PHASE 3 PROJECTS

Student name: Pupity Mwendwa

Student pace: Part time

Instructor name: Samuel Karu

Topic: SyriaTel Customer Churn Prediction.

SyriaTel Customer Churn Prediction

Project Task Build a binary classification model to predict whether a SyriaTel customer will churn (i.e. stop doing business) in the near future.

The questions this project seeks to answer include:

- What are the factors that are contributing in customers churning?
- What attributes do the customers who churn have?
- How can SyriaTel increase customer retention?

PROJECT OVERVIEW

- Business Problem
- Data loading and understanding
- Data preparation
- Modelling
- Evaluation
- Model of choice: Decision tree
- Recommendation and future investigation
- Conclusion

PROJECT OBJECTIVES

The main objective for this project is to build predictive model with 85% accuracy that will help to:

- Gain insight in the factors that are contributing to customer churning

- Gain insights on the attributes the customers who churn have increase customer retention

BUSINESS PROBLEM

I have been tasked by SyriaTel, a telecommunication company. They are seeking to build a classification model that can predict whether a customer is likely to stop doing business with them soon. The primary goal is to reduce financial losses due to customer churn.

BUSINESS QUESTIONS

The key questions that the project aims to address:

1. What are the factors that are contributing to customer churning? This question seeks to identify the specific reasons why customers are leaving SyriaTel.
2. What attributes do the customers who churn have? This question focuses on understanding the characteristics and behaviors of customers who are more likely to churn.
3. How can SyriaTel increase customer retention? This question aims to find effective strategies and solutions to prevent customers from leaving and maintain their loyalty.

DATA LOADING AND UNDERSTANDING

Discusses two findings that were discovered during the Exploratory Data Analysis (EDA) phase:

Finding 1: Converting data type

•The area code column is currently represented as an integer data type.

•However, the values in this column are essentially labels or placeholders, not numerical values that can be used for mathematical calculations.

•Therefore, it is necessary to convert the data type of the area code column to a categorical or ordinal data type to accurately represent its nature.

Finding 2: High correlation – Multicollinearity

•The heat map analysis revealed that several columns in the dataset exhibit high levels of correlation with each other.

•This indicates the presence of multicollinearity, a condition where independent variables are highly interrelated.

•Multicollinearity can cause issues in modeling, as it can make it difficult to determine the individual impact of each variable on the target variable.

•Addressing multicollinearity may be necessary before proceeding with the modeling process.

Heat map to check how the columns are correlated.

Heat map Interpretation: Colors in the heat map represent the correlation coefficient between features.

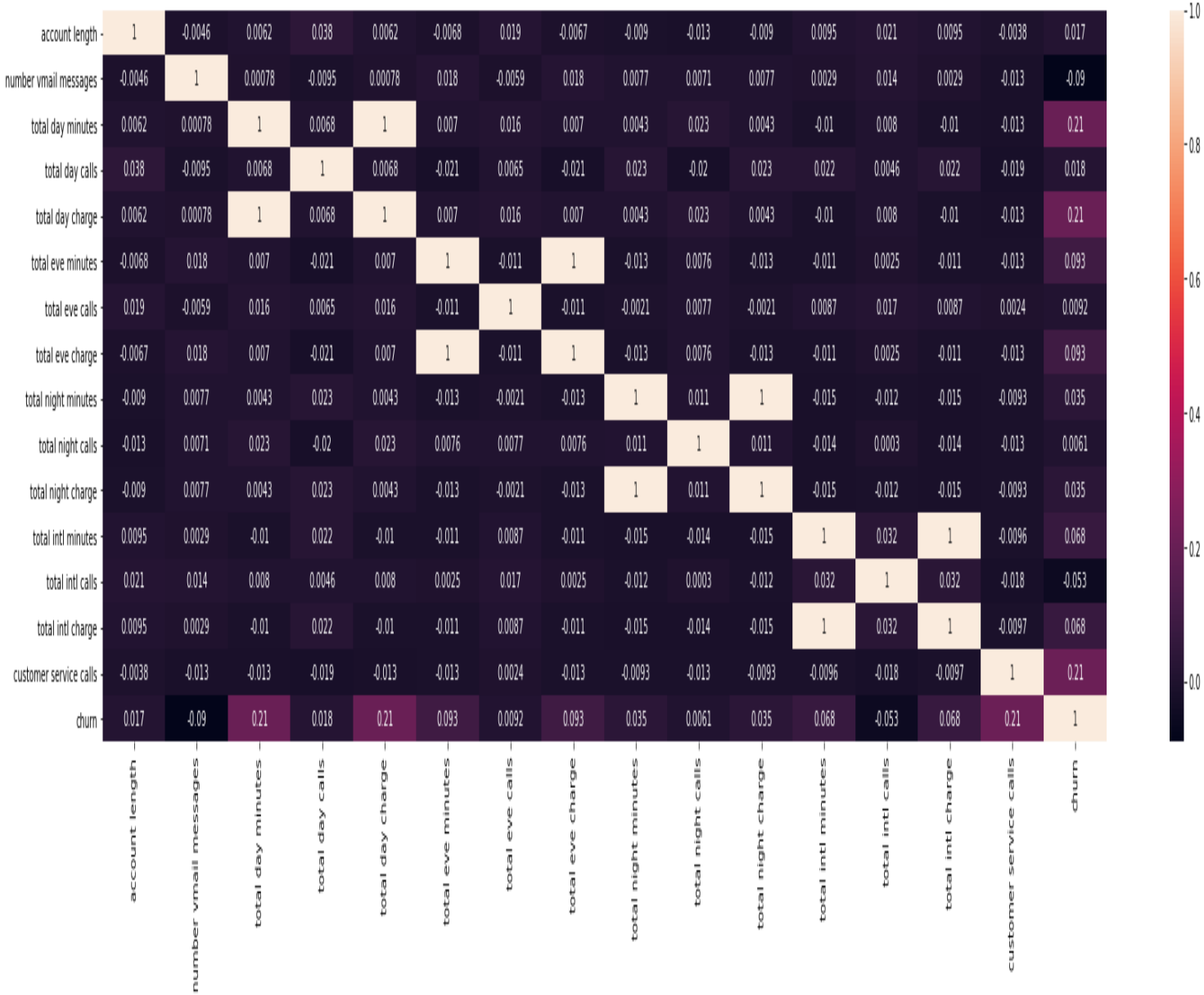
Red indicates a positive correlation, where higher values in one feature correspond with higher values in another feature.

Blue indicates a negative correlation, where higher values in one feature correspond with lower values in another feature.

White close to zero indicates no correlation between the features.

The intensity of the color represents the strength of the correlation.

Darker colors represent stronger correlations (positive or negative)

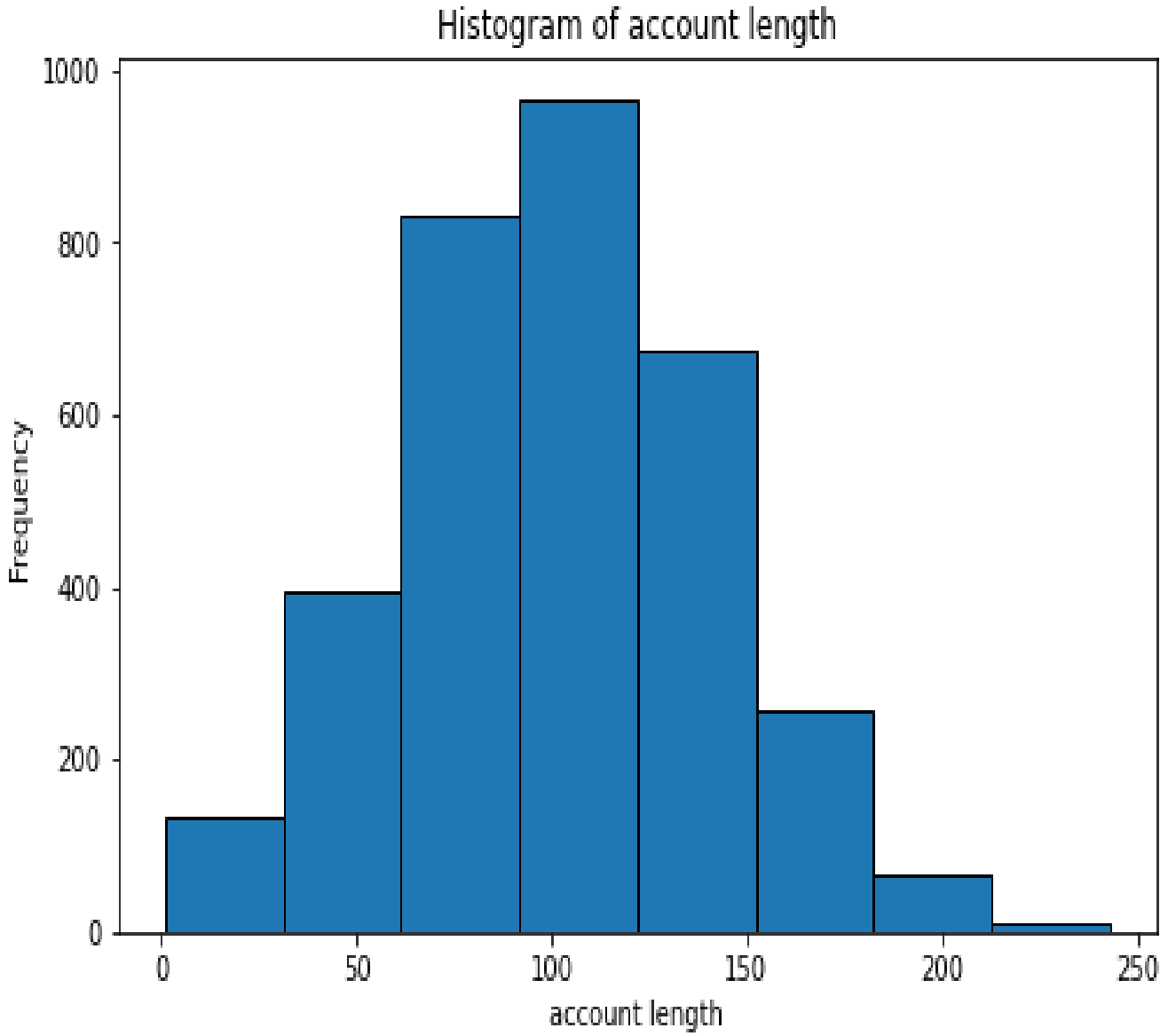


Histograms for numerical features to visualize their distribution

The histogram helps visualize the distribution of numerical features in the data.

By looking at the histograms, you can see if the features are normally distributed, skewed, or have outliers.

You can adjust the number of bins (bins) to achieve a better visualization of the distribution.

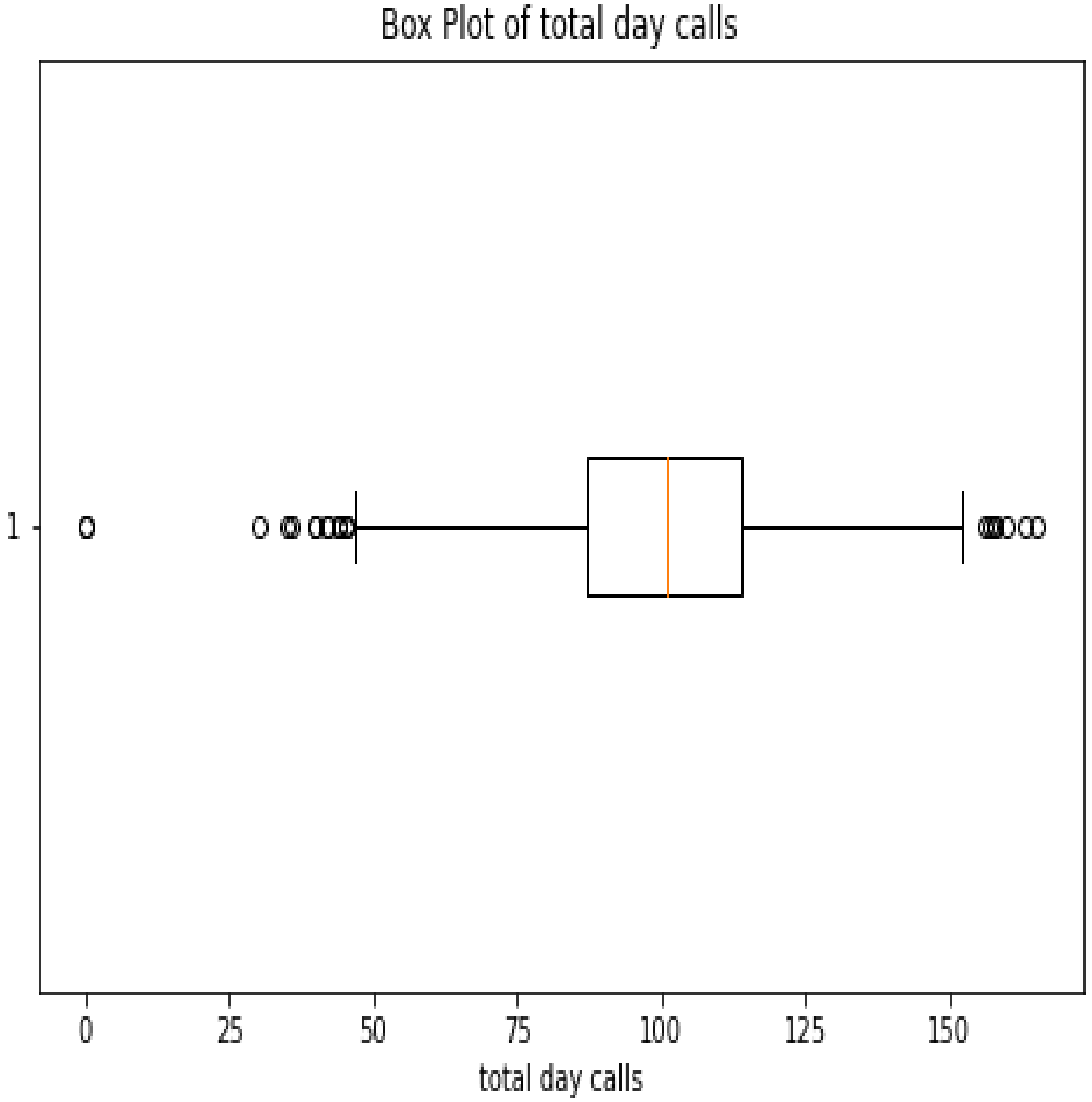


FINDINGS

We observed the presence of a significant number of outliers in our dataset, as indicated by the boxplots. These outliers have the potential to impact our modeling process. However, it is important to note that, in this case, these outliers are not anomalies that should be removed. Instead, they are a noteworthy aspect of our dataset that we should be aware of during our modeling process.

Box plots to identify outliers and visualize the spread of data.

The boxplot visualizes the distribution of the data for a single feature. The box part of the plot represents the interquartile range (IQR), which is the range between the first quartile (Q1) and the third quartile (Q3) of the data. The line in the middle of the box represents the median (Q2) of the data. The whiskers extend from the box towards the minimum and maximum values, up to 1.5 times the IQR. Points beyond the whiskers are considered outliers. By analyzing the boxplots, you can identify features that have outliers, skewness, or a large spread in their data.



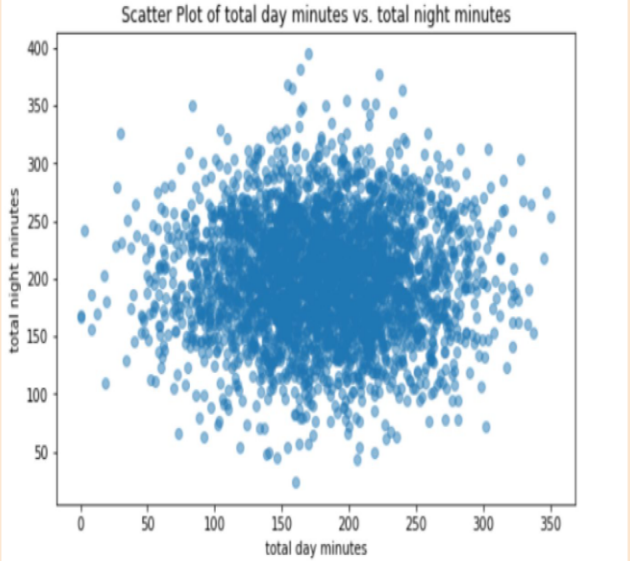
Scatter plots showing multicollinearity

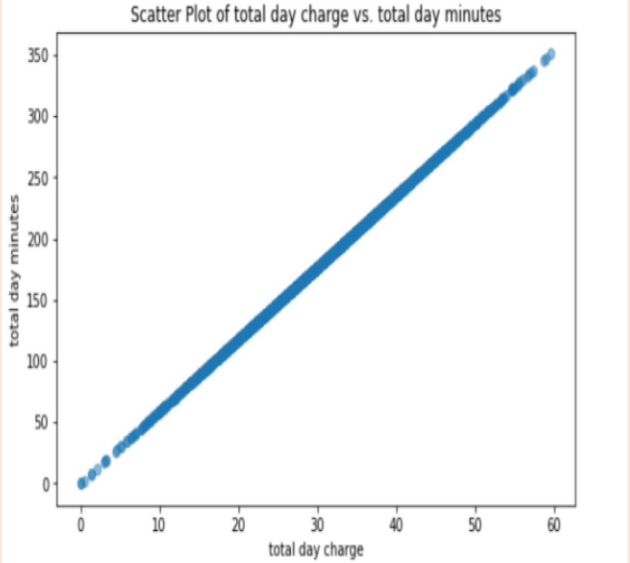
Perfect Correlation: The scatter plots have revealed features that exhibit a perfect correlation with each other.

Multicollinearity: This perfect correlation indicates multicollinearity, a situation where independent variables are highly interrelated, often moving in perfect synchronization.

Modeling Challenges: Multicollinearity poses a significant challenge to the modeling process and can lead to less reliable statistical inferences.

In essence, the analysis has identified a strong linear relationship between certain features, suggesting that they may be redundant or providing similar information.





Data Preparation

Below are some steps followed during data preparation

Step 1: Removing irrelevant columns Given that our dataset does not contain any missing values or duplicates, the next step is to streamline our data by removing columns that are not essential for our analysis or modeling.

Step 2: Feature Engineering In our dataset, we have identified that both our target variable and certain feature columns are categorical in nature. To effectively use this data in our modeling process, it is advisable to encode these categorical variables into a numerical format.

Encoding

Finding 5: Encoding

Encoding categorical variables is a crucial step, as many machine learning algorithms require numerical inputs for model training. By performing appropriate encoding techniques, such as one-hot encoding or label encoding, we can convert the categorical values into a numerical representation that the model can understand.

Finding 6: Imbalanced data

The 'churn' column represents a binary outcome, where 'False' represents customers who did not churn (i.e., stayed), and 'True' represents customers who did churn (i.e., left). The 'churn encoded' column is a numerical representation of 'churn,' where 0 typically corresponds to 'False,' and 1 corresponds to 'True'. The majority of the data (about 85.51%) falls into the 'False' or 0 category, while the minority of the data (about 14.49%) falls into the 'True' or 1 category. This indicates that the dataset might be imbalanced, with a higher proportion of non-churned customers.

MODELLING

Develop a predictive model designed to anticipate whether a customer is on the verge of discontinuing their engagement with Syria. The primary objective is to curtail financial losses stemming from customers who have a short-lived association with the entity.

Model Used:

- Logistic Regression
- Decision Trees
- KNN Classifier Model

EVALUATION

- The logistic regression model shows a balanced performance with reasonably good accuracy, precision, recall, and F1 score. It captures positive cases effectively while maintaining precision. The ROC AUC score is also decent.
- The decision tree model exhibits high accuracy, especially for the majority class. However, it has lower precision, recall, and F1 score for the minority class. This suggests that it may not perform as well on classifying the minority class.
- The model is well-suited for imbalanced datasets. The KNN classifier model achieves decent accuracy and performance for the majority class but struggles with the minority class, similar to the decision tree model.

Model of Choice: Decision Trees

- The decision tree model was evaluated with precision, recall, and F1 score for both Class 0 and Class 1.
- From the accuracy results above, our model correctly predicts the class labels for the majority of instances in the test data. The precision metric is very important as it measures how accurate the model is at identifying the majority class, which is the customers who don't churn.
- The downside of our model is that it has lower precision, recall, and F1 score for the minority class. This suggests that it may not perform as well on classifying the minority class, the same way it does the majority class. But this is majorly attributed to the fact that our data is also highly imbalanced. But compared to the other two models, this one performs much better.
- Our model performs well in terms of precision and recall for both classes on the holdout test data, thus it can be deployed in a real-world scenario.

Recommendations and Future Investigations

- Customer Service Calls Investigation: Dig deeper to understand why some customers need to contact customer service frequently. This will help in finding ways to better assist them.
- International Plan Churn Investigation: Since some of the customers with international plans are leaving, it's essential to explore ways to retain these customers.
- High Churn States Analysis: Look into the states where many customers are leaving to identify any patterns or reasons for the high churn rates.
- Incentives for High Bill Customers: Find ways to encourage customers with high daily charges (over $55) to stay with Syria Tel. This might involve offering extra benefits and perks. Currently, all of these high-bill customers are leaving, which is a concern.
- Incentives for Customers who stay more than 6 months: Find ways to encourage customers to stay with the company even longer, e.g., giving them loyalty points, offers, etc., as this will help in creating a form of loyalty.

CONCLUSION.

- In conclusion, the decision tree model appears to be providing reasonably good results, especially for the majority class, which is typical for imbalanced datasets and is the best option out of the three models built above.
- The decision tree model is the most suitable model for the given task, especially considering the imbalanced nature of the dataset. This means that the decision tree model is able to effectively predict the target variable for the majority class (i.e., the class with the most instances) while maintaining reasonable performance for the minority class.This is a common scenario in real-world datasets where one class is significantly more prevalent than the other. In such cases, models that are specifically designed to handle imbalanced data, like the decision tree model, often outperform other models.

THE End!!!

THANK YOU

NAME: Pupity Mwendwa

EMAIL: puritykimathi26@gmail.com

PHONE:0700314247.