Cheptoi-Millicent /
DS-PHASE3-PROJECT

Code Issues Pull requests Actions Projects Wiki Security Insights Setting

0 stars 0 forks 1 watching Branches Activity
Tags

Public repository

1 Branch 0 Tags Go to file t Go to file Add file Code

**Cheptoi-Millicent** ipynb file 4ffe549 · 3 minutes ago

| | | |
|---|---|---|
| .gitignore | Initial commit | 3 days ago |
| Area Code.png | Image | 5 minutes ago |
| Churn Distribution.png | Image | 5 minutes ago |
| Notebook.ipynb | ipynb file | 3 minutes ago |
| Presentation.pdf | Slides | 4 minutes ago |
| README.md | Update README.md | 6 minutes ago |
| bigml_59c28831336c6604c80... | Data | 3 days ago |

README

# DS-PHASE3-PROJECT

## Overview

This repo helps identify customers who are at risk of churn, thus helping the company proactively identify high-risk customers and take actions to retain them.

## Problem Statement

The objective of this project is to build a classifier that predicts whether a customer will "soon" stop doing business with SyriaTel, a telecommunications company. This is a binary classification problem where the goal is to identify customers who are at risk of churn, i.e., customers who will leave the service in the near future. The model will help the company proactively identify high-risk customers and take actions to retain them.

# Data

The data i worked with was accessed throught the link: (https://www.kaggle.com/datasets/becksddf/churn-in-telecoms-dataset)

# Business Objectives

- Churn Prediction: Identify the factors that are most likely to lead to customer churn.
- Customer Retention: Develop a model that can accurately predict which customers are at risk of churning.SyriaTel can take proactive steps to retain customers who are at risk of churning.
- Cost Reduction: Predicting churn allows SyriaTel to allocate resources effectively, targeting the customers who require intervention before they leave. This reduces the costs associated with acquiring new customers to replace those lost.
- Revenue Retention: Ultimately, reducing churn will contribute to higher customer lifetime value (CLV) and revenue retention, ensuring the long-term profitability and sustainability of SyriaTel.

# Success Metrics

The success criteria for this project include:

- Developing a robust churn prediction model with high recall score of 0.8
- Identifying the key features and factors that significantly contribute to customer churn.
- Providing actionable insights and recommendations to the telecom company for reducing churn and improving customer retention.
- Demonstrating the value of churn prediction models in enabling proactive retention strategies and reducing revenue losses due to customer churn.

# Source of data

- Telecom's Dataset to an external site (Kaggle

# Description of Data

Telecom's Data - The datatypes include: bool(1), float64(8), int64(8), object(4)

- Data Manipulation, the module used include: pandas and numpy
- Data Visuaalization, the module used include: seaborn, matplotlib, plotly( graph_objs, express)
- Modelling, the modules were accessed from the sklearn they include: LabelEncoder, MinMaxScaler, train_test_split,cross_val_score, GridSearchCV,

f1_score,recall_score,precision_score,confusion_matrix,roc_curve,roc_auc_score,classification_report: from scipy, it is the stats.

- Algorithms for supervised learning methods: The modules were accessed from sklearn they include DecisionTreeClassifier, RandomForestClassifier, LogisticRegression

# Loading of the Data

data = pd.read_csv("bigml_59c28831336c6604c800002a.csv")

# Data Preparation

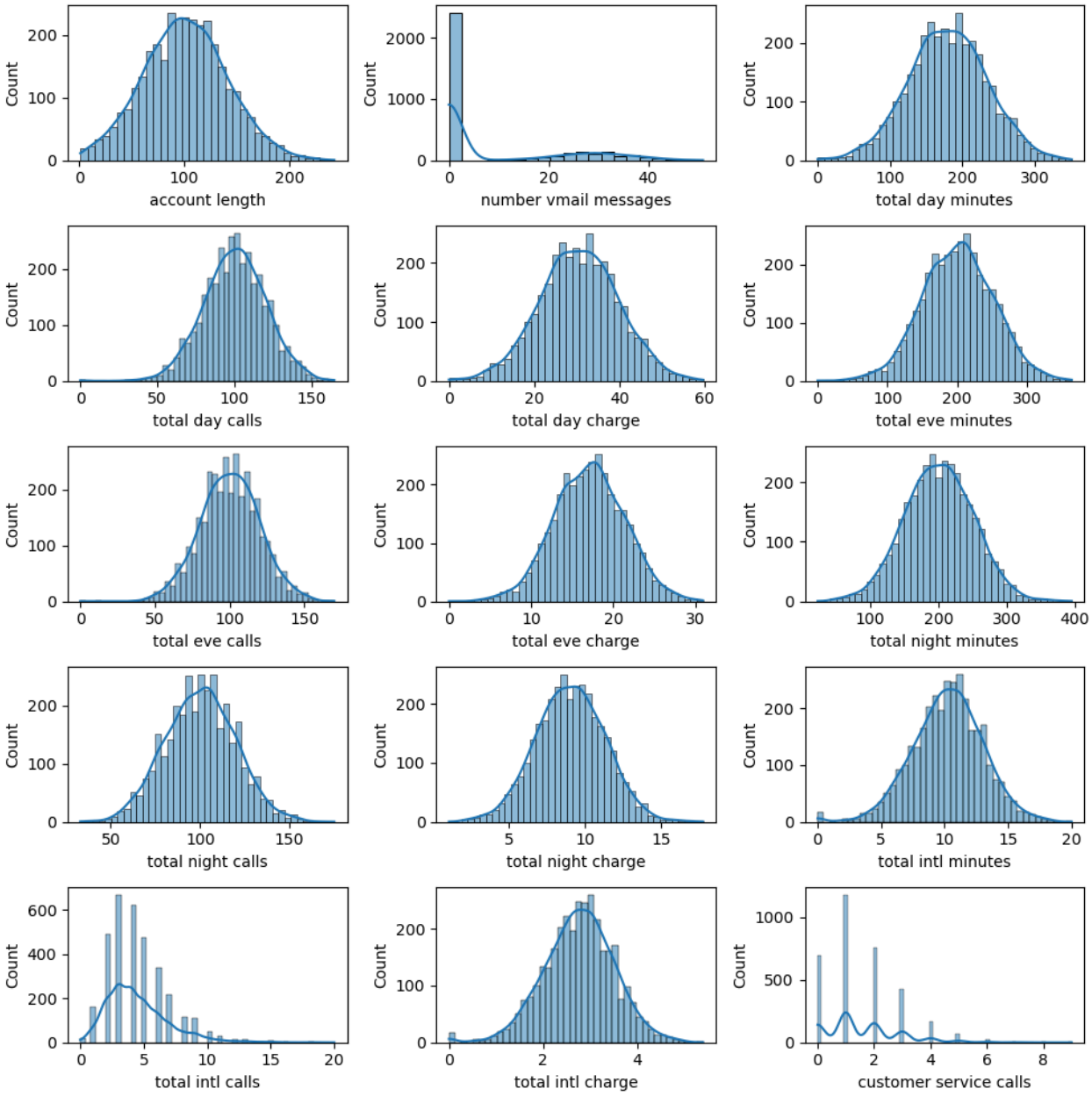
## Cleaning of Data

The data did not have any missing values or duplicates, so no cleaning was carried out apart from deleting the 'phone number column and converting the 'area code' datatype to 'object'

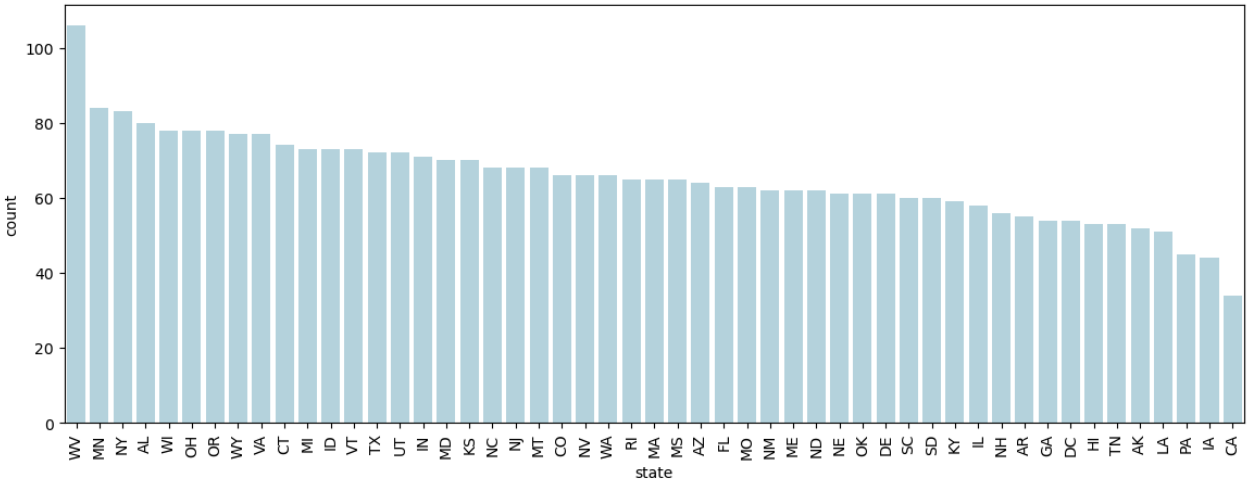
## Exploratory Data Analysis(EDA)

### Univariate Analysis

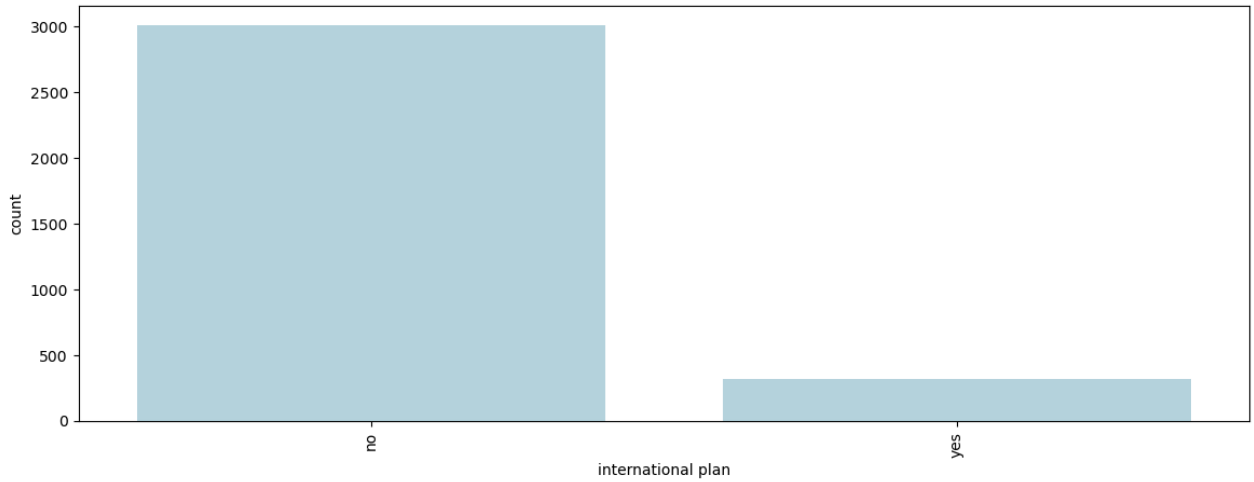
- Distribution of 'Churn' Feature
- Distribution of 'Area Code' Feature
- Distribution of Numerical Features



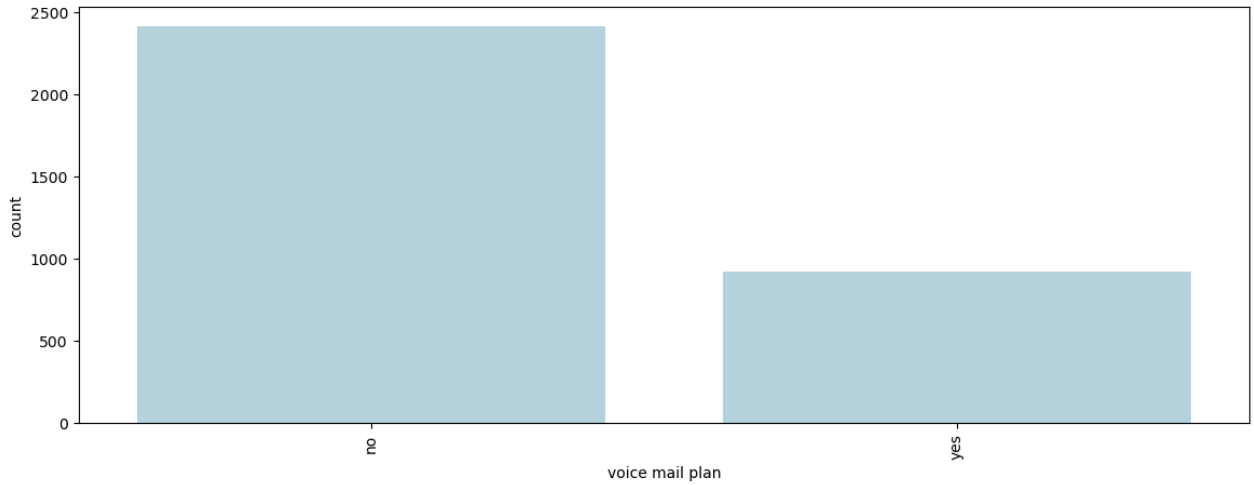

- Distribution of Categorical Features
  - State



- International plan

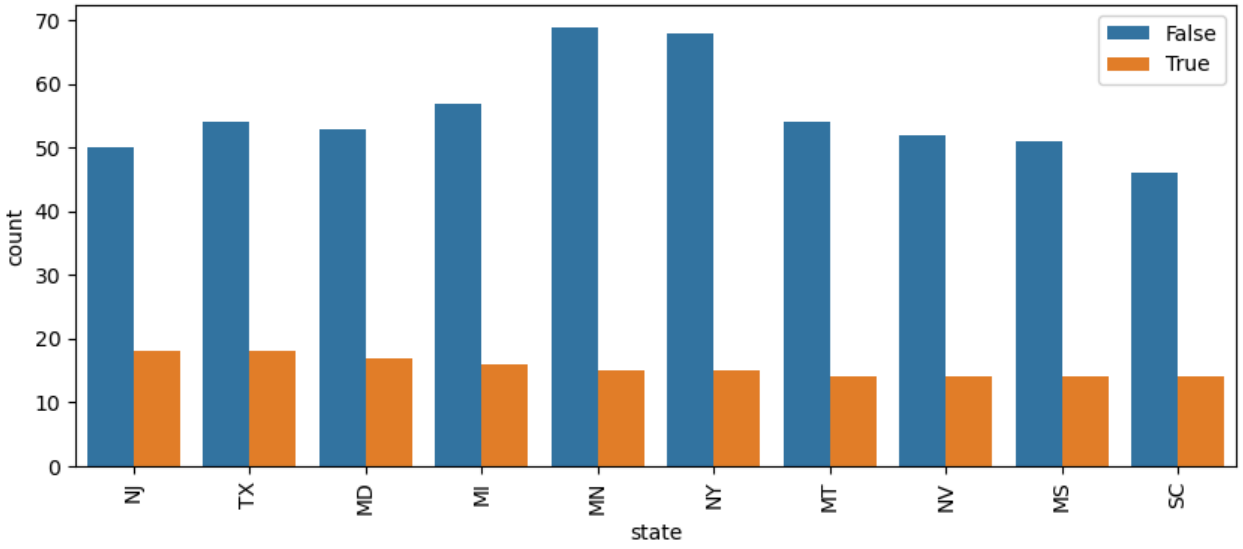



- Voicemail plan

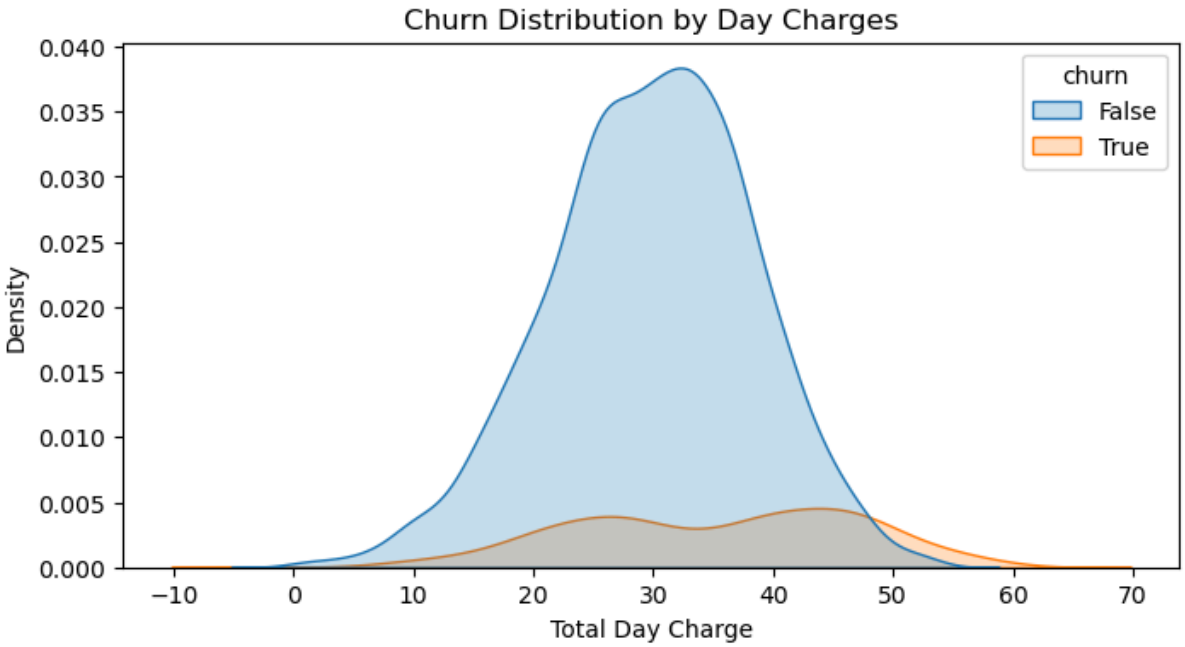



## Bivariate Analysis

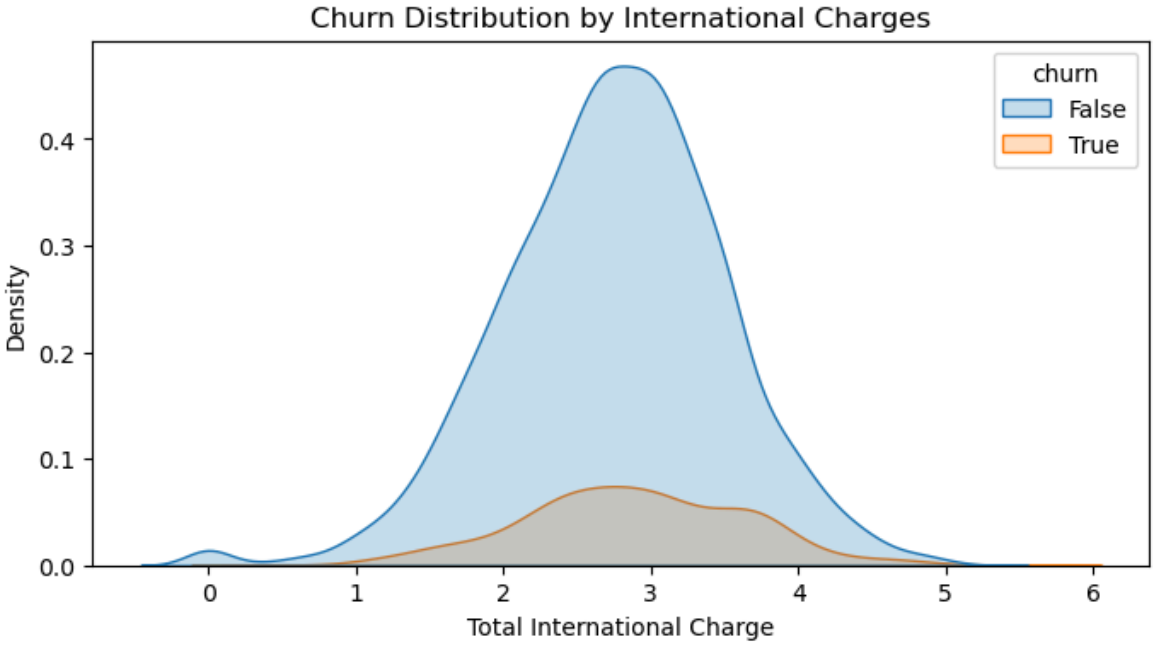
- State



- Churn Distribution By Day Charges




- Churn Distribution By International Charges




## Outliers

The outliers were dropped since they can disproportionately impact the performance of predictive models by introducing noise or skewing the training process.

## Features Correlation

Most of the features are not correlated however some do share a perfect correlation.

- `Total day charge` and `total day minutes` features are fully positively correlated.
- `Total eve charge` and `total eve minutes` features are fully positively correlated.
- `Total night charge` and `total night minutes` features are fully positively correlated.
- `Total int charge` and `total int minutes` features are fully positively correlated.

It makes sense for these features to be perfectly correlated because the charge is a direct result of the minutes used.

## Multicollinearity check

To check for multicollinearity among features, the dataset was analyzed using correlation matrix,. Multicollinearity occurs when two or more features in the dataset are highly correlated with each other, which can cause issues during modeling such as instability, overfitting, or inaccurate coefficient estimates.

## Feature Engineering

The process of transforming raw data into features that better represent the underlying problem to the predictive models, resulting in improved model accuracy on unseen data. In this phase, we'll perform Label Encoding, One Hot Encoding and Scaling the data.

### Label Encoding

Convertion of columns with 'yes' or 'no' to binary.

### One Hot Encoding

This is a technique used to convert categorical variables into a set of binary features. This is done by creating a new feature for each category, and then assigning a value of 1 to the feature if the category is present and 0 if it is not.

# Modelling

The model will be evaluated on the recall score. Specifically, if it achieves an recall score of 80% or higher, it will be considered a success.

In order to achieve the targets stipulated in the project proposal, we will be using the following algorithms:

- Logistic Regression
- Decision Tree
- Random Forest

The `ROC_AUC` metric to evaluate the performance of our models

To deal with class imbalance, we will be using `SMOTE` to generate synthetic examples of the minority class in our dataset

# Train-Test Split

Splitting data into train and test sets using a test_size of 0.25

# Scaling the data

Scaling is a technique used to transform numerical features into a comparable range. It helps in reducing the impact of outliers and standardizing the variables. In this process, the minimum value of the variable is transformed to 0, and the maximum value is transformed to 1, while the remaining values are scaled proportionally in between.

# Applying SMOTE to Resolve Unbalanced 'churn' Feature

Synthetic Minority Oversampling Technique ("SMOTE") is an oversampling technique where synthetic samples are generated for the minority class. This algorithm helps to overcome the overfitting problem posed by random oversampling.The technique aims to balance class distribution by randomly increasing minority class examples by replicating them.

# Logistic Regression

A statistical model used for binary classification tasks. It is a type of regression analysis where the dependent variable is binary. The goal of logistic regression is to estimate the probability of an instance belonging to a specific class based on the values of the independent variables.

# Decision Tree Classifier

It is a supervised machine learning algorithm that can be used to classify data. Decision trees work by splitting the data into smaller and smaller subsets until each subset contains only data of a single class. The decision tree then predicts the class of a new data point by following the path down the tree that corresponds to the values of its features.

# Random Forest Classifier

It is an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting a class prediction or regression value by averaging the predictions of the individual trees.

# Model Evaluation

In this phase, we'll evaluate models based on recall score and ROC_AUC. After, we will the best two models to tune them for better perfomance.

# Models Comparison - Recall Score

The recall score is a measure of how many of the positive instances the model correctly identifies. A higher recall score indicates that the model is better at identifying positive instances.

# Models Comparison - ROC Curve

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.

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

# Model Tuning

Based on the evaluation of the models using recall scores and ROC AUC, it is observed that theRandomForest classifier and the DecisionTree classifier have shown promising performance. To further improve their performance, model tuning can be performed using GridSearch.

# 1Tuning RandomForest and Tuning Decision Tree

Based on the ROC curve and the recall metric, the tuned Decision Tree model performs well in distinguishing between positive and negative classes (churned and non-churned customers) and in correctly identifying churned customers.The model has a recall score of 0.75, which means model is able to capture 75% of the actual churned customers.

# Conclusion.

The recall score of our DecisionTree was 75%. While this is still a good predictive model, we would like to undertake further feature engineering to boost this recall score if we had more time. We achieved our objectives to be able to predict customer churn and had an acceptable recall score.

**Releases**

No releases published
Create a new release

**Packages**

No packages published
Publish your first package

## Languages

- **Jupyter Notebook** 100.0%