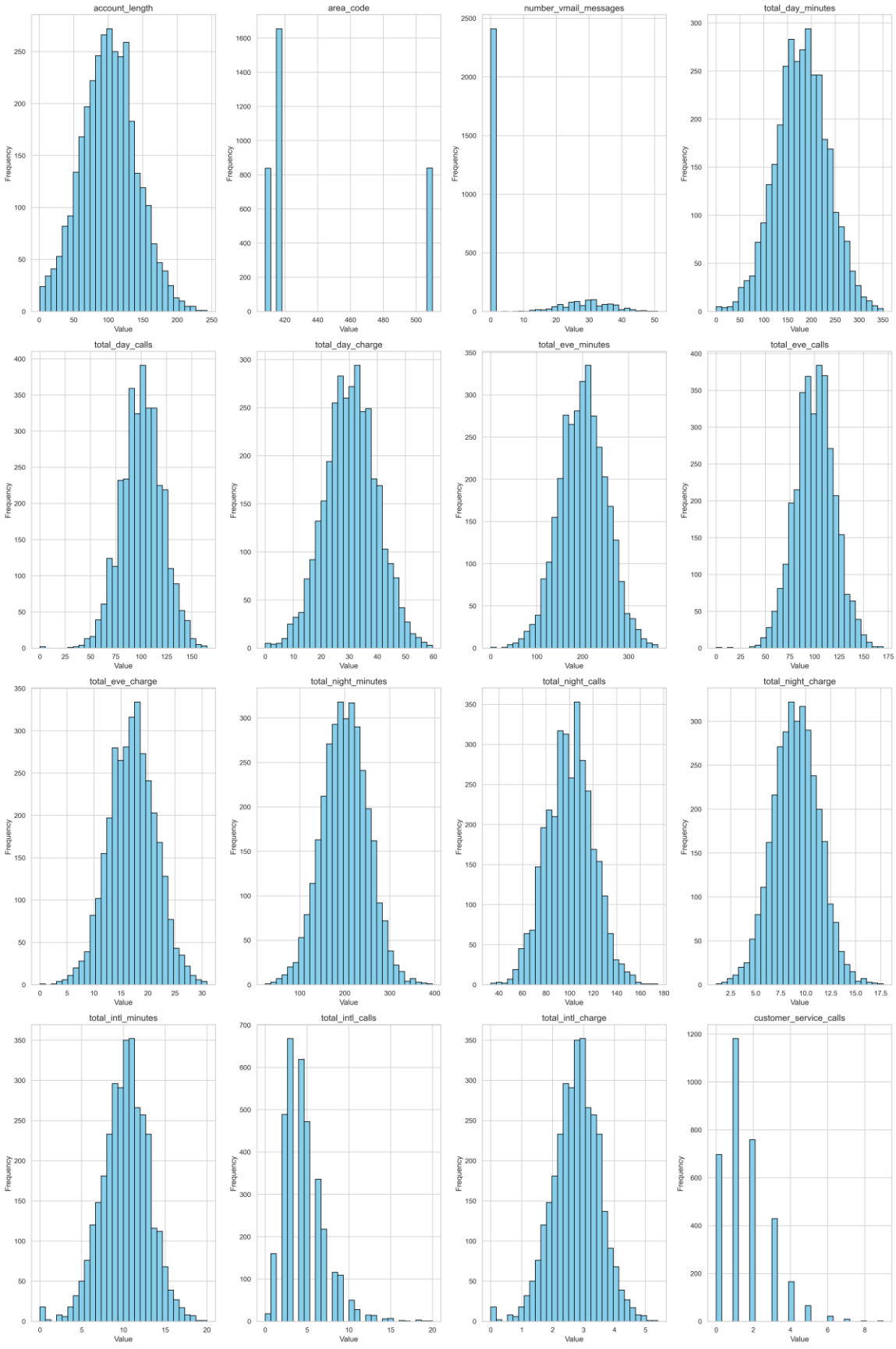
[1591]:



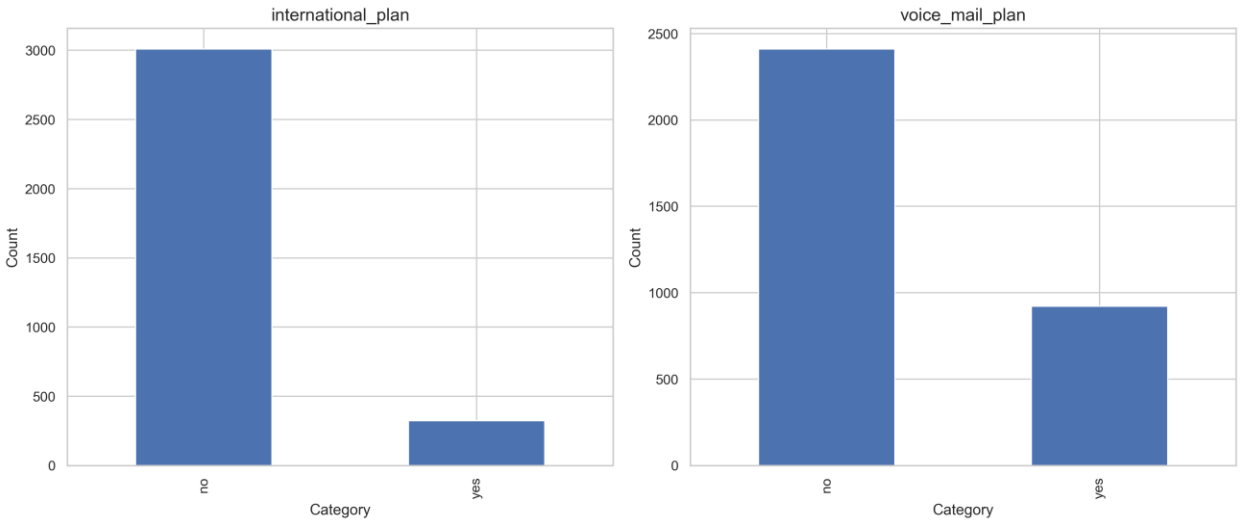
In box boxplots, we can see that the columns have numerous outliers, which may affect the performance of machine learning models such as k-nearest neighbors (knn).

As for our data, all these outliers contain valuable information, which will be very important to our models. As such, we will not be eliminating data.

Numerical dataset


account_length
area_code
number_vmail_messages
total_day_minutes
total_day_calls
total_day_charge
total_eve_minutes
total_eve_calls
total_eve_charge
total_night_minutes
total_night_calls
total_night_charge
total_intl_minutes
total_intl_calls
total_intl_charge
customer_service_calls
Frequency
Value

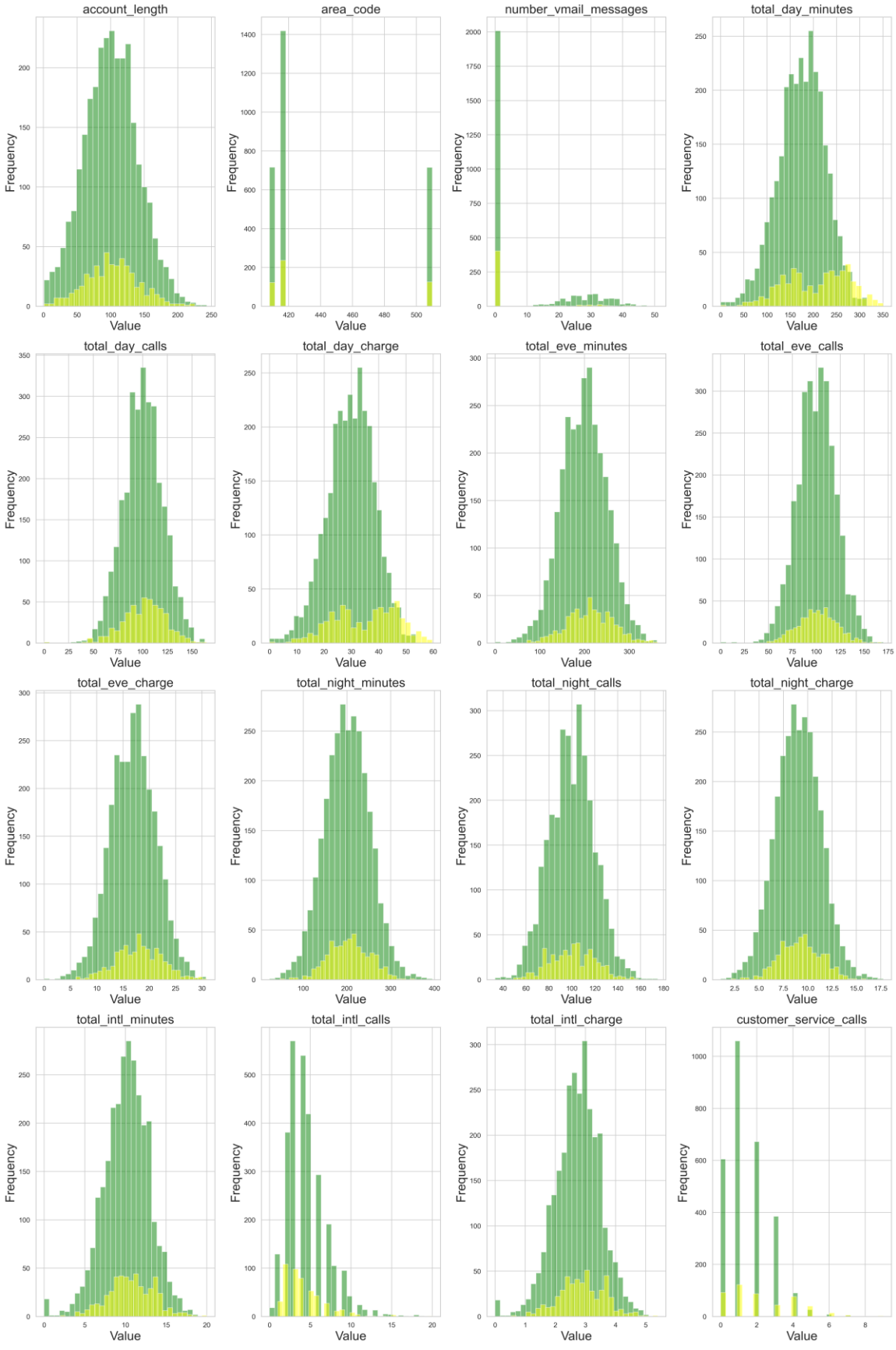
Categorical dataset



Bivariate analysis

Bivariate analysis involves analyzing the relationship between two variables. For our project, we examine the relationship between each feature and the target variable (customer churn) to understand how they are related.

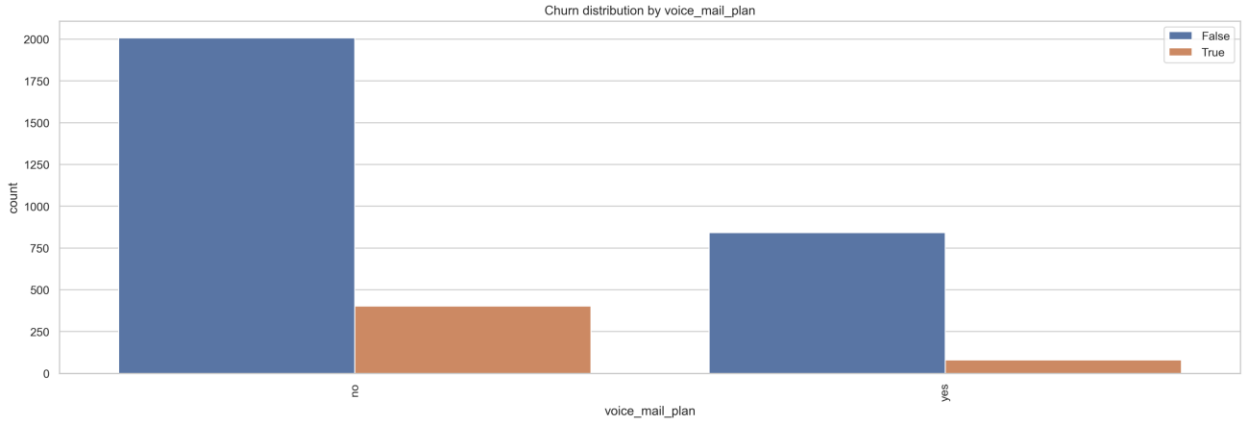
Numerical dataset

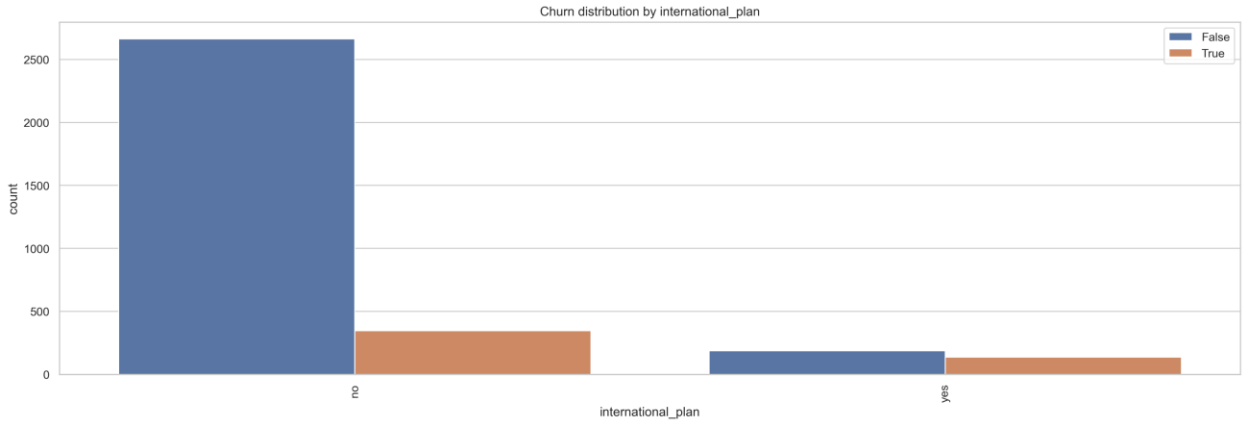

account_length
area_code
number_vmail_messages
total_day_minutes
total_day_calls
total_day_charge
total_eve_minutes
total_eve_calls
total_eve_charge
total_night_minutes
total_night_calls
total_night_charge
total_intl_minutes
total_intl_calls
total_intl_charge
customer_service_calls
Frequency
Value

There seems to be strong relationship between customer service calls and true churn values. After 4 calls, customers are a lot more likely to discontinue their service. Besides, most customer calls are are associated with disatisfaction with customer service. At this point more than 4 customer calls indicate that it takes long for their issues to be addressed, and thus a possibility of them leaving increases.

Categorical dataset

Here, we are doing some analysis of the customer churning in relation to international plan, and voice mail plan. We are trying to understand whether there are correlations between the categorical columns and the customer churning





Multiveriate Analysis

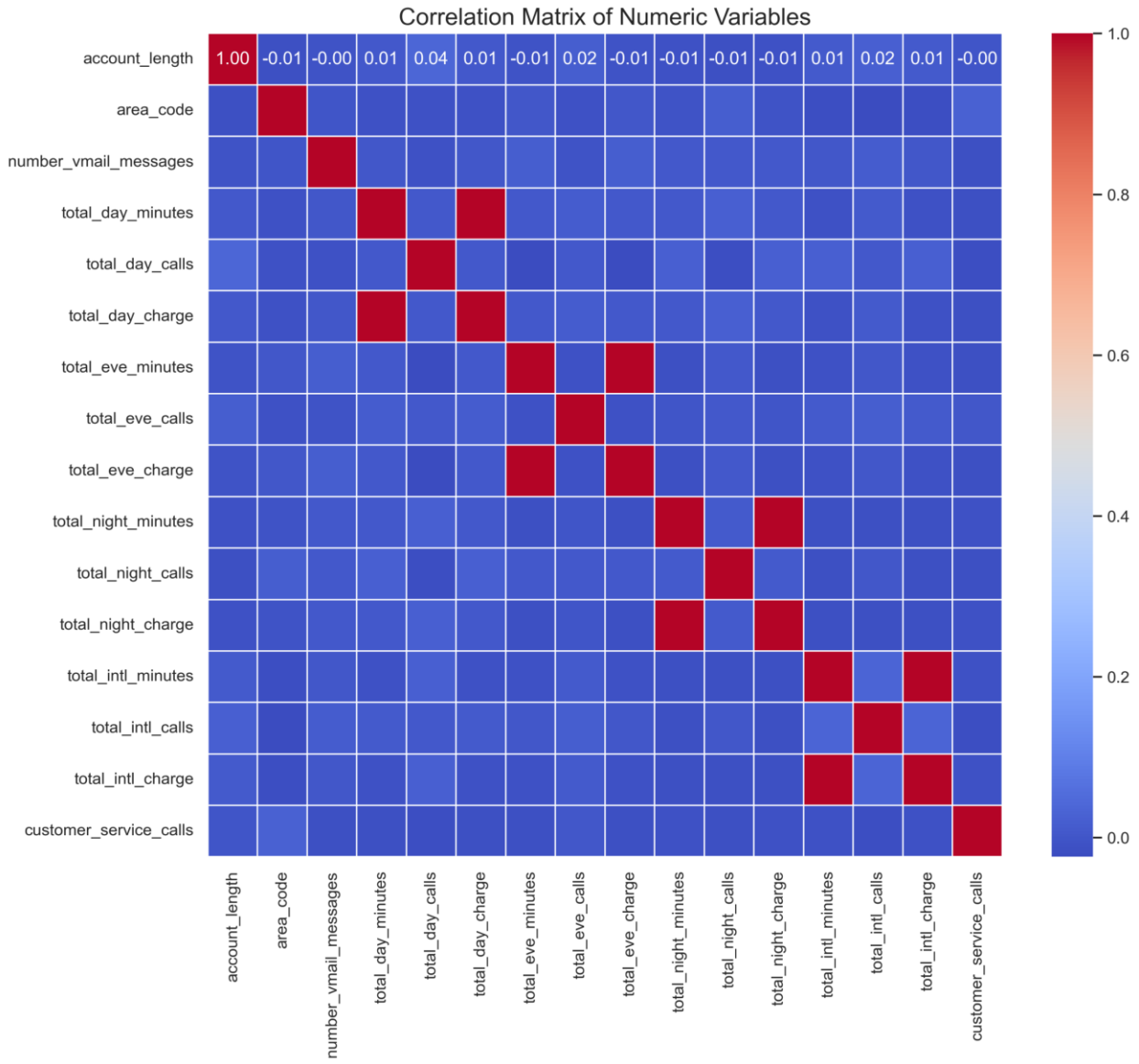
Multivariate analysis involves analyzing the relationship between multiple variables simultaneously. In this case, we explore the relationship between multiple features and the target variable (customer churn) to understand how they are related when considered together.

We used a correlation matrix to identify the correlation between different variables in the dataset.

-There is a very low correlation between most features.

-However, there is a perfect positive correlation between total evening charge and total evening minutes, total day charge and total day minutes, total night charge and total night minutes, and total international charge and total international minutes. This is expected since the charge of a call depends on the length of the call in minutes. One correlated variable will have to be dropped from each pair to handle multicollinearity. total day minutes, total day charge and customer service calls have a weak positive correlation with churn.

-The other features have a negligible correlation with churn, approximately 0.



-There is a very low correlation between most features.

-However, there is a perfect positive correlation between total evening charge and total evening minutes, total day charge and total day minutes, total night charge and total night minutes, and total international charge and total international minutes. This is expected since the charge of a call depends on the length of the call in minutes. One correlated variable will have to be dropped from each pair to handle multicollinearity. total day minutes, total day charge and customer service calls have a weak positive correlation with churn.

-The other features have a negligible correlation with churn, approximately 0.

Data Preprocessing

In this section, we proprocess the data to prepare it for modelling. In the dataset, we have categorical and numeric data columns, some of which must be tranformed into a datatype acceptable by the different machine learning models used in the modelling section.

The dataset must also be split into different sets, the training and testing sets. We will use the training set to train the different models and evaluate the performance using the test data. Cross-validation is used.

We also drop features that have minimal or no effect on the target variables using ridge or lasso regression. We may also identify other frameworks for choosing the best features.

step 1 : Transform columns to numeric¶

Step 2: Separate features and target variable

Step 3: Conduct a Train-test-split on the data

Step 4: Scaling

Step 5: SMOTE

SMOTE is used to handle class imbalance problems by oversampling the minority class with replacement

Creating our models

We create several models, evaluate them, then do some hyper-parameter tuning to try and improve the models. Our intention in this case is to find the model and parameters that perform the best.

We train and evaluate the following models:

Logistic Regression Model, Decision Trees and Random Forests.

Model 1: Logistic Regression Model

Logistic regression is a type of generalized linear model that can be used to predict the probability of a binary outcome, such as whether a customer will churn or not.

In our case, we use logistic regression to model the relationship between the our features and the likelihood of a customer churning.

Train Accuracy: 0.81
Test Accuracy: 0.82

Classification Report (Test Set):

	precision	recall	f1-score	support
0	0.93	0.85	0.89	566
1	0.43	0.65	0.52	101
accuracy			0.82	667
macro avg	0.68	0.75	0.70	667
weighted avg	0.86	0.82	0.83	667

Confusion Matrix (Test Set):
[[480 86]

[35 66]]

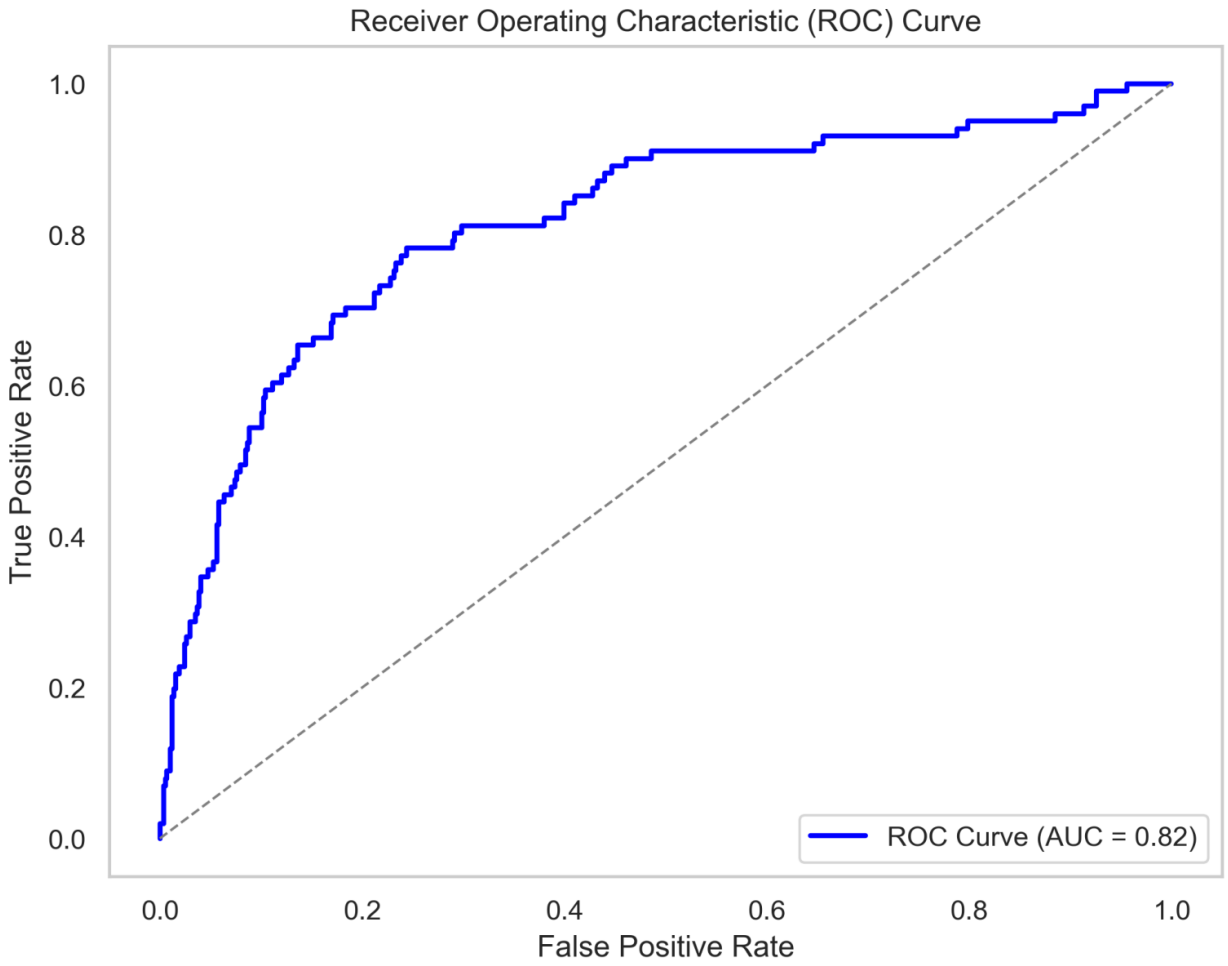
Comments and notes on model Accuracy: The accuracy of the model is 82% Train Accuracy: 0.81 Test Accuracy: 0.82

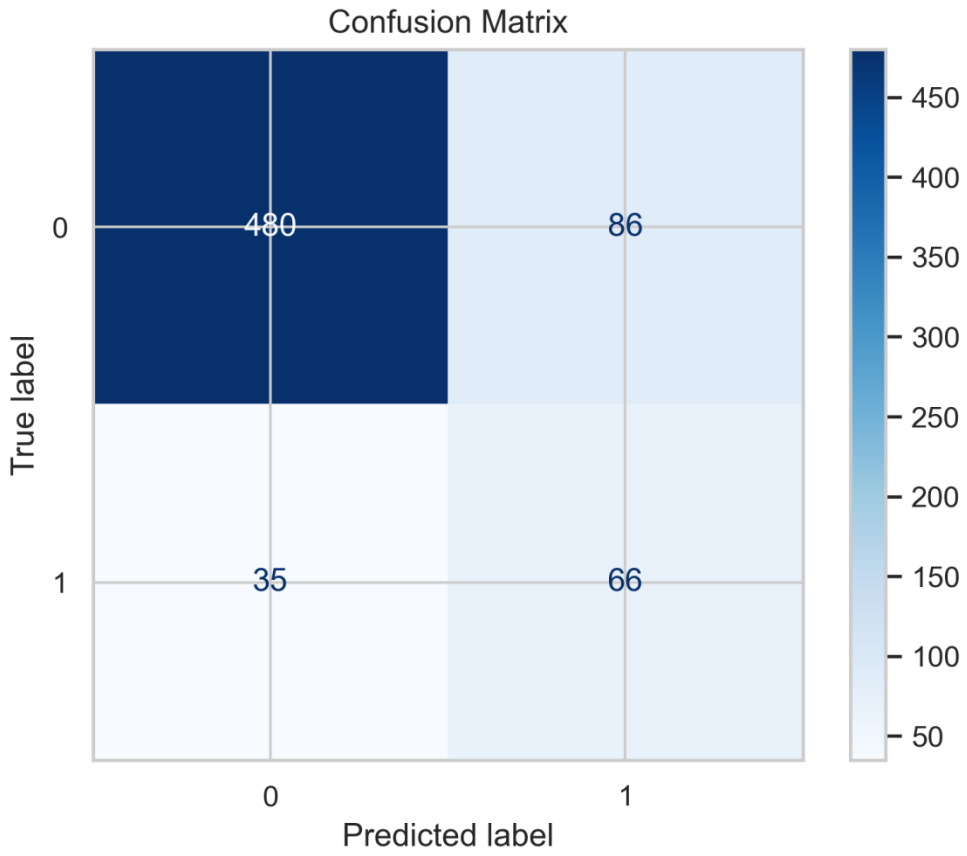
Classification Report:

Precision: The precision for class 0 (not churned) is 93%. The precision for class 1 (churned) is 43%

Recall: The recall for class 0 (not churned) is 85% but the recall for class 1 (churned) is only 65%.

F1-score: The F1-score for class 0 (not churned) is 89% and for class 1Co (churned) is only 52%. The F1-score for class 1 is low due to the low recall.


Receiver Operating Characteristic (ROC) Curve
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC Curve (AUC = 0.82)



The confusion matrix shows a total of 667 samples in the test set.

True Positives (TP): The model correctly predicted 66 samples as Not churned (class 0).

True Negatives (TN): The model correctly predicted 480 samples as churned (class 1).

False Positives (FP): The model incorrectly predicted 86 samples as churned when they were not churned.

False Negatives (FN): The model incorrectly predicted 35 samples as not churned when they were churned.

The ROC curve & The AUC

They provide a measure of how well the model can distinguish between positive and negative samples. A model with an AUC of 1 is perfect, while an AUC of 0.5 indicates that the model is no better than random guessing.

AUC = 0.5: The model's performance is equivalent to random guessing, and it is not useful for classification.

AUC > 0.5: The model performs better than random guessing, and the higher the AUC, the better the model's discriminatory power.

AUC = 1: The model perfectly distinguishes between positive and negative samples, making it an excellent classifier.

The AUC is 0.83, which is greater than 0.5 and closer to 1. An AUC of 0.83 suggests that the model has a good ability to rank the predictions, and it performs significantly better than random guessing.

Intepretation :

The model performs well in predicting the negative class (not churned) as evidenced by high accuracy, precision, and recall for class 0. However, it performs poorly for the positive class (churned) as indicated by the low values for precision, recall, and F1-score for class 1.The model is missing a substantial number of customers who are actually churned, leading to false negatives. It is failing to correctly identify those customers who have churned

Model 2 : Decision Trees¶

Decision Tree Train Accuracy: 1.00

Decision Tree Test Accuracy: 0.87

Classification Report (Test Data):

	precision	recall	f1-score	support
0	0.95	0.90	0.92	566
1	0.56	0.74	0.64	101
accuracy			0.87	667
macro avg	0.76	0.82	0.78	667
weighted avg	0.89	0.87	0.88	667

Confusion Matrix (Test Set):

[[507 59]
[26 75]]

Comments and notes on model Accuracy: The accuracy of the model is 87% Train Accuracy: 1.00 Test Accuracy: 0.87

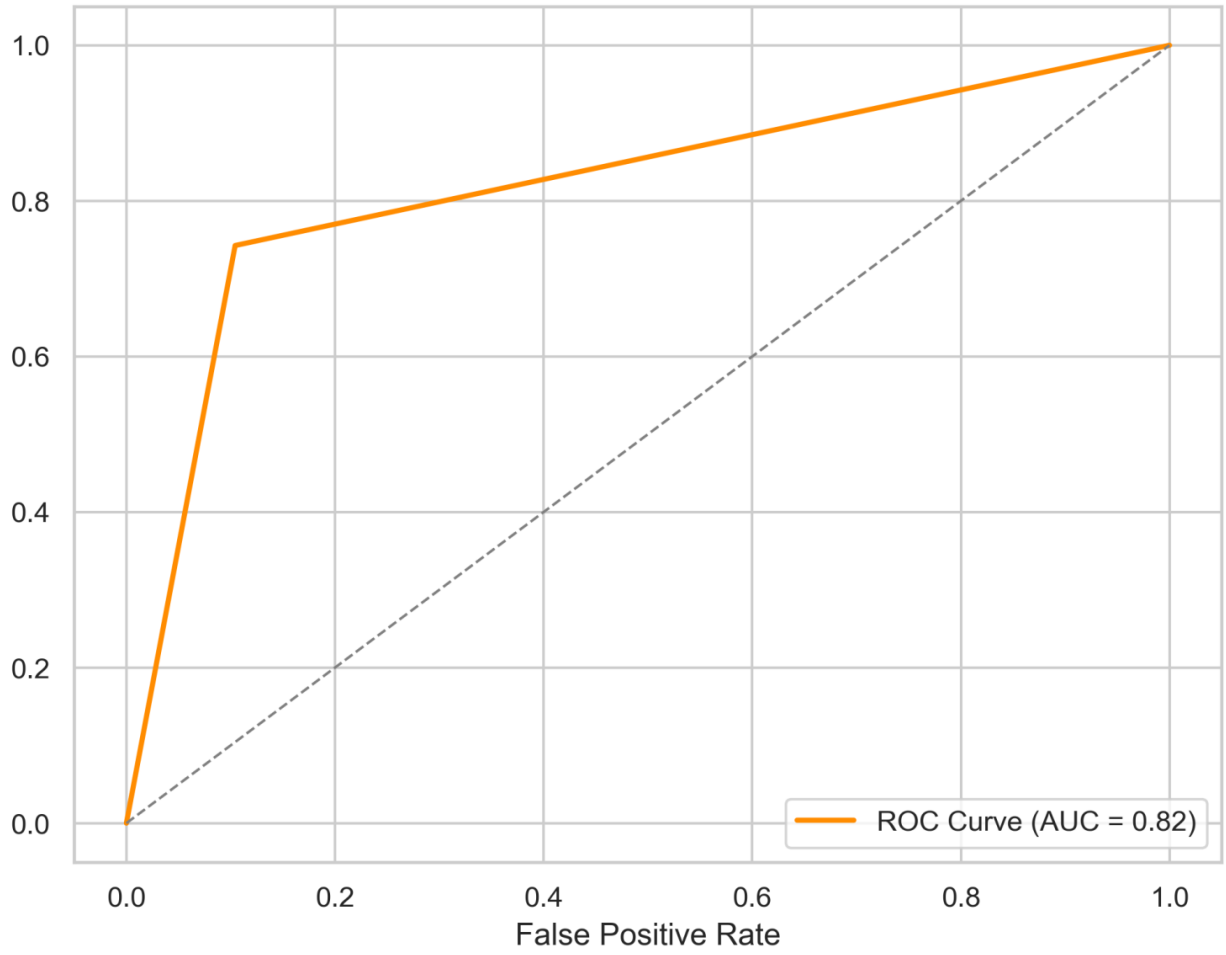
Classification Report:

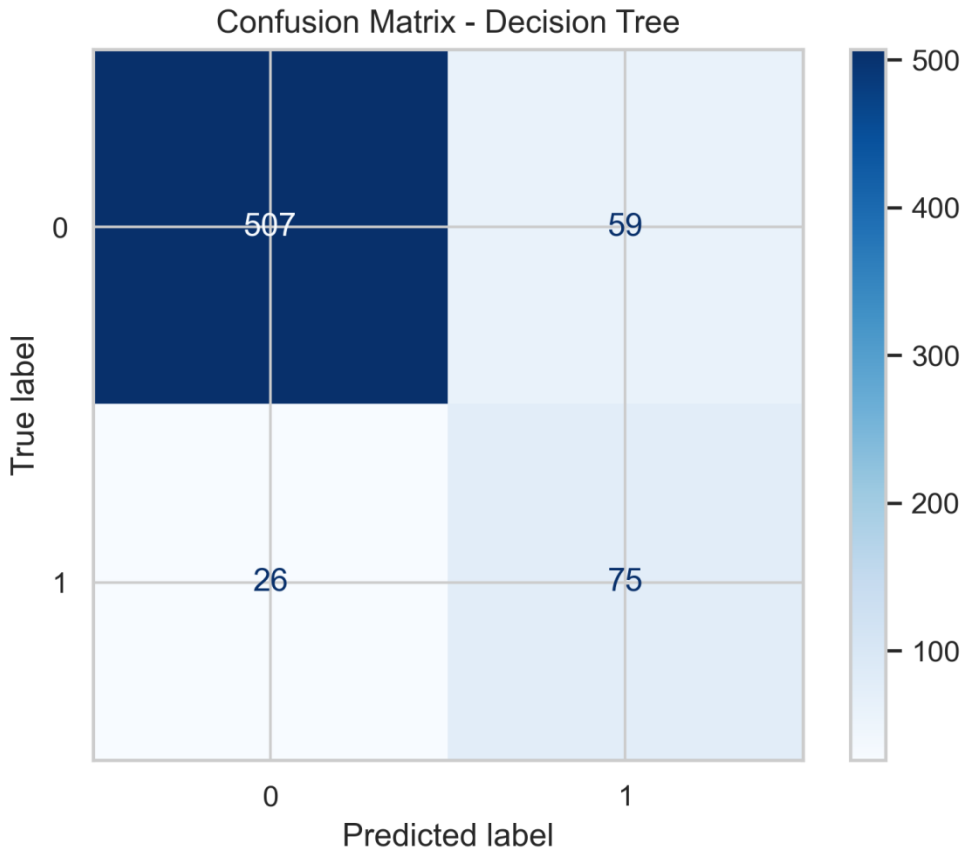
Precision: The precision for class 0 (not churned) is 95%. The precision for class 1 (churned) is 56%

Recall: The recall for class 0 (not churned) is 90% but the recall for class 1 (churned) is only 74%.

F1-score: The F1-score for class 0 (not churned) is 92% and for class 1 (churned) is only 64%.

In overall the decision tree model outperformed the logistic regression model across all metrics





The confusion matrix shows a total of 667 samples in the test set.

True Positives (TP): The model correctly predicted 59 samples as Not churned (class 0).

True Negatives (TN): The model correctly predicted 507 samples as churned (class 1).

False Positives (FP): The model incorrectly predicted 75 samples as churned when they were not churned.

False Negatives (FN): The model incorrectly predicted 26 samples as not churned when they were churned.

Model 3 : Random Forest

Baseline Model*

Train Accuracy: 1.00

Test Accuracy: 0.87

Random Forest Classification Report:

	precision	recall	f1-score	support
0	0.96	0.96	0.96	566
1	0.79	0.78	0.79	101
accuracy			0.94	667
macro avg	0.88	0.87	0.87	667
weighted avg	0.94	0.94	0.94	667

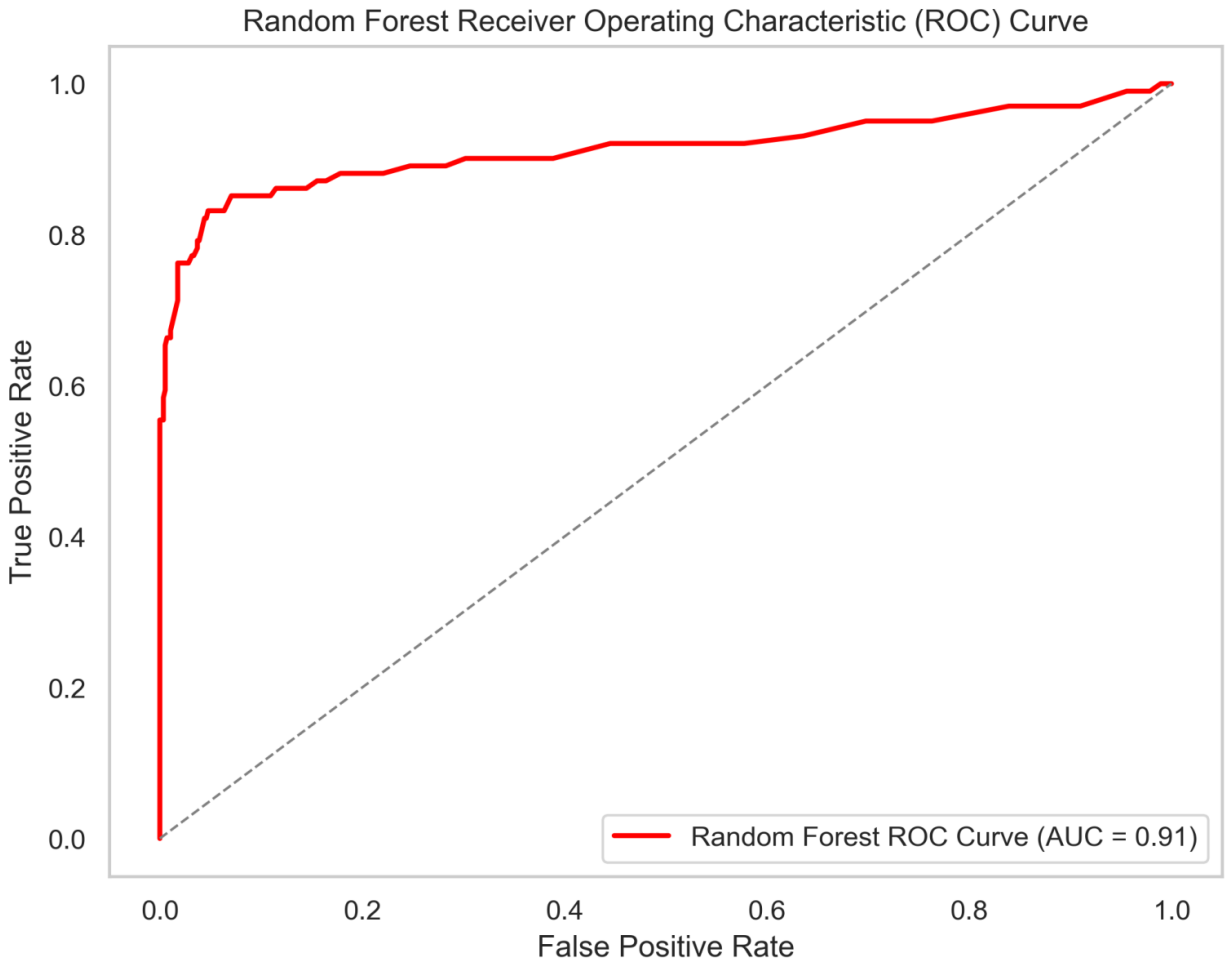
Comments and notes on model Accuracy: The accuracy of the model is 87% Train Accuracy: 1.00 Test Accuracy: 0.87

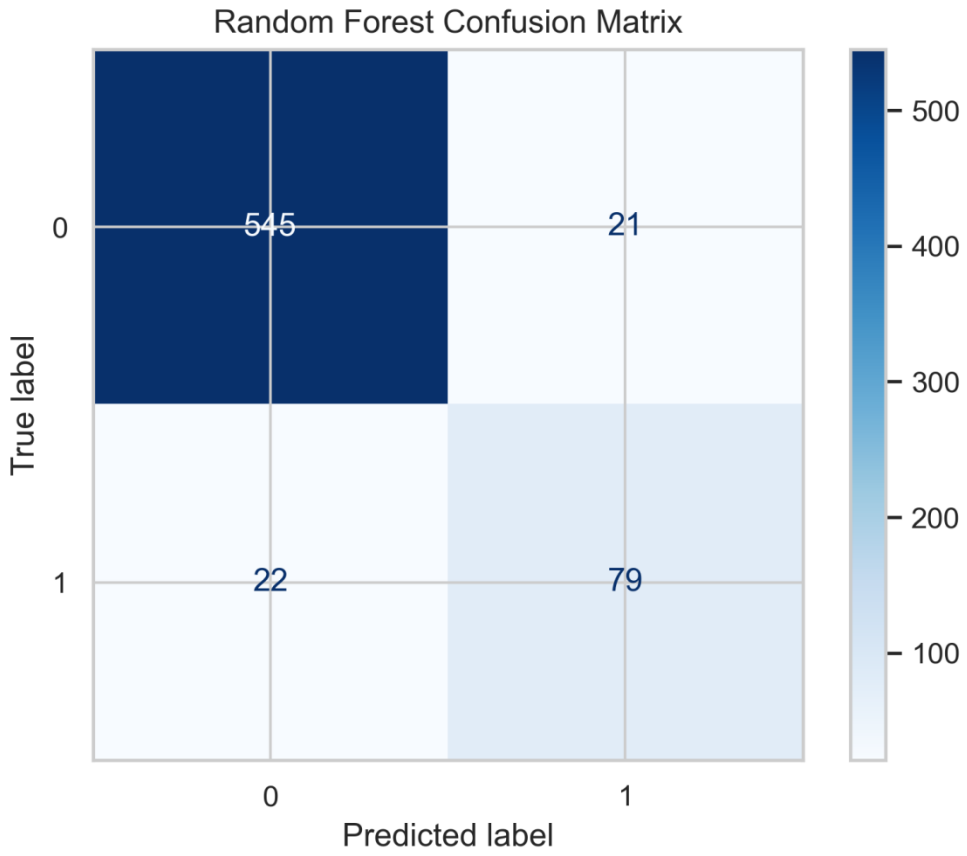
Classification Report:

Precision: The precision for class 0 (not churned) is 96%. The precision for class 1 (churned) is 79%

Recall: The recall for class 0 (not churned) is 96% but the recall for class 1 (churned) is only 78%.

F1-score: The F1-score for class 0 (not churned) is 96% and for class 1 (churned) is only 79%.


Random Forest Receiver Operating Characteristic (ROC) Curve
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Random Forest ROC Curve (AUC = 0.91)



The confusion matrix shows a total of 667 samples in the test set.

True Positives (TP): The model correctly predicted 21 samples as Not churned (class 0).

True Negatives (TN): The model correctly predicted 546 samples as churned (class 1).

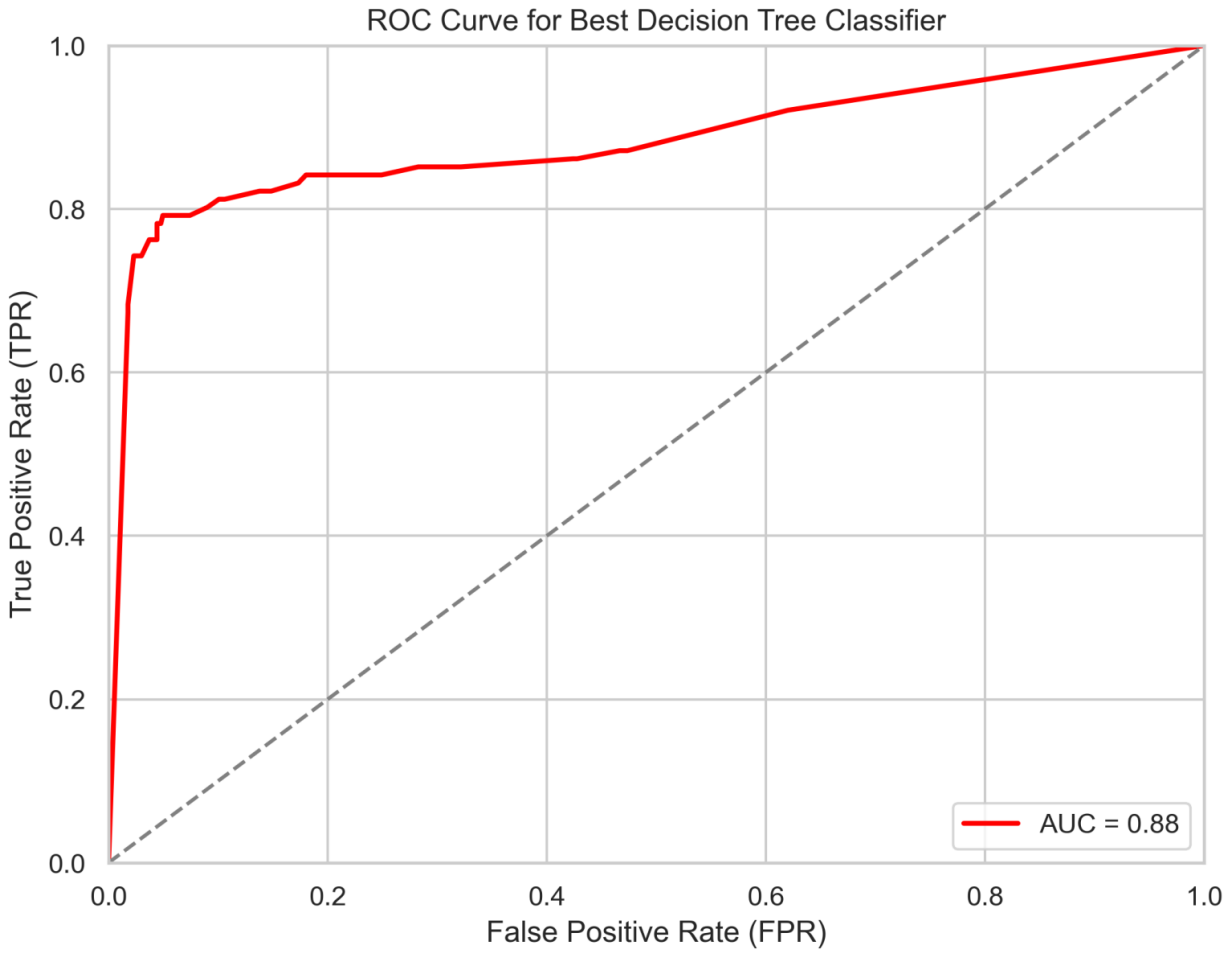
False Positives (FP): The model incorrectly predicted 79 samples as churned when they were not churned.

False Negatives (FN): The model incorrectly predicted 22 samples as not churned when they were churned.

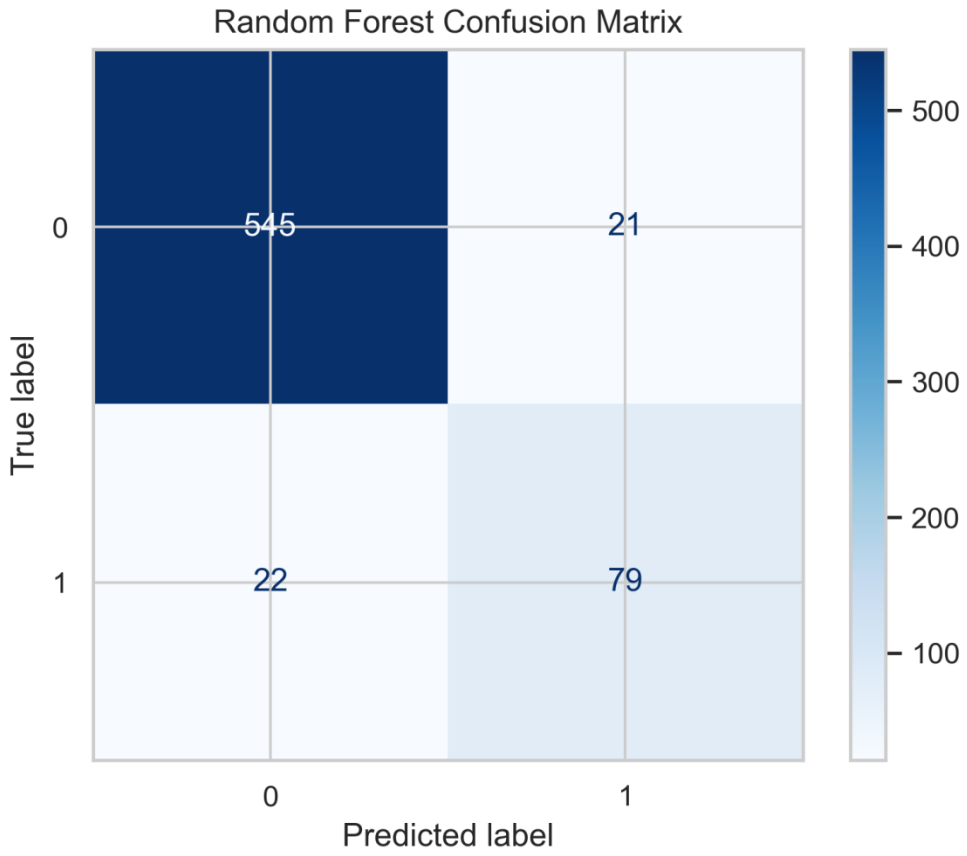
The random forest model outperformed both the logistic regression and decision tree models, exhibiting higher accuracy, precision, recall, and F1-score, indicating its superior predictive capability for churn prediction.

Hyper parameter tuning

There are some parameters of decision trees that can be tuned for the model's better performance. This includes n_estimators, max_depth, min_samples_split, min_samples_leaf and max_features.



From the curve above, we see that the curve follows the left part of the border, implying the more accuracy on the test



train score 0.9358581436077058

test score 0.9190404797601199

Random Forest Classification Report:

Accuracy: 0.9190404797601199

Precision: 0.7079646017699115

Recall: 0.7920792079207921

F1 Score: 0.7476635514018691

total_intl_charge: 0.2127527367025068

total_day_minutes: 0.2094645744666386

total_intl_calls: 0.16493829107022803

total_day_calls: 0.07758568448185565

voice_mail_plan: 0.049270866267131244

total_eve_calls: 0.038586559430451466

total_night_calls: 0.038389175009939415

customer_service_calls: 0.03504036024674709

total_night_charge: 0.030551144502230596

total_eve_minutes: 0.027290642466317367

area_code: 0.012901613028732722

number_vmail_messages: 0.012273171092812252

total_intl_minutes: 0.007333329154956915

total_day_charge: 0.005029296285431321

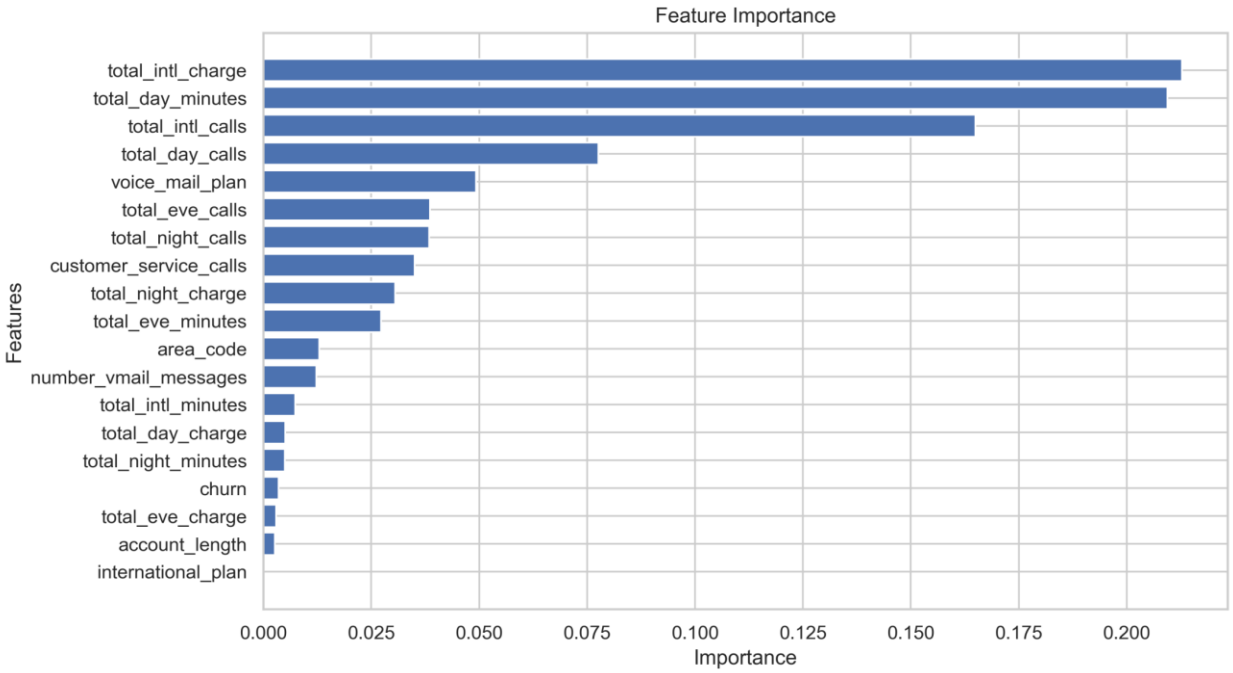
total_night_minutes: 0.004992802260004188

churn: 0.0035410213975318842

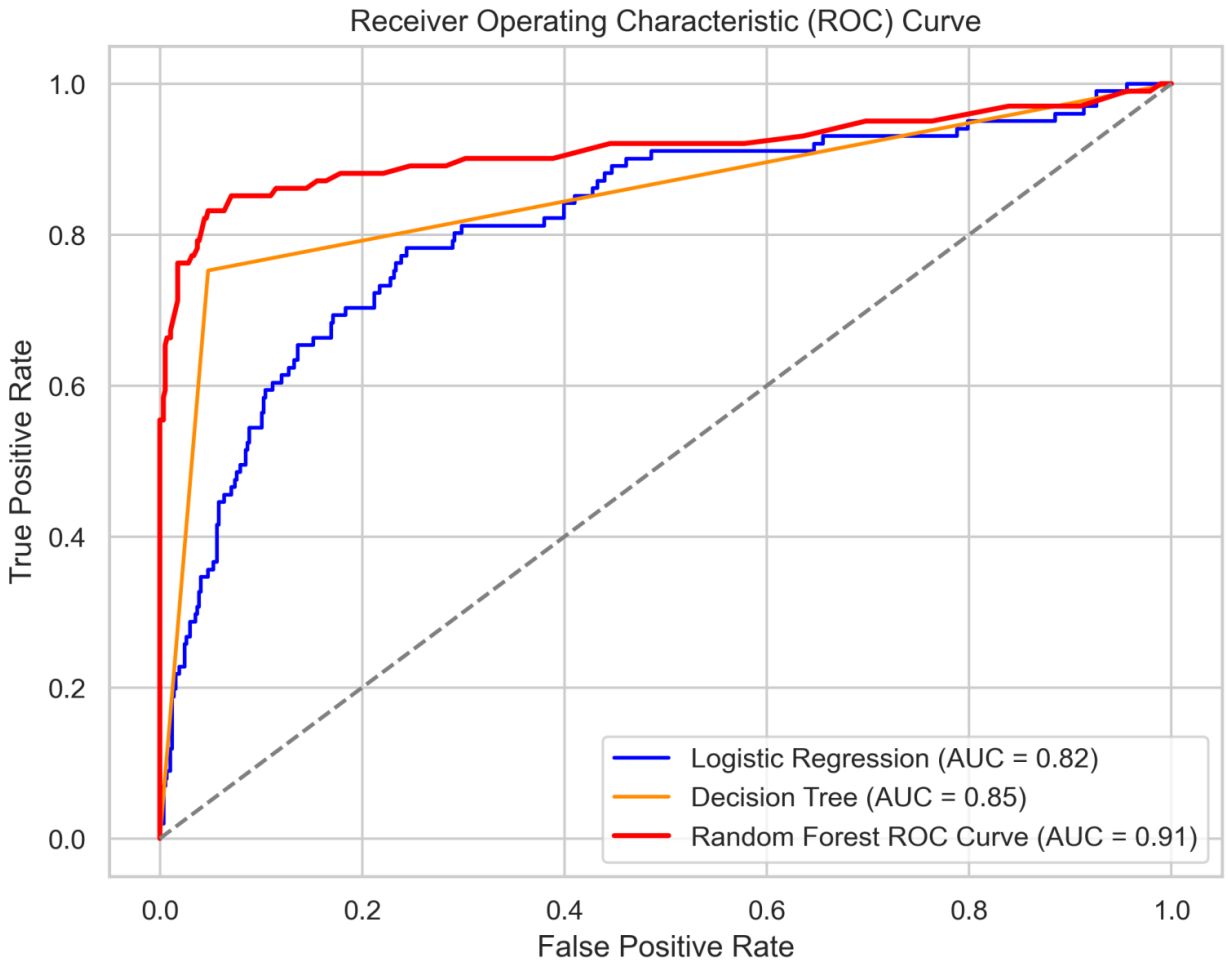
total_eve_charge: 0.002942740694801391

account_length: 0.0026407882262165843

international_plan: 0.0



The AUC values of Logistic Regression, Random Forest, Decision Tree model



Model evaluations

Logistic Regression: Accuracy: 0.82 Precision: 0.72 Recall: 0.65 F1-score: 0.52 AUC-ROC Score: 0.83 Summary: Logistic Regression achieves moderate accuracy and precision but lower recall compared to other models. The AUC-ROC score indicates good overall performance.

Decision Tree: Accuracy: 0.87 Precision: 0.56 Recall: 0.74 F1-score: 0.62 AUC-ROC Score: 0.82 Summary: Decision Tree exhibits lower accuracy and precision than Random Forest. However, it also has lower recall and a slightly lower AUC-ROC score, indicating suboptimal performance.

Random Forest: Accuracy: 0.92 Precision: 0.71 Recall: 0.791 AUC-ROC Score: 0.91 Summary: Random Forest improves accuracy, precision, and recall compared to Decision Tree. Its AUC-ROC score is the highest.

Random Forest with tuned hyperparameters is the model with the best performance. The model has the highest recall score. The accuracy and precision scores are also high. However, the recall score achieved is below the set score of at least 85%. The evaluation metric values of the tuned decision tree are shown below:

train score 0.9358581436077058

test score 0.9190404797601199

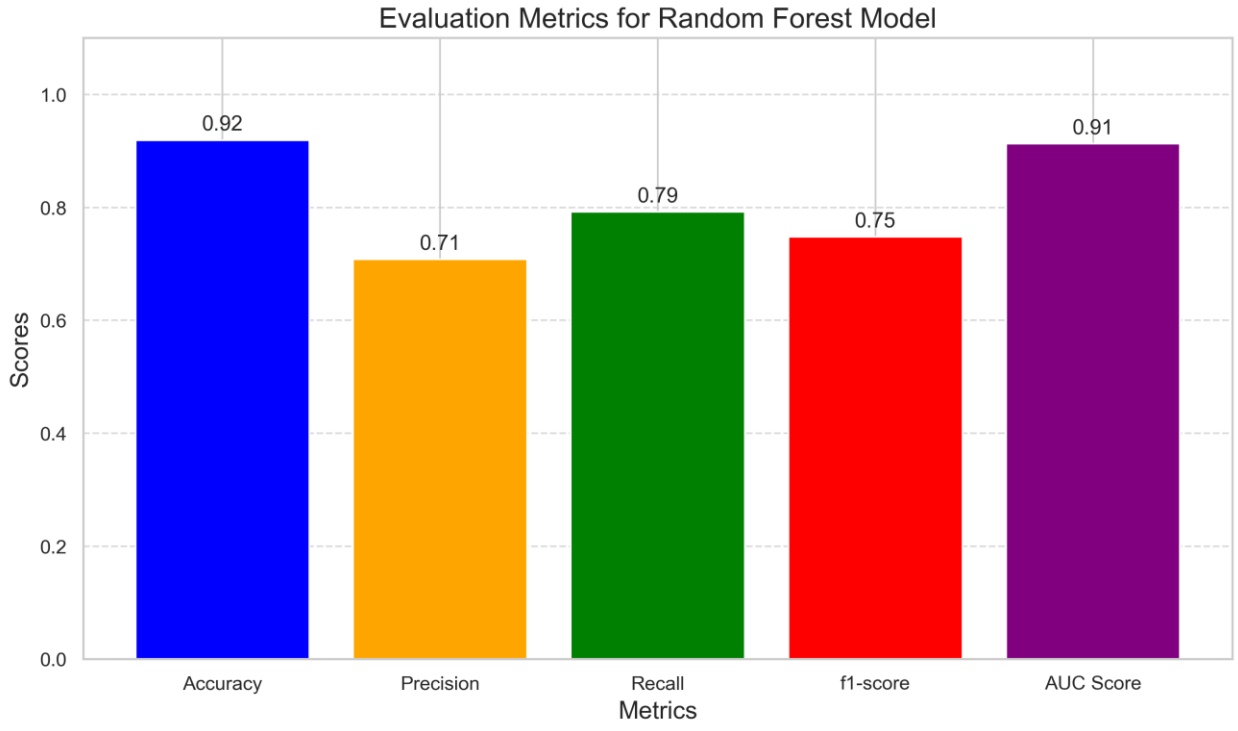
Random Forest Classification Report:

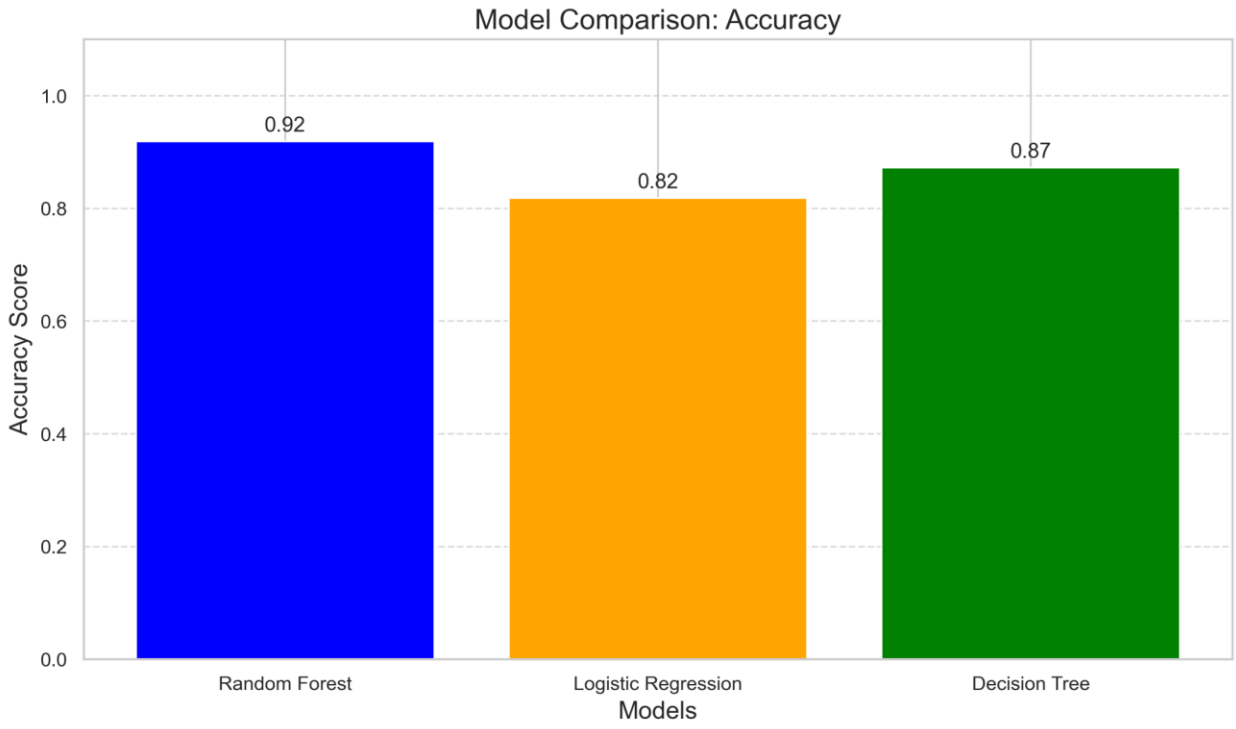
Accuracy: 0.9190404797601199

Precision: 0.7079646017699115

Recall: 0.7920792079207921

F1 Score: 0.7476635514018691





Analysis of Results

- The Decision Tree model's superior performance indicates its ability to handle complex interactions between variables.
- The high correlation of churn with customer service calls suggests an area for immediate intervention.

Challenges

- Imbalanced data: Required techniques like SMOTE to balance churn vs. non-churn classes.
- Feature selection: Ensured that highly correlated variables did not distort results

Conclusion

Model Evaluation: Rigorous evaluation of various machine learning models, including Logistic Regression, Decision Trees and Random Forest was conducted to anticipate customer churn accurately.

Performance Comparison: Among the models assessed, Random Forest emerged as the most effective, exhibiting superior accuracy compared to other algorithms.

The most important features for predicting customer churn are:

total day minutes: total number of minutes the customer has been in calls during the day

total evening minutes: total number of minutes the customer has been in calls during the evening

customer service calls: number of calls the customer has made to customer service

total international minutes: total number of minutes the user has been in international calls

Recommendation

Customer Retention Strategies: Implement proactive measures such as personalized offers, loyalty programs, and targeted marketing campaigns to incentivize customer retention and foster brand loyalty.

Service Improvement Initiatives: Continuously monitor and improve service quality, addressing pain points and enhancing customer satisfaction across all touchpoints.

Enhanced Communication Channels: Establish effective communication channels to gather customer feedback, address concerns promptly, and provide timely support, thereby building trust and loyalty.

Data-Driven Decision Making: Leverage advanced analytics and machine learning models to gain deeper insights into customer behavior, preferences, and churn drivers, enabling data-driven decision-making and strategic interventions.

References

- Dataset Source: Provided by SyriaTel.
- Tools: Python (pandas, scikit-learn, matplotlib), Jupyter Notebook.
- Documentation: Scikit-learn User Guide.

This report serves as a comprehensive overview of the customer churn analysis project for SyriaTel, providing insights and strategies to enhance customer retention.