PREDICTION MODEL ON SYRIA TEL CUSTOMERS CHURN

Overview

Syria Tel, a telecommunications company has commissioned an analysis to better understand the customers' preferences and predict whether a customer will stop doing business with the company or not. The aim of the project is to leverage historical customer data to identify key factors influencing customer behavior with a focus on predicting customer churn. Churn refers to when customers stop or discontinue their relationship with a business or a service and for this case telecommunication services.

DATA UNDERSTANDING

This dataset has been derived from Churn in Telecoms Dataset on Kaggle containing detailed information on customer usage, service preferences and interaction history. The dataset enables a comprehensive analysis of factors that may influence customer satisfaction, preferences and the likelihood of churn offering SyriaTel detailed recommendations to improve service offerings and maintain their customer base.

OBJECTIVE

Build a classification model to predict whether a customer will churn or not churn providing insights to the company to reduce dissatisfaction among customers and maintain their customer base.

DATA PREPARATION AND PRESENTATION

- Import libraries will enable manipulation of imported datasets.
- Once the libraries have been imported, load and review the dataset.
- **Check for duplicates** using the function duplicated()**.**value_counts() and missing values using the function isna()**.**sum() in the data frame. The data frame contains no duplicate entries or missing values across the columns.
- **Check for outliers using z-score**. The data shows that most of the z-scores are within the range of -3 to 3, indicating there is no significant outliers.
  Row 3 has relatively high evening usage as it is indicated in the total eve minutes and total eve charge with z-scores above 2.7, suggesting long evening calls.

EXPLANATORY DATA ANALYSIS(EDA)

- **Scaling of the numeric columns using z-score**

Scaling changes the units of the dataset to be uniform

- **Convert the target variable(churn) from True and False to binary**
- **Binary encoding for voice mail plan and international plan columns**

MODELING

a) Univariate Model

This refers to a model with one independent/ input variable to explain one dependent/output variable. The model for consideration in this case is Linear Regression model which will also be the baseline model.

Evaluate the Model Evaluation using confusion matrix

From the above analysis:

- True Negatives(TN) are 551.These customers who did not churn, and the model correctly predicted them as false.
- False Positives(FP) are 15.These customers who did not churn, but the model incorrectly predicted them as true.
- False Negatives(FN) 79. These customers who churned, but the model incorrectly predicted them as false.
- True Positives(TP) 22. These customers who churned, and the model correctly predicted them as true.
- **Determine the accuracy of the model**

The logistic regression model predicted approximately 85.45% correctly of the test set labels.

Determine the model's precision

For each positive prediction the model made, about 55.88% were actually correct positive cases.

Determine how many true churn cases were successfully detected by the model using the recall function

The model only predicted about 18.81% of correctly as churn.

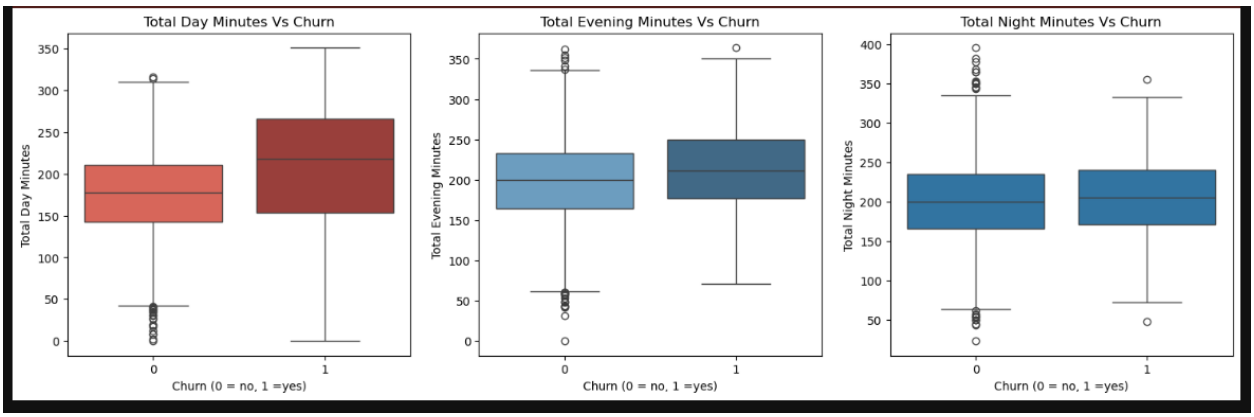
EVALUATION OF MODEL PERFORMANCE

From the above Logistic Regression model, its current performance shows that it not very accurate with detecting churn due to the low recall shown (18.81%). Improving the recall would be crucial for successfully predicting and having measures against customer churn.

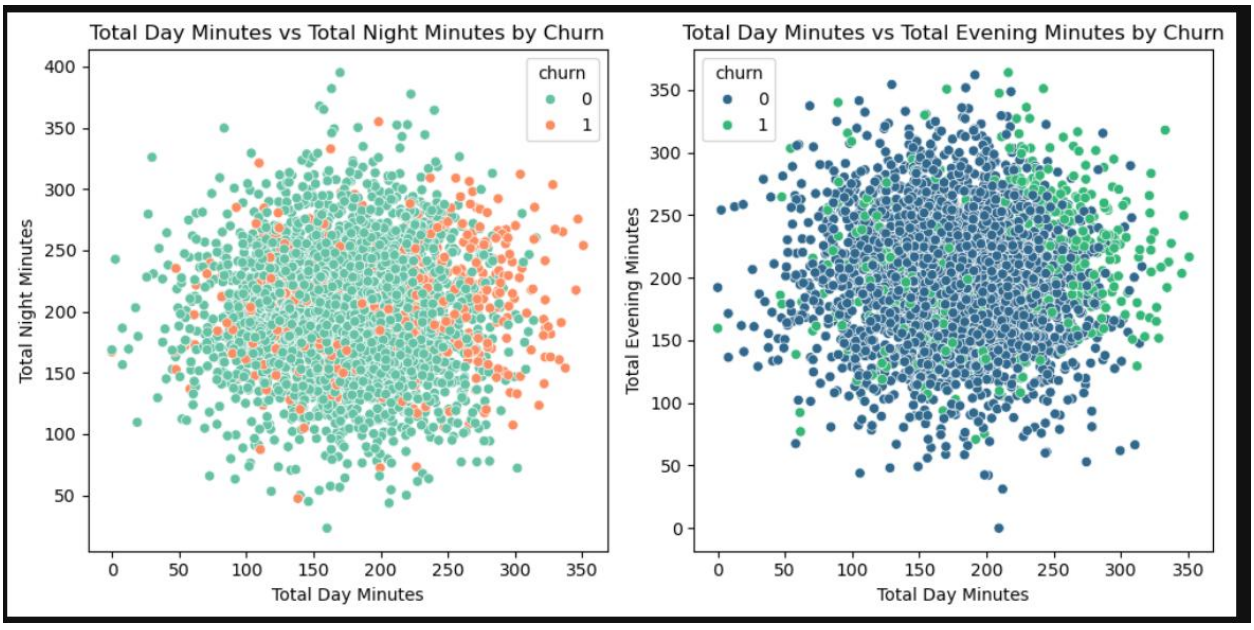
b) Bivariate Model

A bivariate model compares the relationship between two variables. To get a better understanding of customer call patterns and churn, plot boxplots of total day minutes, evening

call minutes and night call minutes side by side. This allows for visual comparison of call usage patterns differ from customers who churn and those who doesn't churn.



From the above box plots, day minutes have the strongest association with churn as individuals tend to have higher usage. Evening and night Minutes do not show significant differences for churn implying they may not be strong predictors. Plot a scatter diagram of Total day minutes, Total evening minutes and total night minutes vs churn to analyze the relationship between total day minutes, total evening minutes and total night minutes.



Both scatter plots show distinctively that customers have higher total day minutes, but the overlap in the data suggests that these features alone may not be sufficient to predict churn accurately. A decison tree model will be incorporated to illustrate the predictive relationship between total day minutes, total evening minutes, total night minutes and churn.

To begin with configure the DecisionTreeClassifier with hyperparameters to reduce overfitting.

Evaluate the Model

a) using confusion matrix

From the above analysis:

- True Negatives(TN) are 821.These customers who did not churn, and the model correctly predicted them as false.
- False Positives(FP) are 36.These customers who did not churn, but the model incorrectly predicted them as true.
- False Negatives(FN) 80. These customers who churned, but the model incorrectly predicted them as false.
- True Positives(TP) 63. These customers who churned, and the model correctly predicted them as true.

b) Determine the accuracy of the model

The logistic regression model predicted approximately 88.4% correctly of the test set labels.

c) Determine the model's precision

For each positive prediction the model made, about 63.64% were actually correct positive cases.

d) Determine how many true churn cases were successfully detected by the model using the recall function.

The model only predicted about 44.05% of correctly as churn.

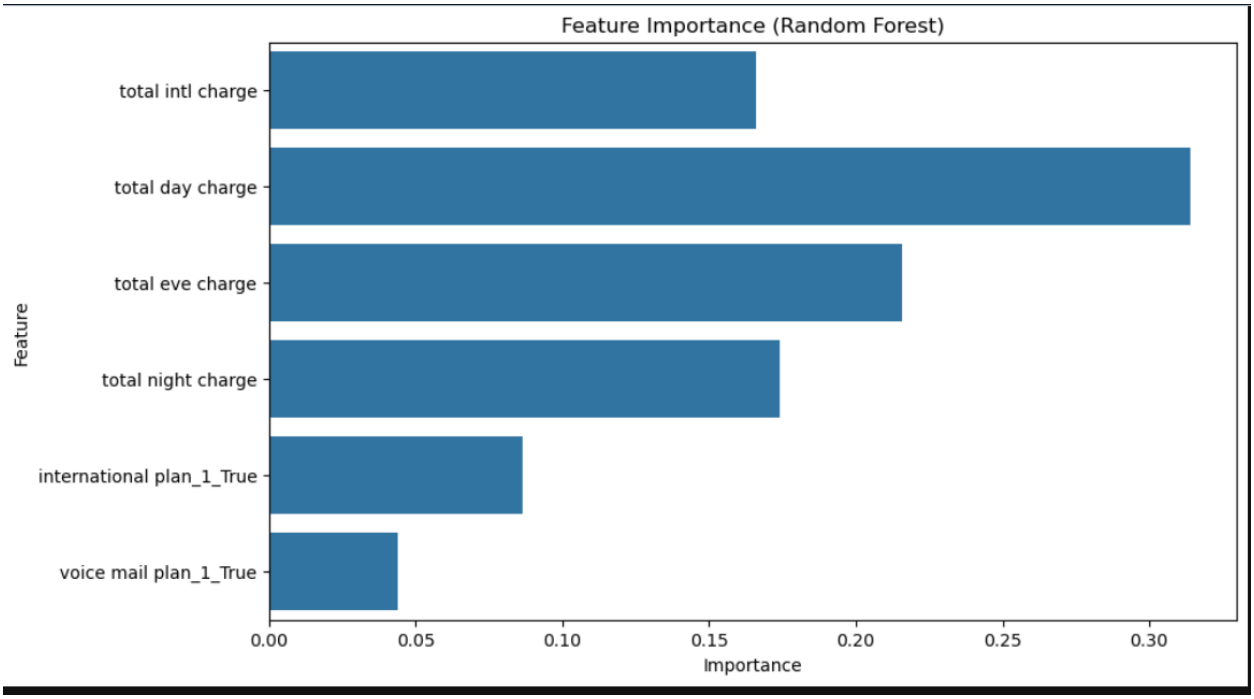
MODEL IMPROVEMENT THROUGH HYPERPARAMETER TUNING

Comparison between the baseline model(Logistic Regression) with Tuned Decision tree model

From the analysis of the tuned decision tree and the baseline model(Logistic Regression) there is an improvement in the recall value of 44.05% from 18.88% which clearly indicated that the model predicted about 44.05% of correctly as churn which could be acceptable for churn prediction.

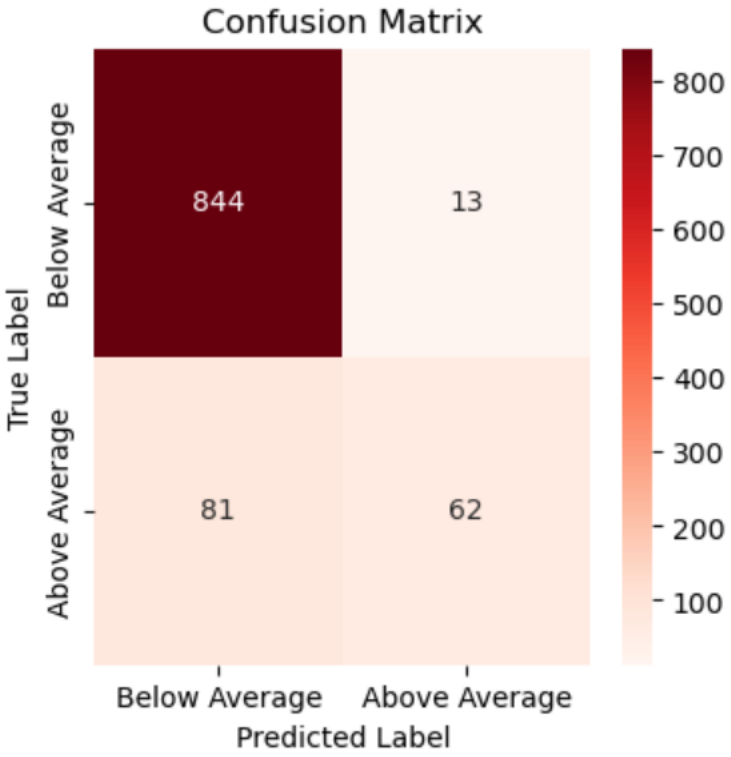
c) Multivariate Model

This refers to a model with multiple input features (predictors) to make predictions on the target. This model of consideration will analyze the relationship between features such as total charges for international, day evening and night calls, international plan, voice mail plan to predict the target variable customer churn. The model for consideration in this case is Random Forest classification model.



Model Evaluation

- **Accuracy -** The Random Forest model predicted approximately 90.6% correctly of the test set labels.
- **Precision** - For each positive prediction the model made, about 82.66% were actually correct positive cases.
- **Recall** - The model only predicted about 43.05% of correctly as churn.
- **Confusion matrix**



Comparison of Logistic regression model, decision tree model and random forest model side by side

	Model	Accuracy	Precision	Recall
0	Logistic Regression	0.866	0.745778	0.575165
1	Decision Tree	0.816	0.653227	0.691647
2	Random Forest	0.908	0.876757	0.713279

RECOMMENDATION

- From the above analysis of Logistic Regression model, Decision Tree Model and Random Forest Model, Random Forest having the highest accuracy of90.8% and strong precision 87.67% along with a good recall of 71.3% it should be selected as the most appropriate model for predicting churn. It provides the best balance between detecting churn and minimizing false positives.
- Logistic Regression model having a high accuracy 86.%6 and a relatively good precision, its recall of 57.5 is low. This suggests the model misses a good portion of the churn cases hence not giving a true record.

CONCLUSION

From the analysis and model evaluation, Random Forest Model had the best predictions being multivariate and including the key metrics of prediction.

NEXT STEPS

Integration of the Random Forest model into a customer relationship management system for real-time churn prediction for the different customer bases.