# CUSTOMER CHURN IN A TELECOMMUNICATION COMPANY


## Project Overview:

This project seeks to **build a classifier** to predict whether a customer will soon stop doing business with **SyriaTel,** which is a telecommunications company. It seeks to identify any predictable patterns among customers who have already left and use these features to predict customers who are likely to leave in the future.

**Author**: Samuel Kiio Kyalo

**GitHub repository:** https://github.com/Samuel-Kiio/Customer-Churn-in-Telecommunication-Company

# 1 BUSINESS UNDERSTANDING

## 1.1 Business Problem

Customer churn is the loss of clients or customers. Predicting churn helps the Telecom company to:

1. Identify at risk customers and implement ***highly targeted*** efforts to stop them churning.

2. Identify pain points and friction across a customer's journey.

3. Identify strategies that target these pain points to lower churn and increase retention rates.

**SyriaTel**, a telecommunication company is facing the problem of an increase in the number of customers who leave the company. A consistently high churn rate could result in the company quickly becoming unsustainable.

Attracting new customers as a strategy is not enough to sustain the company for very long.

It is therefore important for the company to increase the number of loyal and devoted customers by identifying the pain points across a company journey and accurately predicting the customers who are likely to churn and therefore targeting them using aggressive strategies to reduce these points of friction.

## 1.2 Objectives

The main objective for this project is **to build a prediction model that can accurately predict customers who are likely to churn from the company.**

This objective will be achieved through the methods outlined in the **specific objectives:**

1. To build a machine learning model that can accurately predict customers who will churn based on the features in the dataset.

2. To rank features these features in the dataset according to their order of significance

## 1.3 Success criteria

The model performances will be compared using the Recall (Sensitivity) score. This is given by: (TP/TP+FN).

From this, the aim is to reduce the number of false negatives, as it would be detrimental for the model to predict that customers will are not going to churn while they actually will.

A model with a recall of greater than 0.75 would be a successful model.

# 2 DATA UNDERSTANDING

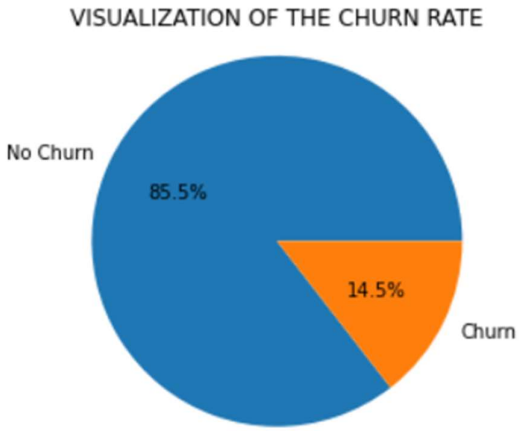
The dataset used is from SyriaTel Telecommunication company. Each row represents a customer, and the columns contain customer's attributes which are described in the following:

- **state:** the state the user lives in
- **account length**: the number of days the user has this account
- **area code:** the code of the area the user lives in
- **phone number**: the phone number of the user
- **international plan:** true if the user has the international plan, otherwise false
- **voice mail plan**: true if the user has the voice mail plan, otherwise false
- **number vmail messages:** the number of voice mail messages the user has sent
- **total day minutes:** total number of minutes the user has been in calls during the day
- **total day calls:** total number of calls the user has done during the day
- **total day charge**: total amount of money the user was charged by the Telecom company for calls during the day
- **total eve minutes**: total number of minutes the user has been in calls during the evening
- **total eve calls:** total number of calls the user has done during the evening
- **total eve charge:** total amount of money the user was charged by the Telecom company for calls during the evening
- **total night minutes:** total number of minutes the user has been in calls during the night
- **total night calls:** total number of calls the user has done during the night
- **total night charge:** total amount of money the user was charged by the Telecom company for calls during the night
- **total intl minutes:** total number of minutes the user has been in international calls.

- **total intl calls:** total number of international calls the user has done
- **total intl charge:** total amount of money the user was charged by the Telecom company for international calls
- **customer service calls**: number of customer service calls the user has done.
- **churn:** true if the user terminated the contract, otherwise false.

### 1.1 The Target Variable

The target variable is "churn". This project seeks to compare the effects of the different variables in respect to the 'churn'. To visualize the churn rate in the organization, a plot was created.




From the chart shown above, there are more samples for customers without churn than for customers with churn.

Therefore, there is **a class imbalance** for the target variable. This class imbalance could lead to predictive models which are biased towards the majority, i.e no churn.

### 2.1 The Predictor Variables

The predictors available to us are:

state, account length, area code, international plan, voice mail plan, number vmail messages, total day minutes, total dat 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 and customer service charge.

# 3 DATA PREPARATION

## 3.1 Detecting and dealing with missing values

There were 0 missing values in the dataset.

## 3.2 Data type conversions

The columns “international plan” and “voice mail plan” were converted to integer data types.

The “state” column was converted from categorical to a numeric column with a unique integer representing each category, by **LabelEncoding**

The column "state" contains categorical data. For this column, there are two options of dealing with these values:

1. **Using One-hot encoding.**
2. **Using LabelEncoder.**

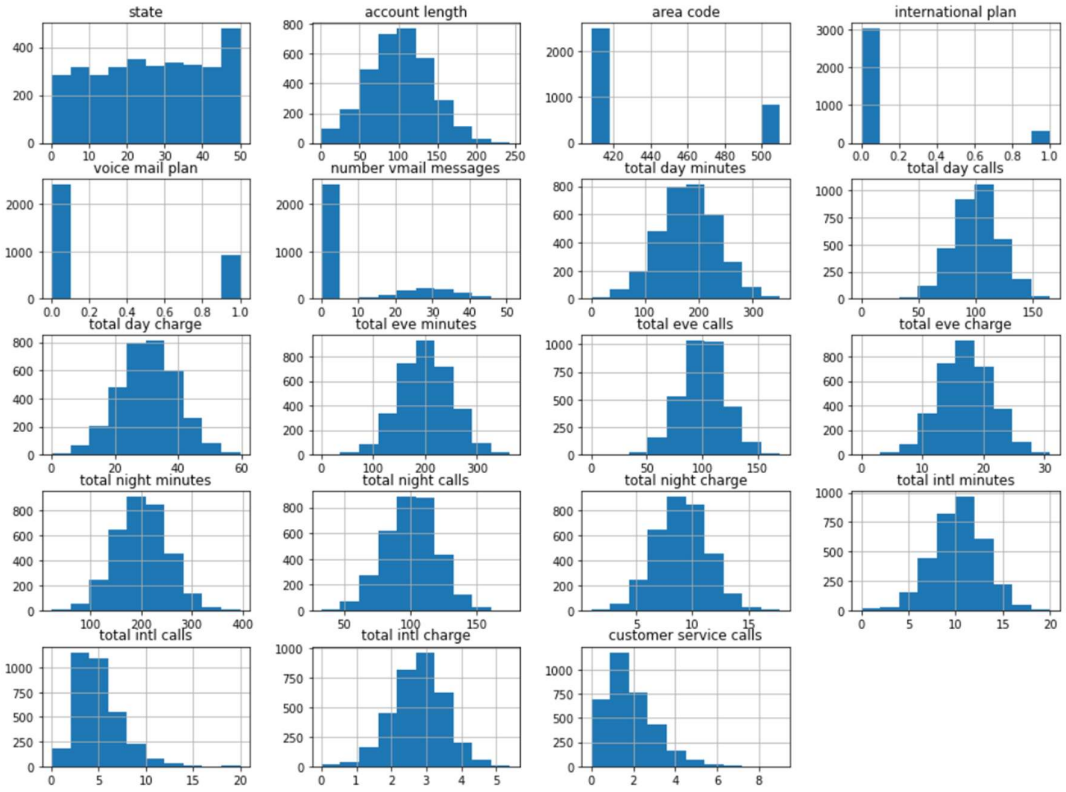
**LabelEncoder** was selected to replace each unique State with a unique integer. It was selected over One-hot encoding as there are 51 unique values that would increase the number of columns, causing the dataset to be messy and would affect the analysis of the importance of features in the models.

There were no duplicate values in the dataset.

## 3.3 Exploratory data analysis

### 3.3.1 Univariate analysis

Some graphs were created to visualize the distribution of different elements of the table:

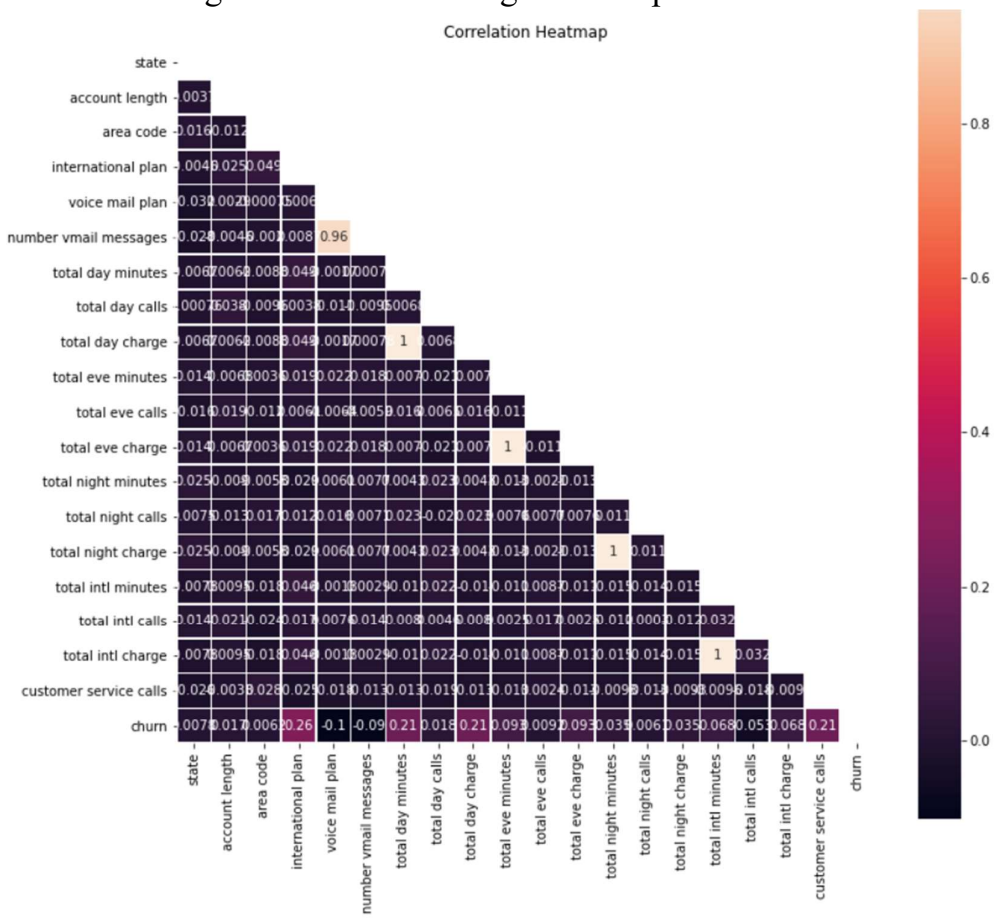



The plots displayed above show that:

1. The scale used across the features is different. Therefore, scaling techniques will be applied to ensure equal treatment of features during model training.
2. Features such as 'total intl calls', 'customer service calls are not normally distributed. They will be normalized.

### 3.3.2 Multivariate analysis

*3.3.2.1* Checking for correlation using a heatmap




From the heatmap above:

1. There is a very low correlation between most features.
2. Five pairs show high correlation. These are:

- Voice mail messages and voice mail plan (0.96)
- Total day charge and total day minutes (1)
- Total evening charge and total evening minutes (1)
- Total night charge and total night minutes (1)
- Total international charge and total international minutes (1)

These pairs of features display **multicollinearity**

The correlations can be explained by:

The **charges are proportional to the number of minutes** the customer spends on a call. Therefore, these **charges depend on the number of minutes a customer spends on a call** and therefore, they can be **dropped.**

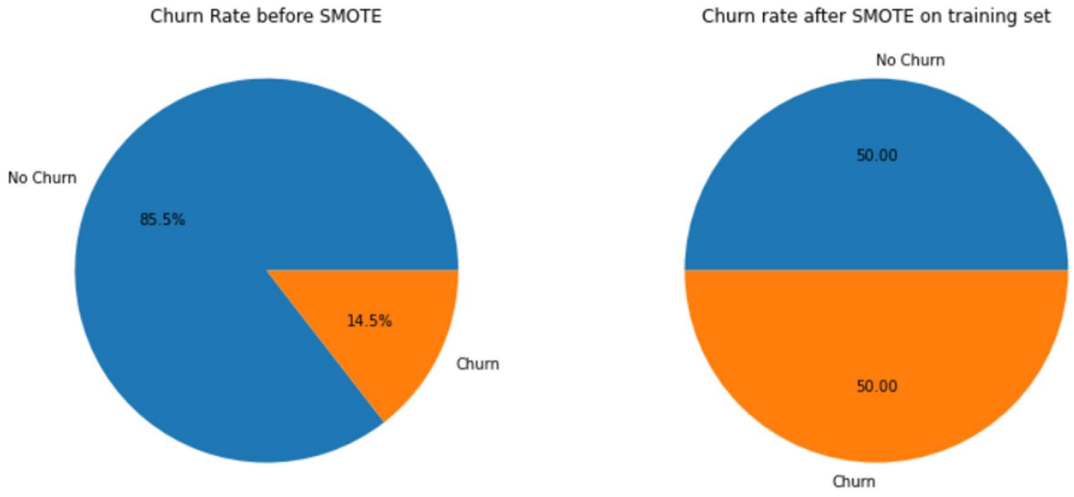
### 3.4 Train-test split

A train test split was then performed. The split was performed in the ratio of a test set of 25% and a training set of 75% were selected.

### 3.5 Handling the class imbalance problem

To handle the class imbalance problem, the Synthetic Minority Over-Sampling Technique (SMOTE) was used.

The result was as shown below:



# 4 MODELING

The project is a **binary classifier task**. To solve this problem, the models that will be tried are:

1. Logistic regression
2. K-Nearest Neighbours
3. Decision Trees
4. Random Forest
5. Support Vector Machine

The testing set therefore maintained an accurate distribution of the sample dataset.

## 4.1 Base Model: Logistic Regression

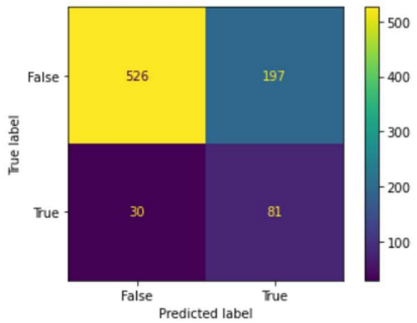
The results of the logistic regression are as shown in the diagram below:

```
Recall Score:
Train: 0.7860836859426422
Test: 0.7297297297297297

Precision Score:
Train: 0.7339771729587358
Test: 0.29136690647482016

Accuracy Score:
Train: 0.7505876821814763
Test: 0.7278177458033573

F1 Score:
Train: 0.7591373439273553
Test: 0.41645244215938304
```




From the results displayed above:

- The metrics displayed by the Recall score, the Precision score the Accuracy and the F1 score are higher in all instances for the Training set compared to the Testing set.
- This shows that the model is overfitting.

To improve the predictions, other types of models were created. These include:

1. KNN (K Nearest Neigbours)
2. Decision Tree
3. Random Forest
4. Support Vector Machine

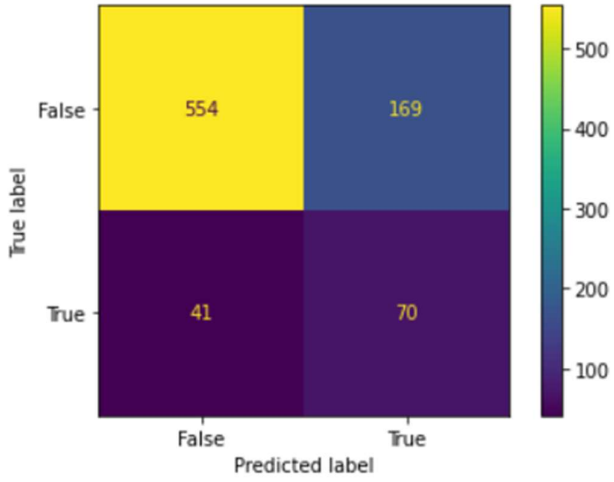
### 4.2 K Nearest Neigbours Model

```
Recall Score:
Train: 0.9797837329572168
Test: 0.6306306306306306

Precision Score:
Train: 0.872331519464211
Test: 0.2928870292887029

Accuracy Score:
Train: 0.918194640338505
Test: 0.7482014388489209

F1 Score:
Train: 0.9229406554472985
Test: 0.4
```




The KNN model **performs worse than the baseline model** on both the recall and the F1 score.

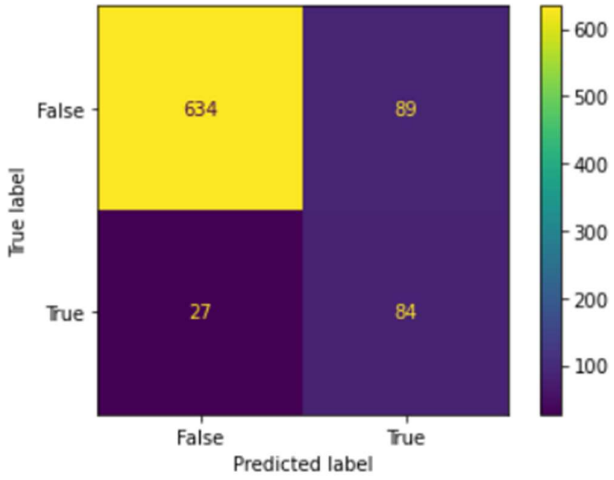
## 4.3 Decision Tree Model

```
Recall Score:
Train: 1.0
Test: 0.7567567567567568

Precision Score:
Train: 1.0
Test: 0.48554913294797686

Accuracy Score:
Train: 1.0
Test: 0.8609112709832134

F1 Score:
Train: 1.0
Test: 0.5915492957746479
```




The decision tree model **performs better than both the base model and the KNN model in all the metrics.**

The Decision Tree Model was then tuned, and the results obtained as shown below:

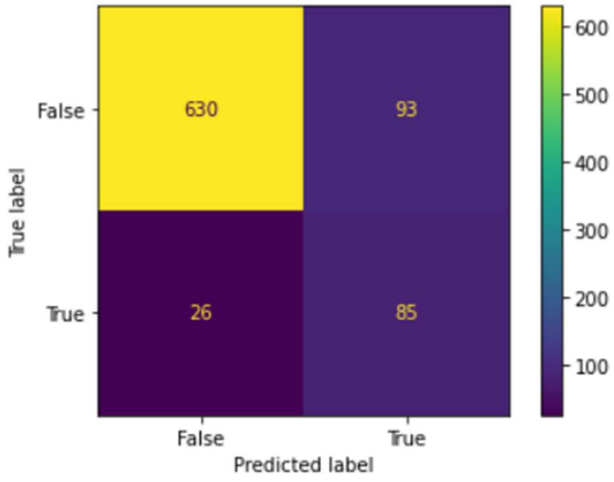
## 4.4 Tuned Decision Tree Model

```
Recall Score:
Train: 0.9793135872120358
Test: 0.7657657657657657

Precision Score:
Train: 0.9806967984934086
Test: 0.47752808988764045

Accuracy Score:
Train: 0.9800188058298073
Test: 0.8573141486810552

F1 Score:
Train: 0.9800047047753468
Test: 0.5882352941176471
```




The tuned model **performs better than the base model, the KNN model and the untuned decision tree model in the recall score.**

## 4.5 Random forest model

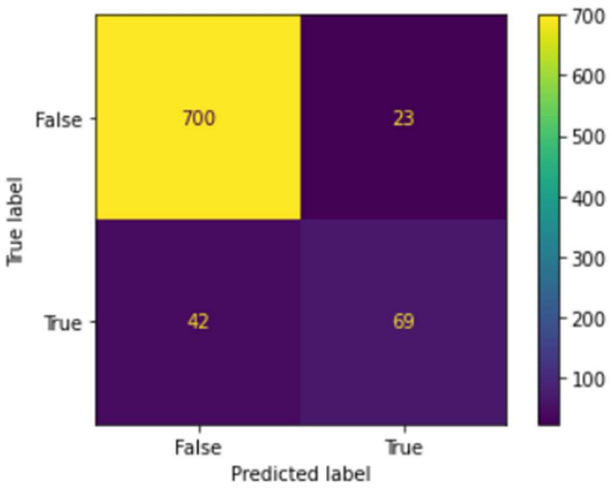
Random forest is used since it is naturally resistant to noise and variance.

```
Recall Score:
Train: 1.0
Test: 0.6216216216216216

Precision Score:
Train: 1.0
Test: 0.75

Accuracy Score:
Train: 1.0
Test: 0.9220623501199041

F1 Score:
Train: 1.0
Test: 0.6798029556650247
```




Random Forest model **performs poorer in the Recall score than the base model, the KNN model and the decision tree models**.

It has a high number of False negatives

The Random Forest model was then tuned and the results obtained as shown below:

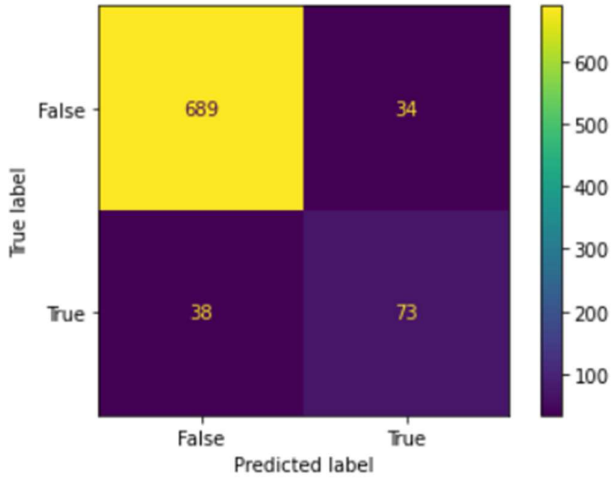
## 4.6 Tuned Random Forest Model

```
Recall Score:
Train: 0.9647390691114246
Test: 0.6576576576576577

Precision Score:
Train: 0.987487969201155
Test: 0.6822429906542056

Accuracy Score:
Train: 0.9762576398683592
Test: 0.9136690647482014

F1 Score:
Train: 0.9759809750297266
Test: 0.6697247706422017
```




On hyperparameter tuning, the recall increases by 2.8%. The model **performs poorer in the Recall score than both the base model and the Decision tree model.**

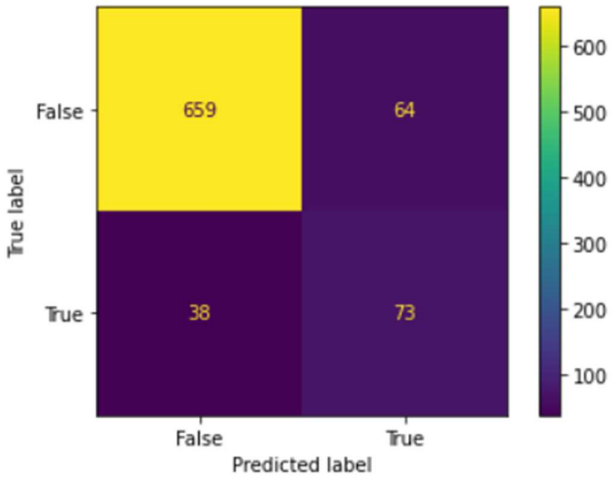
## 4.7 Support Vector Machine model

```
Recall Score:
Train: 0.9050305594734368
Test: 0.6576576576576577

Precision Score:
Train: 0.941320293398533
Test: 0.5328467153284672

Accuracy Score:
Train: 0.924306535025858
Test: 0.8776978417266187

F1 Score:
Train: 0.9228187919463088
Test: 0.5887096774193549
```

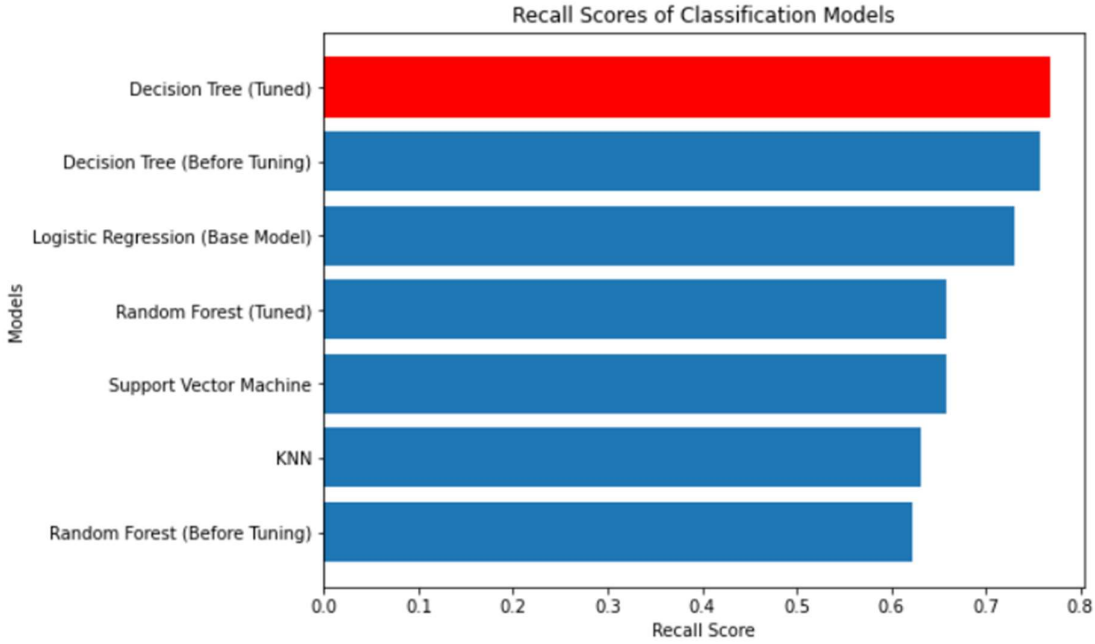



The performance of the model **is similar to the Random Forest model in the Recall score metric**

# 5 Model evaluation

## 5.1 Results

## 5.2 The performances of the models:

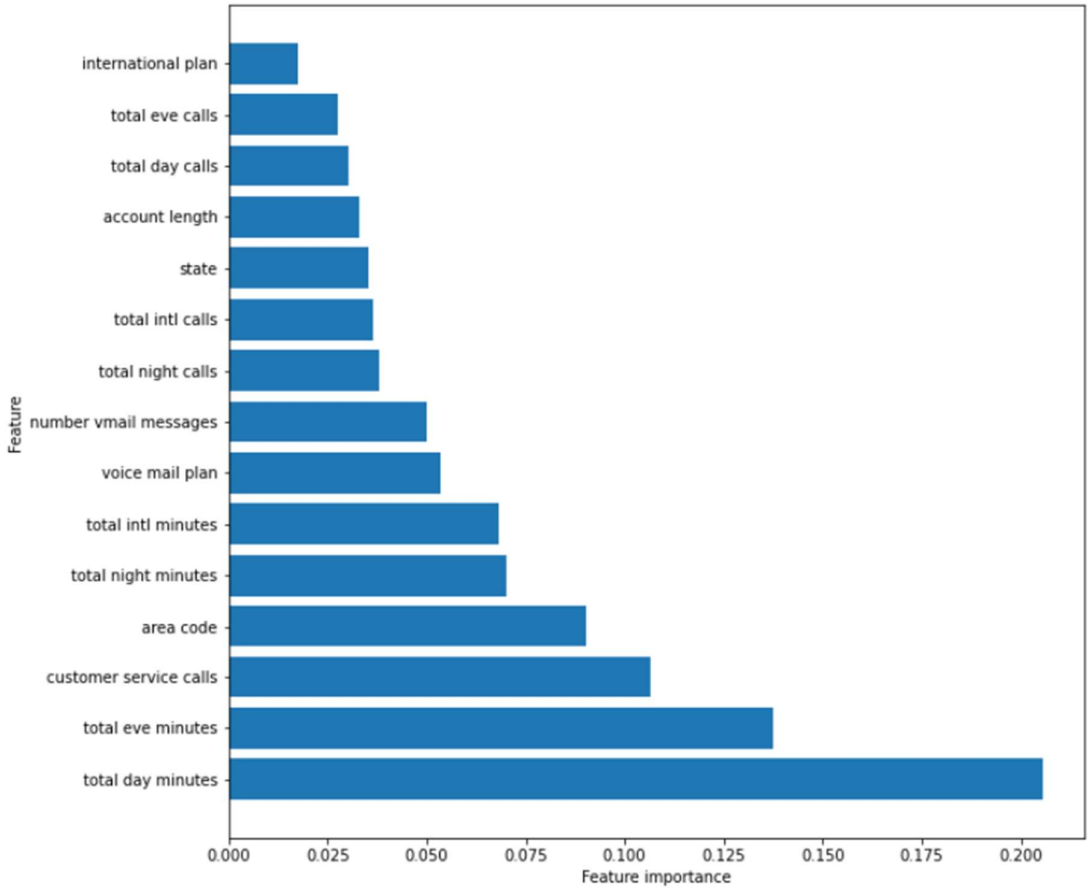



- The **recall score** was chosen as the ideal evaluation metric.
- This is because the goal of this project was **to maximize the number of predictions of customers who are likely to churn**. This means **minimizing the number of false negatives**, i.e the number of customers that the model predicts will not churn but they actually churn.
- The problem is less sensitive to false positives at the expense of false negatives. This is because the problem suggests **that the model would rather predict that a customer will churn and they fail to churn than predict that a customer will not churn and they actually end up churning.**
- The latter would be more **detrimental to the company financially than the former** and thus the choice to select Recall score as the **ideal comparison metric.**

The **tuned Decision Tree Model** performed the **best.**

Therefore, **it was selected**

## 5.3 A rank of features in the dataset according to their order of significance




From the chart shown above, the top 5 most important features in the model prediction are:

1. Total Day minutes
2. Total Evening minutes
3. Number of customer service calls
4. The area code they live in.
5. The total night minutes.

This shows that **the customers who make spend the most amount of time in calls and also make the highest number of customer service calls are most likely to churn.**

# 6 CONCLUSIONS AND RECOMMENDATIONS

The **objective of the project** was **to identify clients who are likely to churn**, in order to deploy strategies that reduce the pain points of customers and increase retention.

**The data** was drawn from the CyriaTel Telecommuication company.

This data was **analysed and cleaned.**

**Class imbalance** was found in the target variable.

This was addressed by the use of the **SMOTE technique** in the test set.

Different classification models were tried out including:

- Logistic regression
- KNN
- Decision tree
- Random Forest
- Support Vector Machine.

The **Decision tree model** offered the **best model.**

**In conclusion,**

1. The **model that meets the success criteria**, in predicting customers who will churn, of a sensitivity/ Recall score of 0.75 is the **Tuned decision tree model.**

2. **The most important feature in predicting the churn rate** is the **number of minutes** a customer spends on a call coupled with the **number of customer service calls.**

# 7 NEXT STEPS

The next steps in this project would be **to investigate how the area code ranks so highly in feature importance** and how it affects customer churn. This could be due to:

1. An area experiencing unstable network connection issues due to poor coverage
2. A lagrer population in certain areas that could cause system overload on existing infrustructure.
3. Demographic issues in a certain area which are being addressed by a competing telecom company.