# Project presentation

Presented by Khadija Omar

# CONTENT

# Project overview

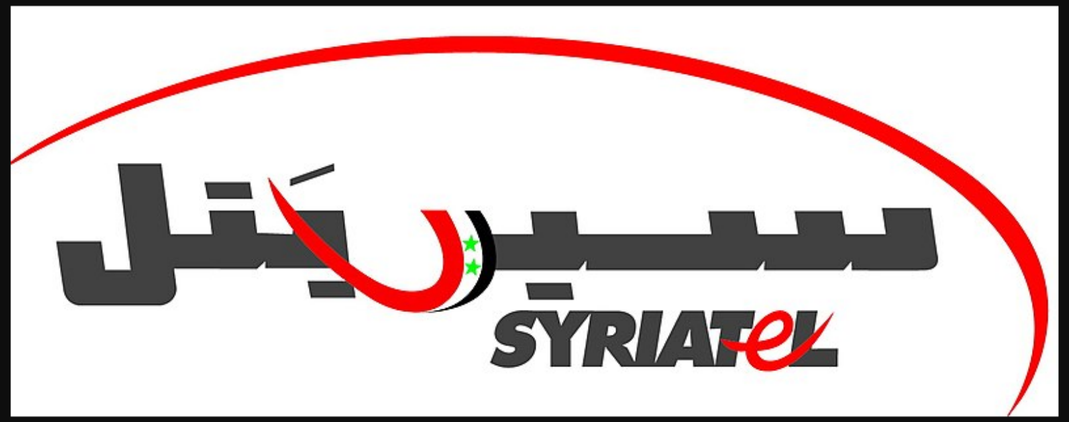
Customer churn, also known as customer attrition or customer defection, is a critical metric in businesses, particularly in subscription-based services and industries with recurring revenue models. Churn refers to the rate at which customers discontinue their relationship with a company or cancel their subscription to a service or product.

I

# Project Overview

In the business context, customer churn is a significant concern due to its potential adverse effects on a company's revenue, growth, and overall sustainability. Understanding and mitigating customer churn is essential for maintaining a healthy customer base and fostering long-term success.This project focuses on analyzing the churn dataset from the telecom industry to predict customer churn and gain insights into the key factors driving churn.

# Your Task: Predict customer churn using a classification algorithm model



# Business Understanding

Syriatel, a telecommunications company, is facing a high churn rate, with many customers discontinuing their services and switching to competitors. The company wants to address this issue by developing a customer churn prediction model. By analyzing the dataset, SyriaTel aims to gain insights into factors associated with churn, with the goal of reducing churn rate, increasing customer retention, and improving overall profitability.

# Data Understanding

This dataset was obtained from kaggle , "https://www.kaggle.com/datasets/becksddf/churn-in-telecoms-dataset" . It contains 3333 entries and 21 columns. 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.

# Data Understanding

The dataset provided information on the following features for each customer:
State, Length of account, Area code, Phone numbers, If the customer has an international plan, If the customer has a voicemail plan, No. of voicemail messages, Breakdown of call minutes for day, evening, night, and international, Breakdown of call charges for day, evening, night, and international, Breakdown of no. of calls for day, evening, night, and international, No. of calls to customer service, If they have churned.

# Data Preparation

In this section, we are going to do several actions to prepare our data for exploratory data analysis and modelling. First, we will import all the necessary libraries, load the dataset using pandas library, preview the data (how many features a records, as well as statistical features), and conduct thorough data preprocessing (checking and removing any missing values and transforming data).

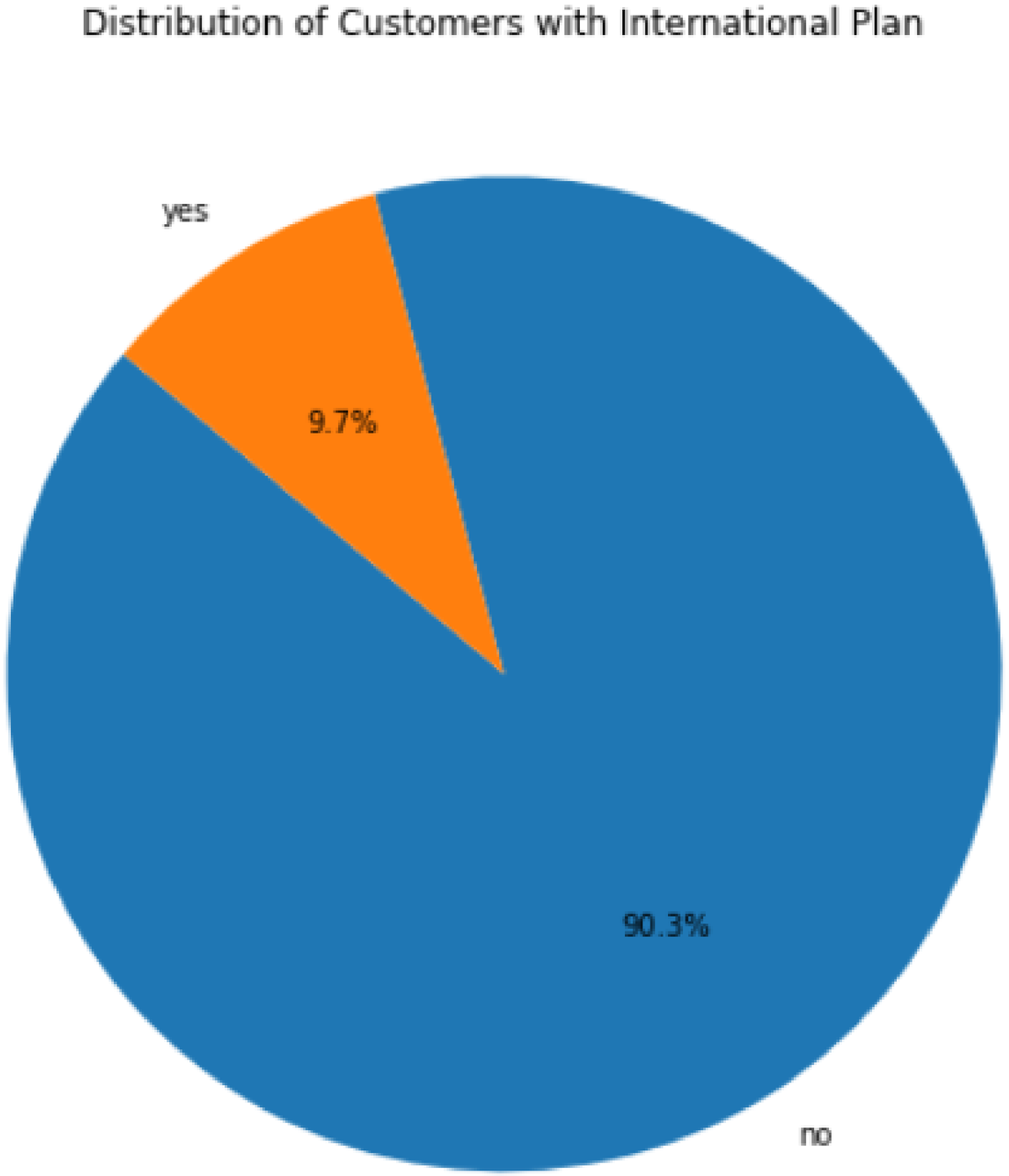
# EDA Analysis

Most of the customers are from West Virginia, Minnesota, NewYork, Alabama, and Wisconsin.

Of the 3333 customers,323 customers have an international plan (that makes up 9.7% of the customers) and 3010 customers do not have an international plan (that makes up 90.3% of the customers)

Of the 3333 customers,922 customers have a voicemail plan and 2411 customers do not have a voicemail plan.

# EDA Analysis

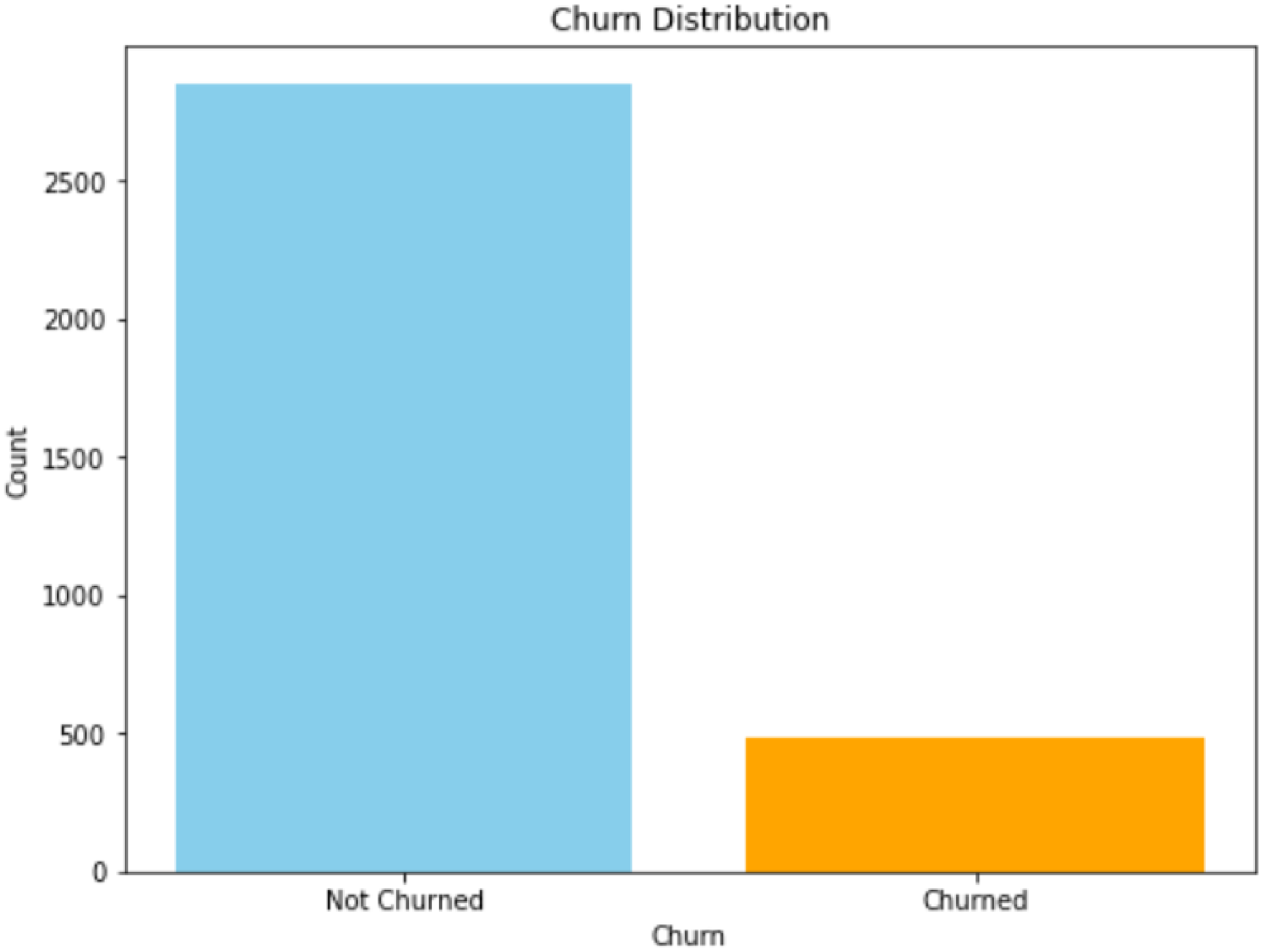


# EDA Analysis

Of the 3333 customers,922 customers have a voicemail plan and 2411 customers do not have a voicemail plan.

Out of the 3,333 customers in the dataset, 483 have terminated their contracts. That is 14.5% of customers lost. The distribution of the binary classes shows a data imbalance. This needs to be addressed before modeling as an unbalanced feature can cause the model to make false predictions.

# EDA Analysis



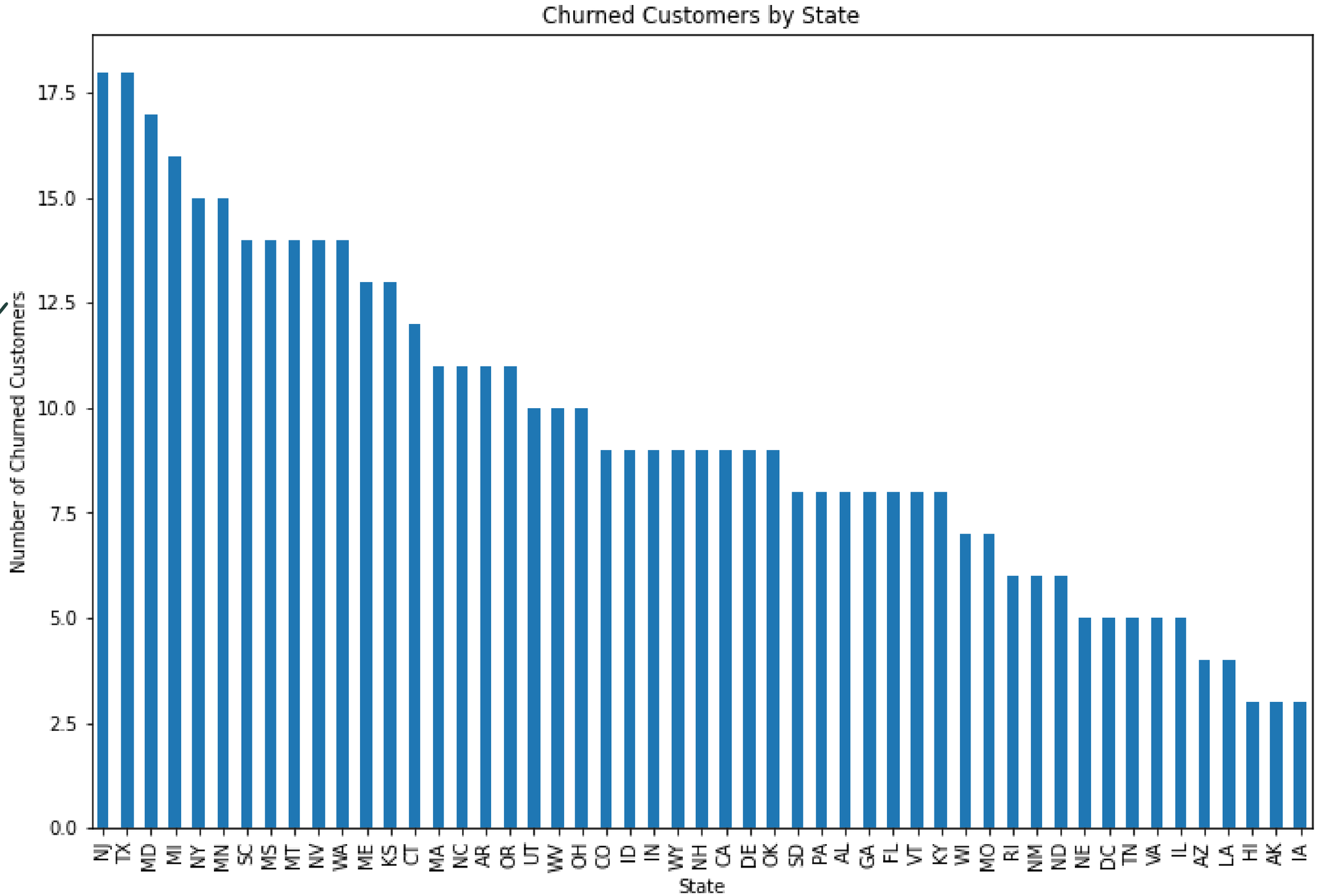
# EDA Analysis

Of the customers who have terminated their accounts, most of them are from area codes 415 and 510.

Of all the customers that churned, the majority are from Texas, New Jersey, Maryland, Miami, and New York.

The majority of customers who churned did not have an international plan nor did they have a voicemail plan.

# EDA Analysis



# Modelling

In this section, we will build a model that can predict customer churn based on the features in our dataset. The model will be evaluated on the recall score.

In order to achieve the targets stipulated in the project proposal, we will be using the following algorithms:
Logistic Regression
Decision Tree
Random Forest
XG Boost

We will also be using the ROC_AUCmetric to evaluate the performance of our models.

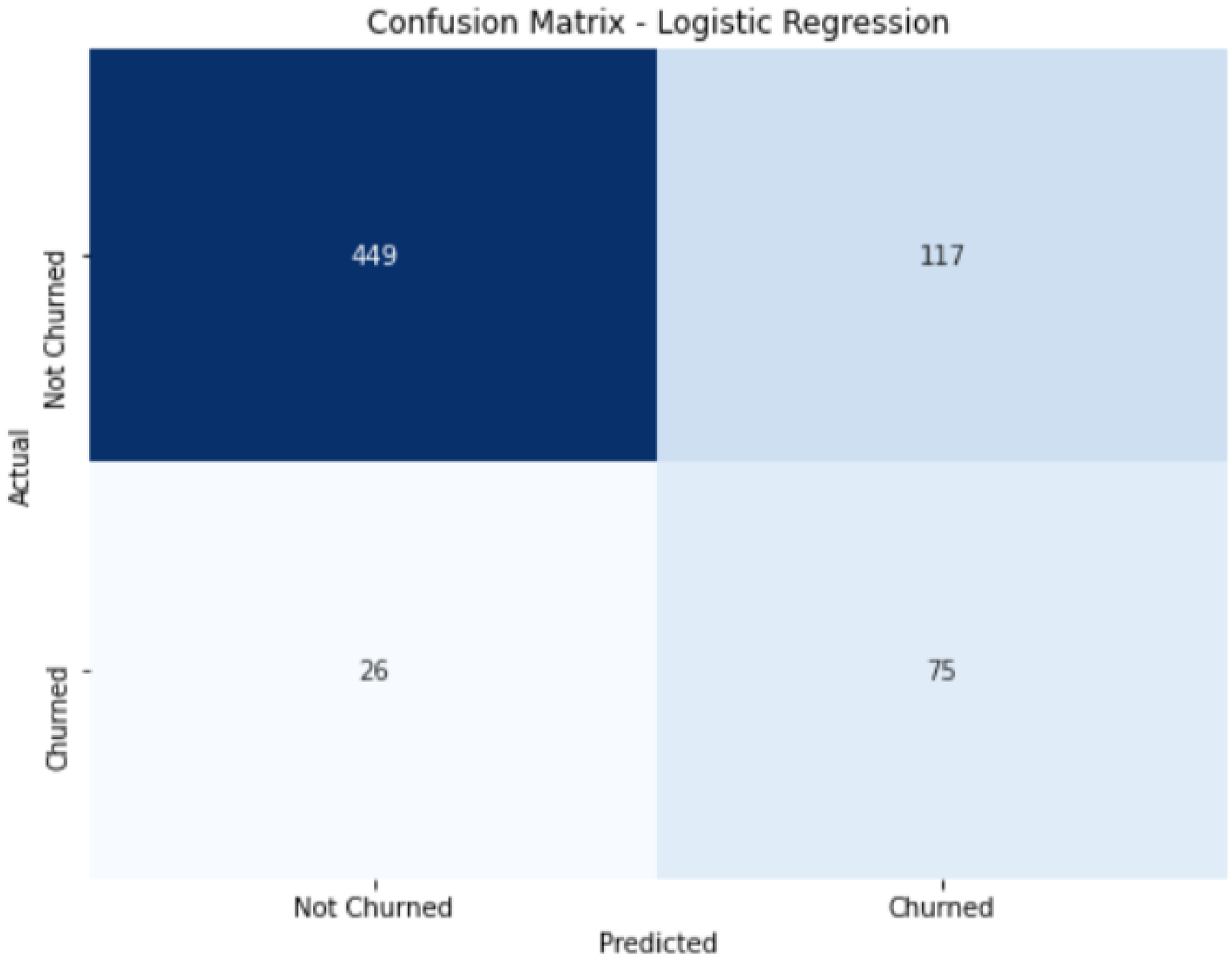
# Modelling

The logistic regression model has a recall score of 0.74, which is actually good for a baseline model. This means that the model can identify around 74% of the actual positive instances correctly.

The confusion matrix evaluation showed that the model had a higher number of true positives and true negatives than false positives and false negatives. This indicates that the model is making correct predictions more often than incorrect ones and is not overfitting.

According to the model, customer service calls, international plan, and total day charges are the top three most important features.

# Modelling


Confusion Matrix - Logistic Regression
449
117
26
75
Not Churned
Churned
Actual
Not Churned
Churned
Predicted

# Modelling

The decision tree model has a recall score of 0.67, which is actually good but not better than our baseline model. This means that the model can identify around 67% of the actual positive instances correctly.

The confusion matrix evaluation showed that the model had a higher number of true positives and true negatives than false positives and false negatives. This indicates that the model is making correct predictions more often than incorrect ones and is not overfitting.

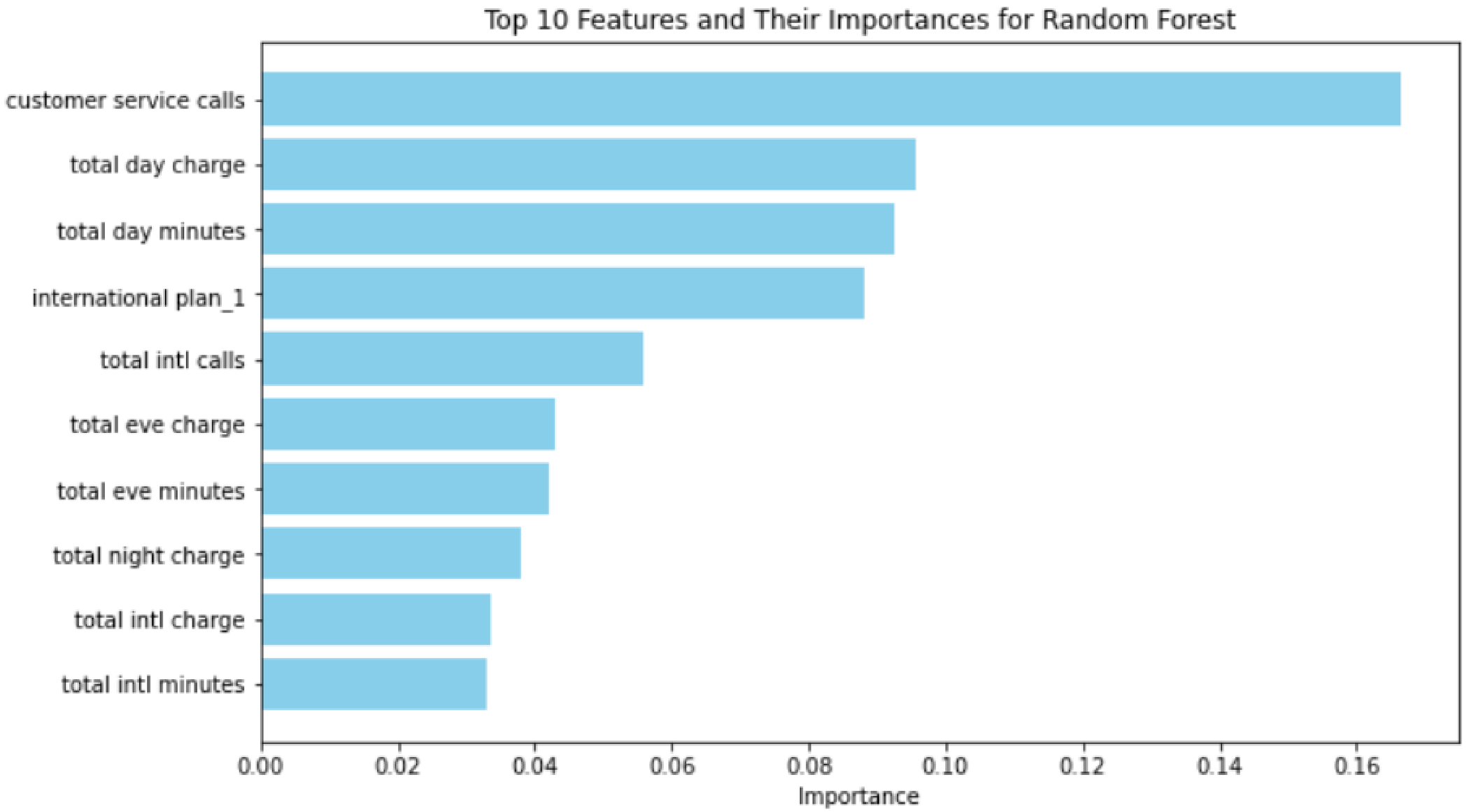
According to the model, customer service calls, total day minutes, and international plan are the top three most important features.

# Modelling

The random forest classifier model has a recall score of 0.73, which is great compared to the previous model. This means that the model can identify around 73% of the actual positive instances correctly.

The confusion matrix evaluation showed that the model had a higher number of true positives and true negatives than false positives and false negatives. This indicates that the model is making correct predictions more often than incorrect ones and is not overfitting.

According to the model, customer service calls, total day charges, and total day minutes are the top three most important features.


Top 10 Features and Their Importances for Random Forest
customer service calls
total day charge
total day minutes
international plan_1
total intl calls
total eve charge
total eve minutes
total night charge
total intl charge
total intl minutes
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
0.16
Importance

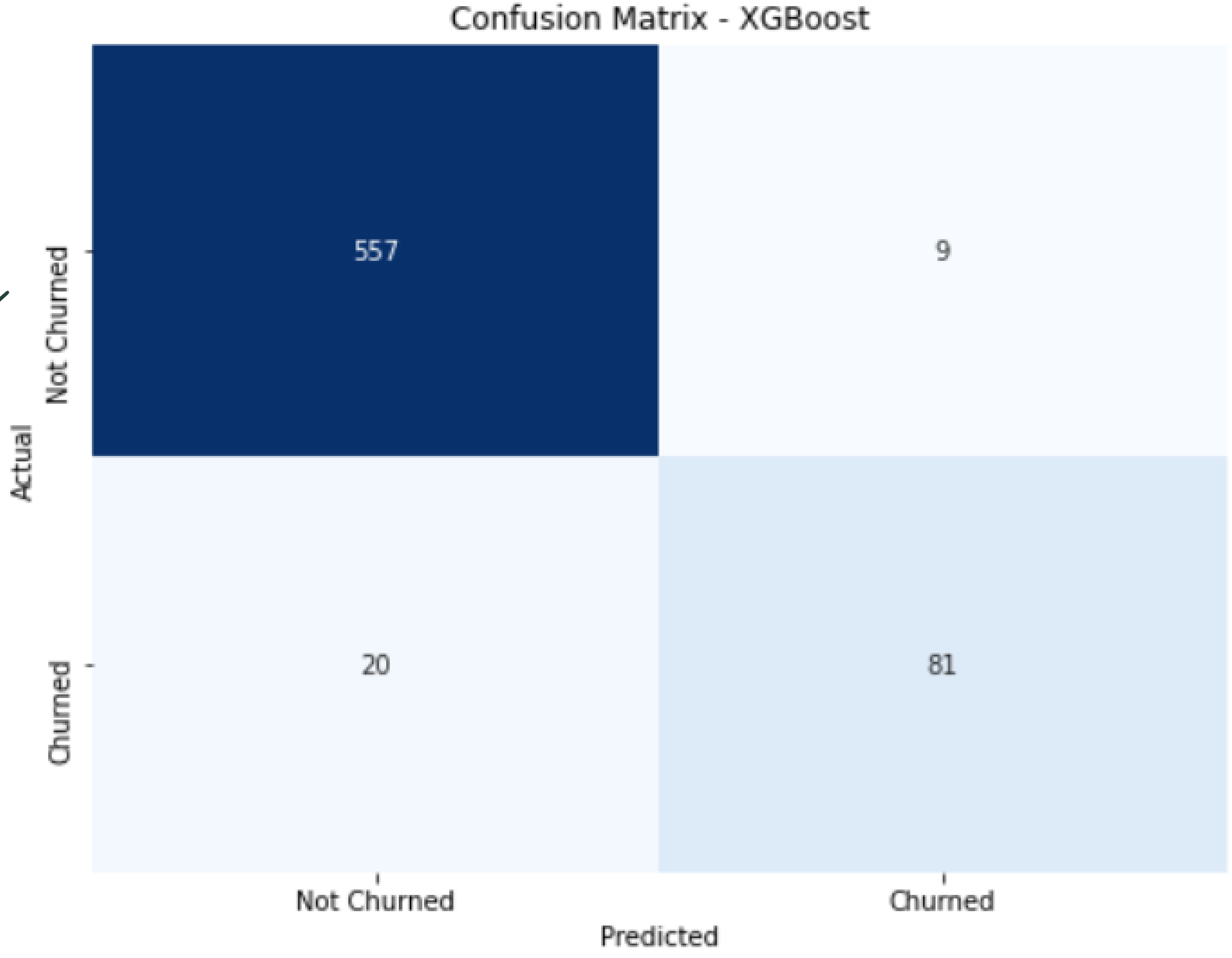
# Modelling

The XGBoost classifier model has a recall score of 0.80, which is actually better than all the previous models. This means that the model can identify around 80% of the actual positive instances correctly.

The confusion matrix evaluation showed that the model had a higher number of true positives and true negatives than false positives and false negatives. This indicates that the model is making correct predictions more often than incorrect ones and is not overfitting.

According to the model international plan, customer service calls and voice mail plan are the top three most important features.

# Modelling

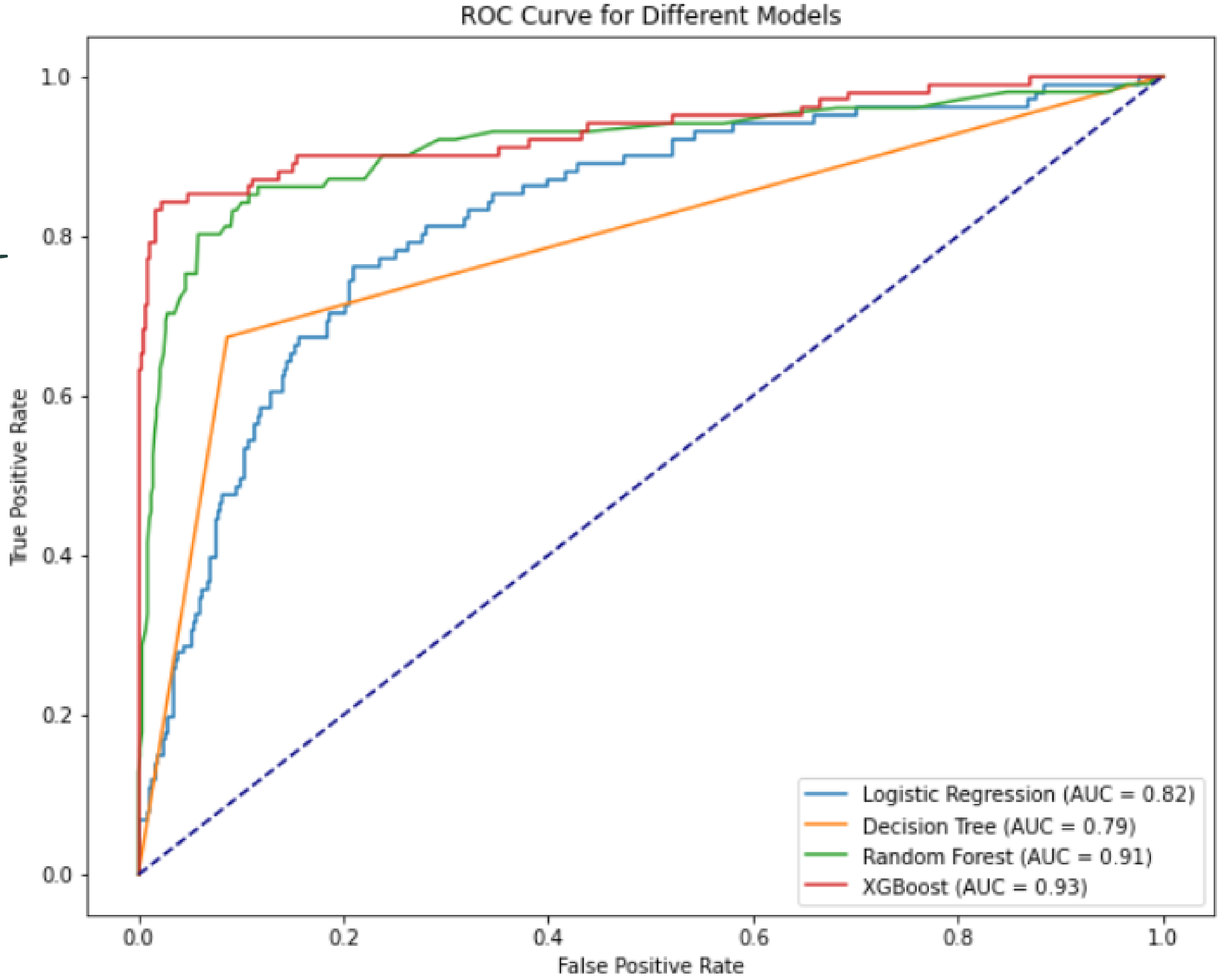

Confusion Matrix - XGBoost
Actual
Not Churned
Churned
557
9
20
81
Not Churned
Churned
Predicted

# Evaluation

The ROC curve analysis shows that the XGBClassifier has the best performance, followed by the RandomForestClassifier, LogisticRegression and Decision Tree Classifier.The XGBClassifier has the highest AUC score of 0.93, while the DecisionTree has the lowest AUC score of 0.79.

The ROC curve is a graphical plot that shows the trade-off between the true positive rate (TPR) and the false positive rate (FPR) for a binary classifier. The TPR is the proportion of positive instances that are correctly classified, while the FPR is the proportion of negative instances that are incorrectly classified. The AUC is the area under the ROC curve, and it is a measure of the overall performance of the classifier.

# Evaluation


ROC Curve for Different Models
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Logistic Regression (AUC = 0.82)
Decision Tree (AUC = 0.79)
Random Forest (AUC = 0.91)
XGBoost (AUC = 0.93)

# Evaluation

A higher AUC score indicates that the classifier is better at distinguishing between positive and negative instances.

The tuned XGBoost classifier model has a recall score of 0.81, meaning that the XGBoost model slightly improved after tuning This means that the model can identify around 81% of the actual positive instances correctly.

According to this model, international plan, voice mail plan, and customer service calls are the top three most important features.

# Recommendations

Following the modelling and analysis that we have done above we can come up with the following recommendations:

1. Offer discounts or promotional offers to customers in area codes 415 and 510, as these areas have a higher churn rate. This can help incentivize customers to stay with the company.

2. Offering attractive international plans might help in retaining customers.

3. Improve customer service quality and reduce the number of customer service calls. Enhance training programs for customer service representatives to ensure prompt and effective resolution of customer issues, leading to higher customer satisfaction and reduced churn.

# Recommendations

4. Focus on customer retention strategies in states with higher churn rates, such as Texas, New Jersey, Maryland, Miami, and New York. This can involve targeted marketing campaigns, personalized offers, or improved customer support tailored to the specific needs and preferences of customers in those states.

5. Enhance the value proposition of the voicemail plan to increase adoption among customers. Highlight the benefits and convenience of voicemail services, and consider offering additional features or discounts to encourage customers to sign up.

6. Monitoring and managing total day charges and minutes is crucial to address customer concerns and potential churn.

# Next Steps

- While the model currently has good accuracy and performance, we should continue monitoring and evaluating its performance on new data. As customer behaviors and preferences change over time, it's important to ensure that the model remains effective and up-to-date.

- Exploring advanced techniques like ensemble methods, gradient boosting methods, and Adaboost, or deep learning to further improve churn prediction performance.

Thank you