SYRIATEL CUSTOMER CHURN ANALYSIS

BY: MARY MWANGI

OVERVIEW

- INTRODUCTION
- EXPLORATORY DATA ANALYSIS
- MODELLING
- MODEL TUNING
- MODEL EVALUATION AND COMPARISON
- FINDINGS, RECOMMENDATIONS AND CONCLUSION

INTRODUCTION

BUSINESS UNDERSTANDING

SyriaTel, a leading telecommunications company, is grappling with the challenge of customer churn. Customer churn, is a situation where subscribers discontinue services, posing a significant financial threat to SyriaTel. As the telecom industry evolves, retaining customers becomes paramount for sustaining revenue and growth.

OBJECTIVES

- **Main Objective**

To develop a predictive model that effectively identifies customers at risk of churning for SyriaTel

- **Specific objectives**

-To Identify the primary factors influencing customer churn in the SyriaTel dataset.

-To build a robust machine learning model for binary classification to predict customer churn.

-To extract meaningful insights from the model to guide SyriaTel in implementing effective retention strategies.

EXPLORATORY DATA ANALYSIS

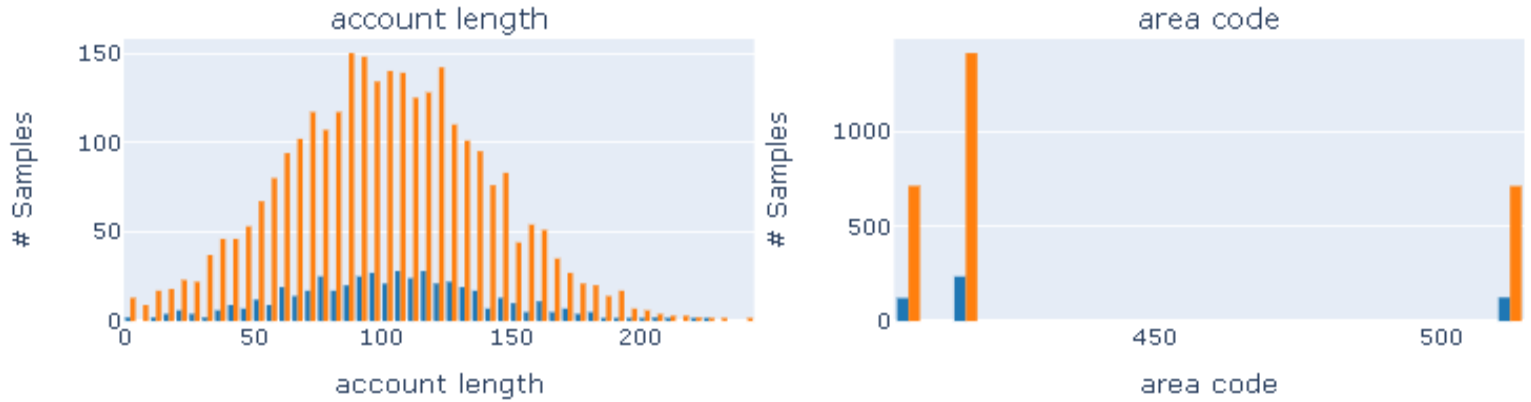
DATA SOURCE AND DESCRIPTION

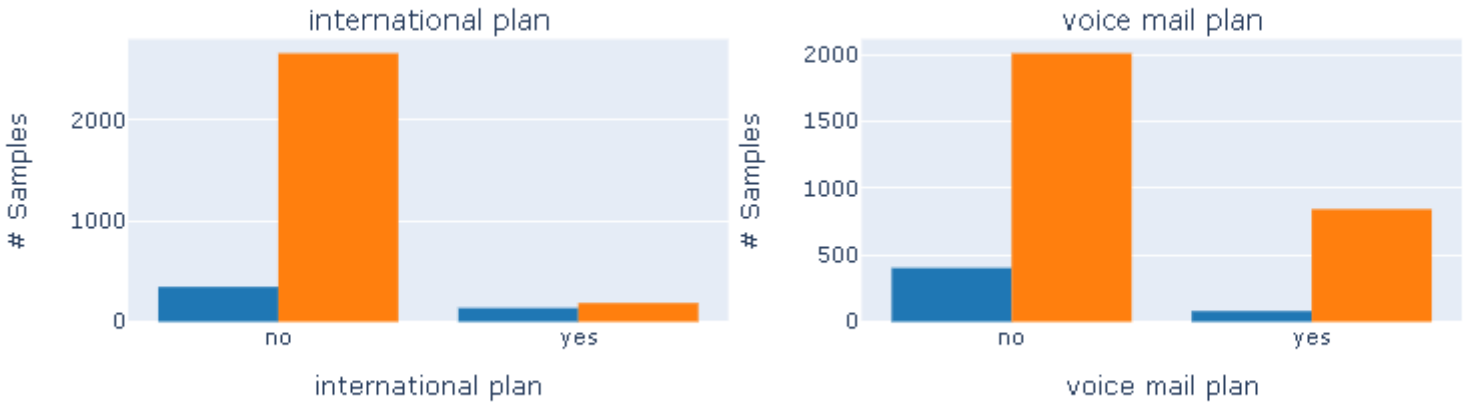
- The dataset has 3333 rows and 21 columns

columns

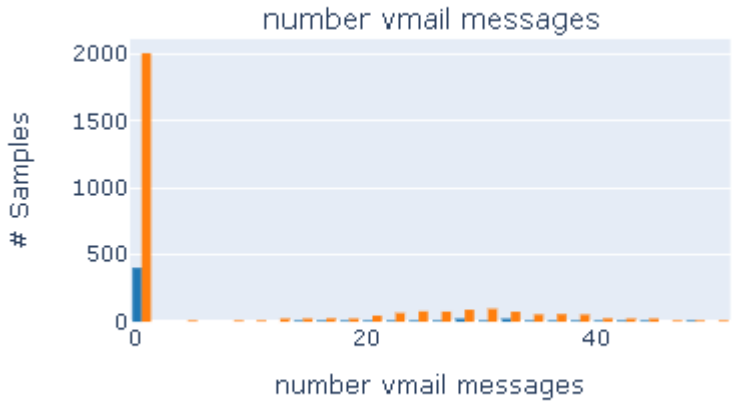
- state, account length, area code, phone number, international plan, voice mail plan, number vmail messages, total day minutes, total day calls, total day charge, total eve minutes, total eve calls, total eve charge, total night minutes, total night calls, total night charge, total intl minutes, total intl calls, total intl charge, customer service calls, churn.

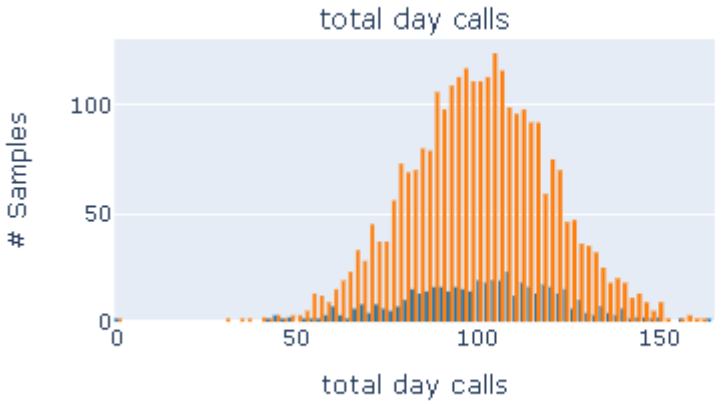
VISUALIZATION

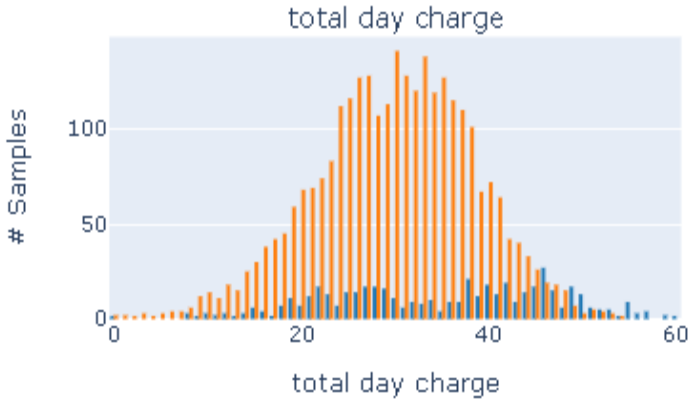


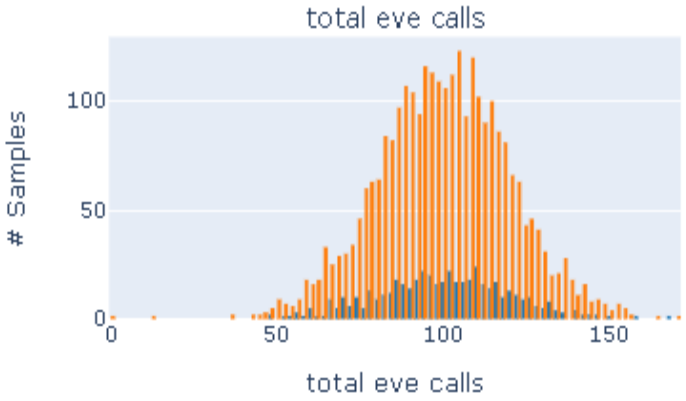


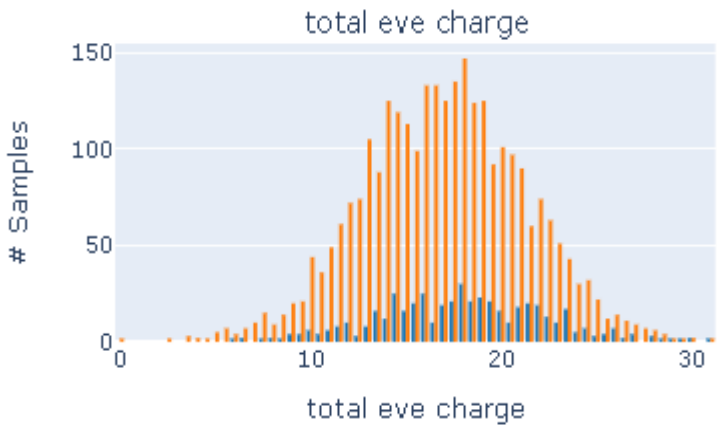
- Interactive graphs displaying the distribution of each feature for customers with churn and those without churn
- Churn is represented by blue, no churn by orange

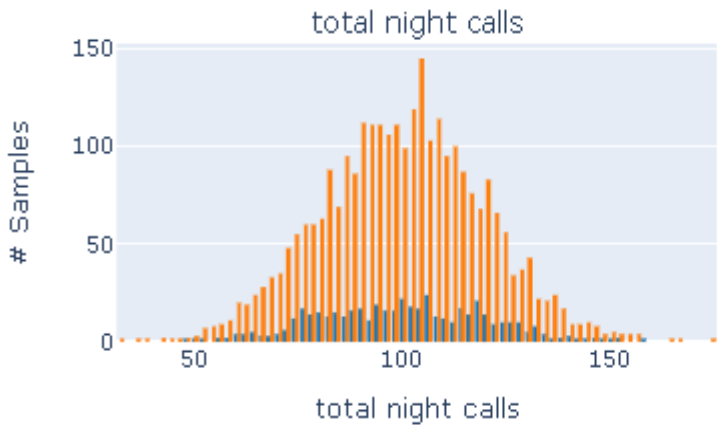

number vmail messages
Samples
2000
1500
1000
500
0
0
20
40
number vmail messages

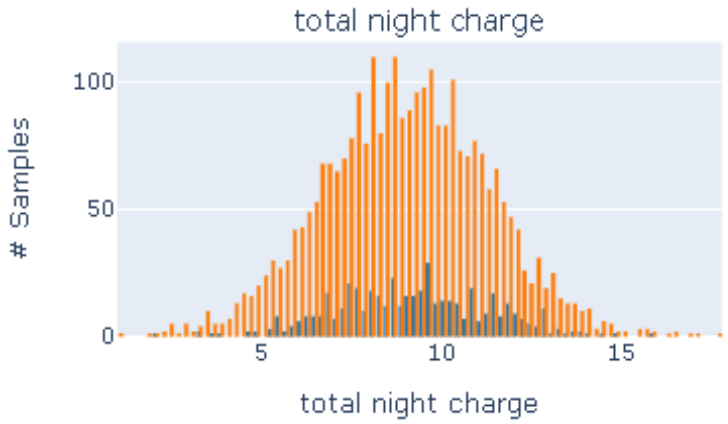

total day calls
Samples
100
50
0
0
50
100
150
total day calls

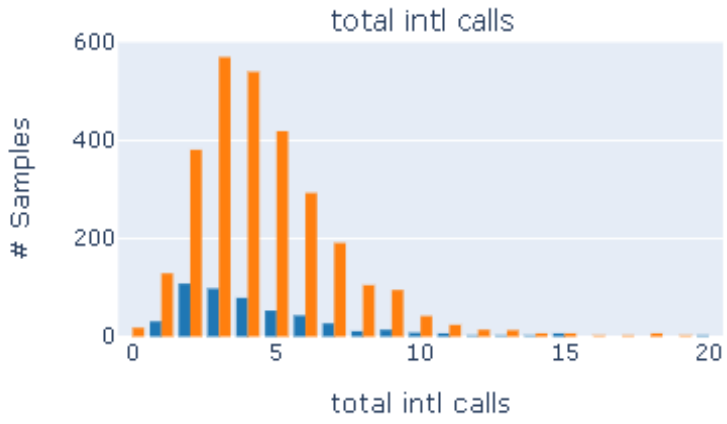

total day charge
Samples
100
50
0
0
20
40
60
total day charge

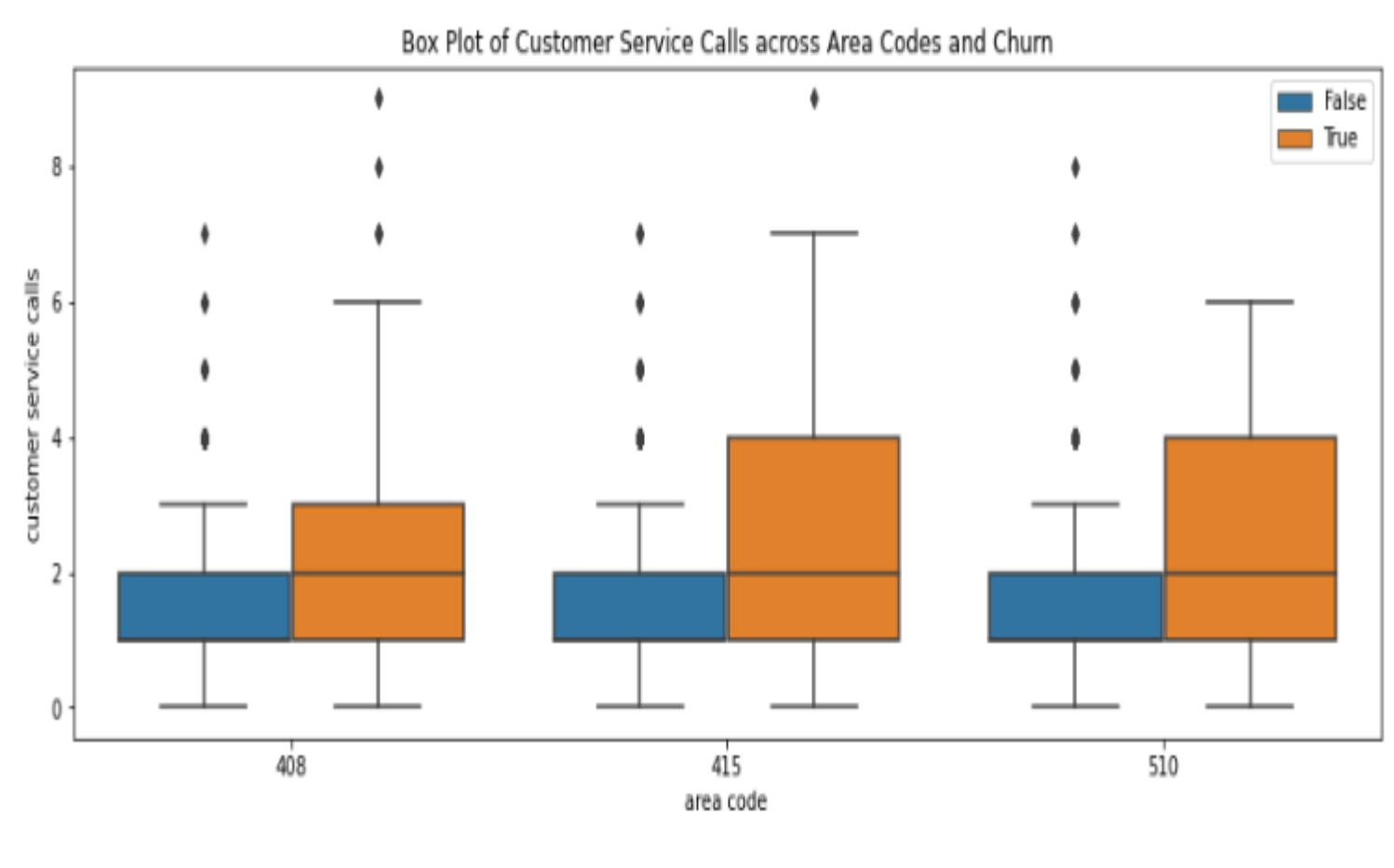

total eve calls
Samples
100
50
0
0
50
100
150
total eve calls


total eve charge
Samples
150
100
50
0
0
10
20
30
total eve charge


total night calls
Samples
150
100
50
0
50
100
150
total night calls

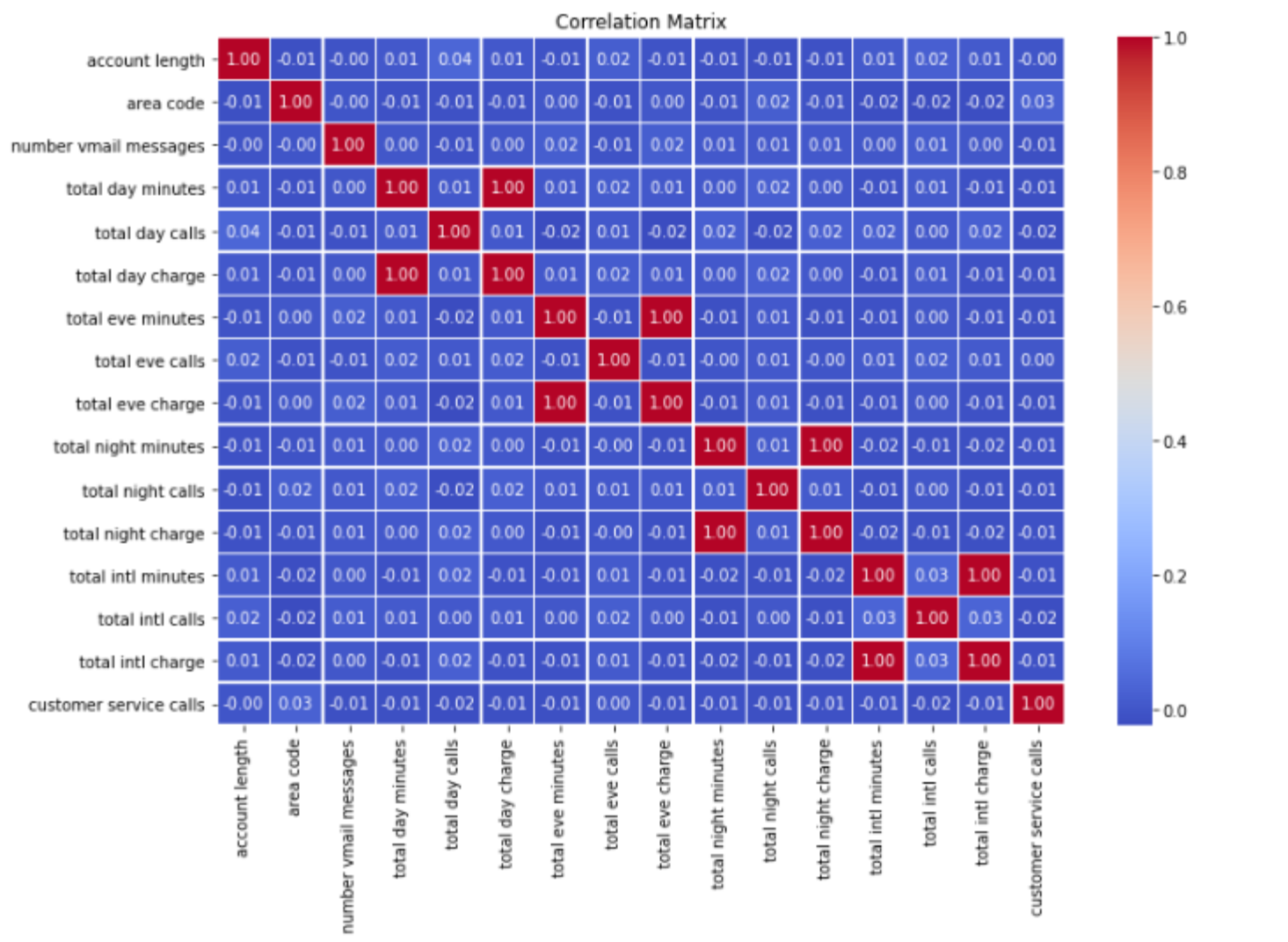

total night charge
Samples
100
50
0
5
10
15
total night charge


total intl calls
Samples
600
400
200
0
0
5
10
15
20
total intl calls



- Boxplots showing the distribution of churn across different area codes

This shows that customers who are likely to have stopped doing business with SyriaTel are within the area code 415 and 510



- heatmap showing the relationship between different numeric variables

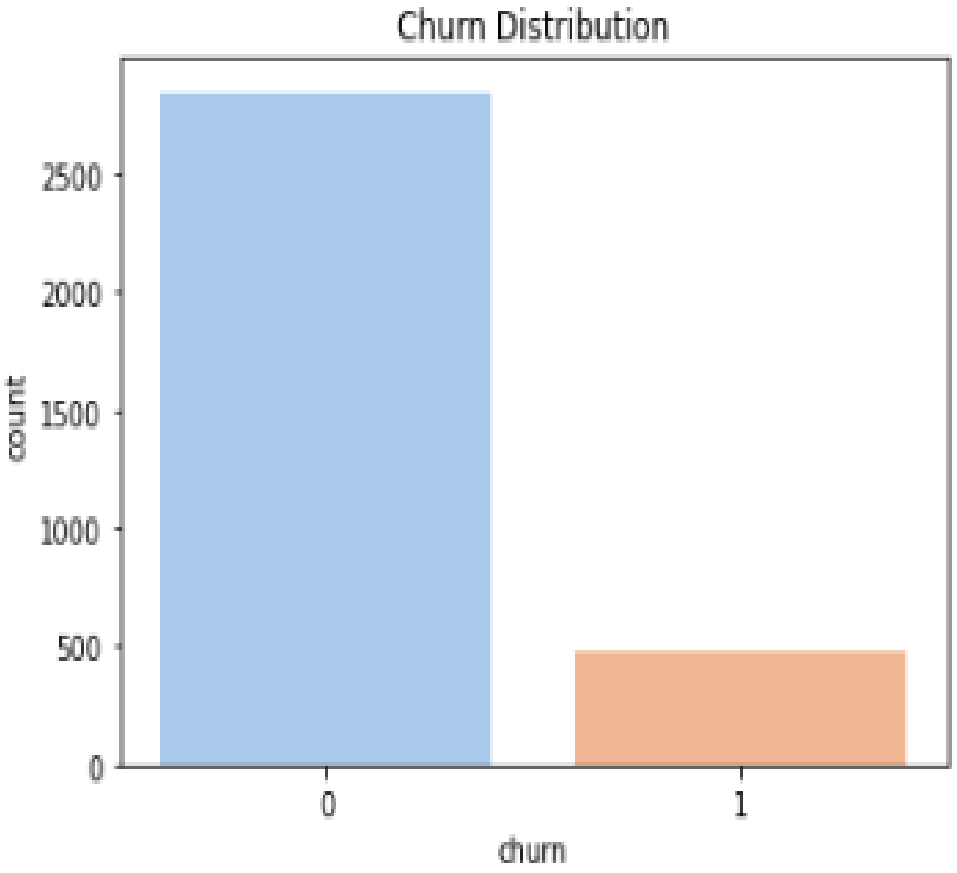
They is a low correlation between most of the features. However, a perfect positive correlation exists between:

-total day charge and total day minutes,

-total evening charge and total evening minutes,

-total night charge and total night minutes,

-total intl charge and total intl minutes



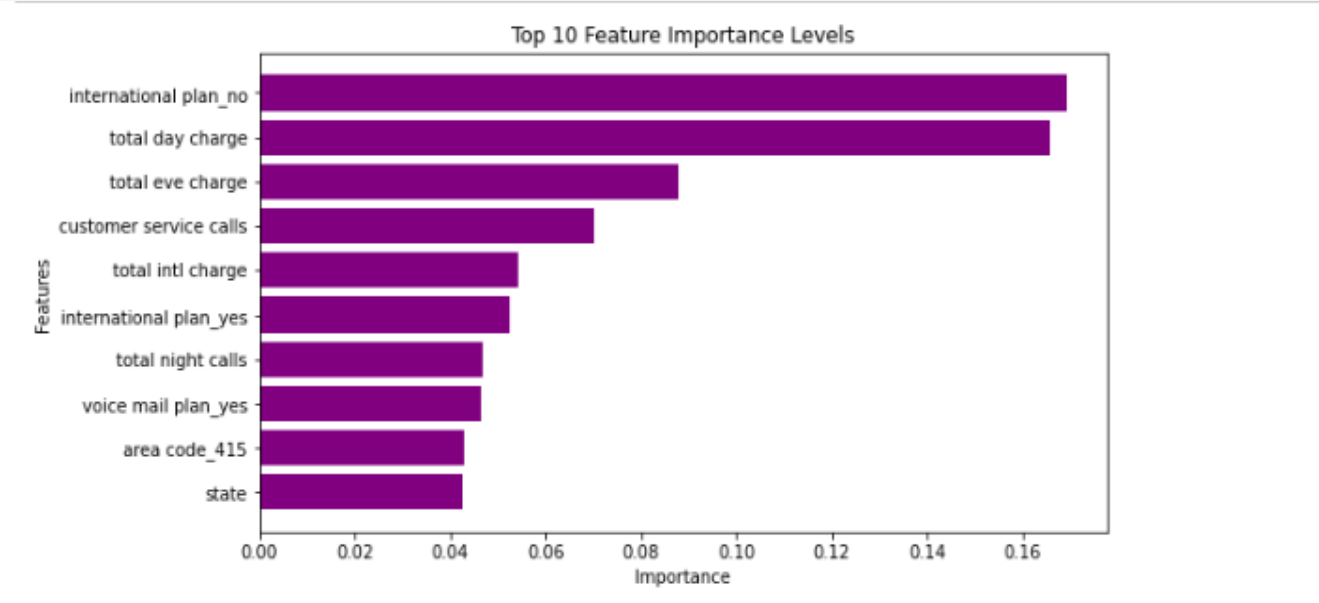
- Bar plot showing how churn has been distributed. With 0 representing customer who have not churned, meaning that they have retained or they are still active,
- while 1 represents customers who have churned.

MODELLING

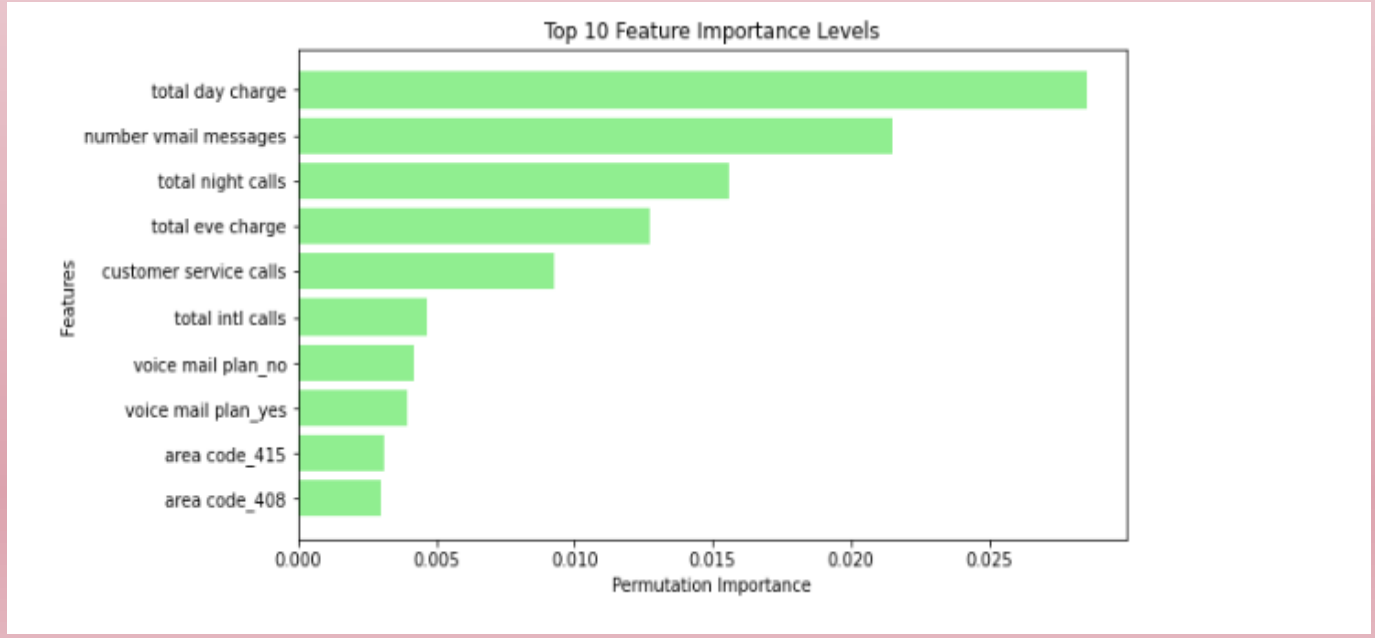
	Logistic Regression	Decision Tree Classifier	K-Nearest Neighbors	Random Forest
precision	0.37306	0.61818	0.21429	0.83019
Recall	0.576	0.816	0.504	0.704
F1-score	0.45283	0.70345	0.30072	0.7619
Accuracy	0.79137	0.89688	0.64868	0.93405

AFTER HYPERPARAMETER TUNING

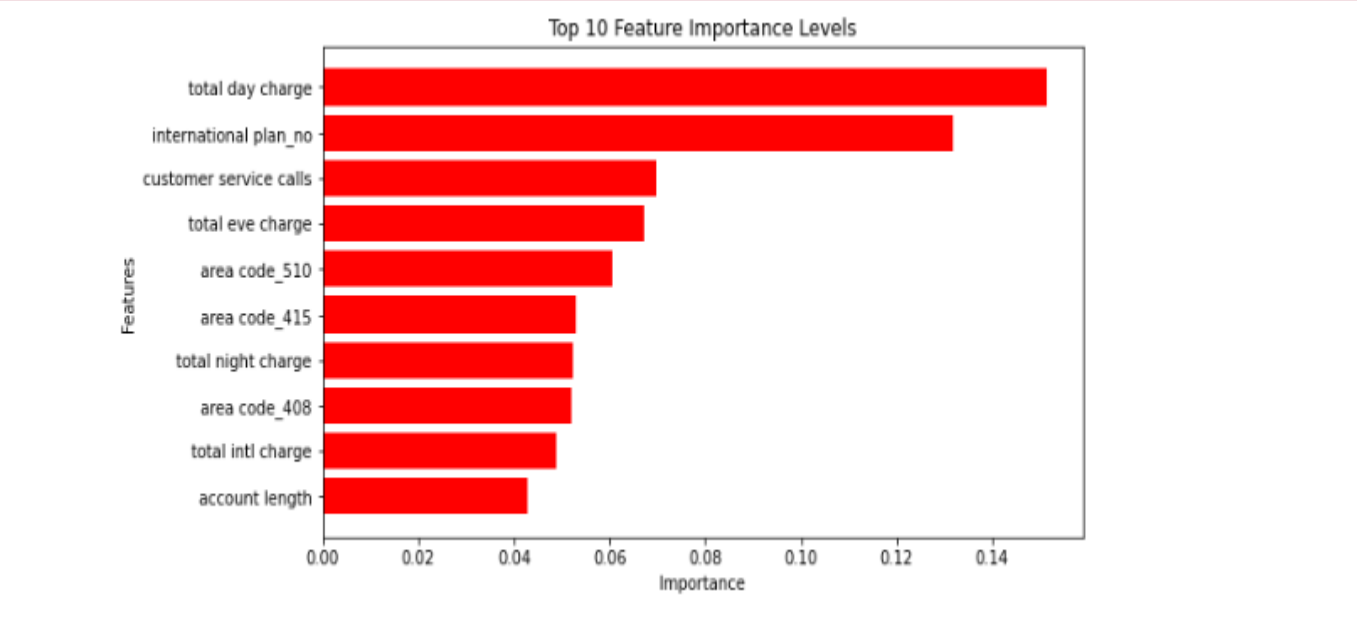
	Logistic Regression	Decision Tree Classifier	K-Nearest Neighbors	Random Forest
precision	0.37306	0.47682	0.21612	0.85047
Recall	0.576	0.576	0.472	0.728
F1-score	0.45283	0.52174	0.29648	0.78448
Accuracy	0.79137	0.84173	0.66427	0.94005



- First 10 important features for the Tuned Decision Tree Model



- First 10 important features for the Tuned KNN classifier

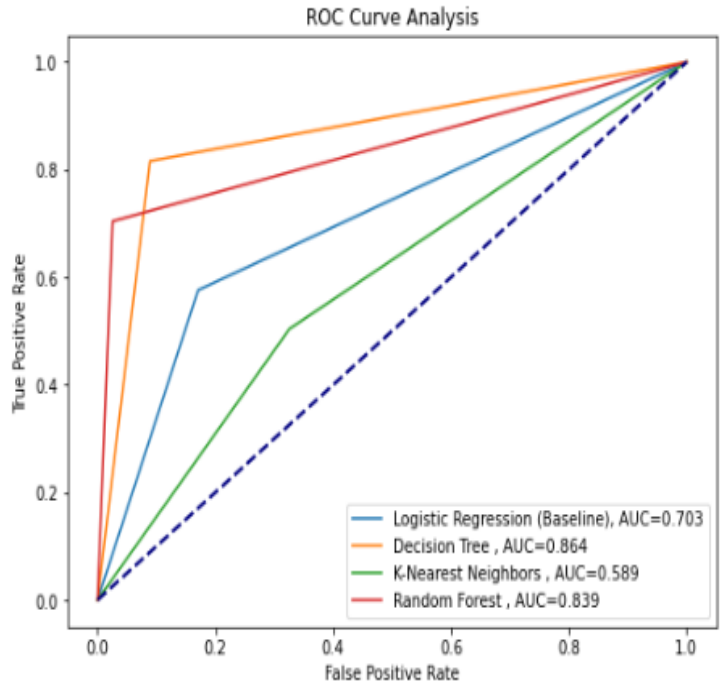


- First 10 important features for the Tuned Random forest classifier

MODEL EVALUATION AND COMPARISON

CURVE ANALYSIS USING ROC

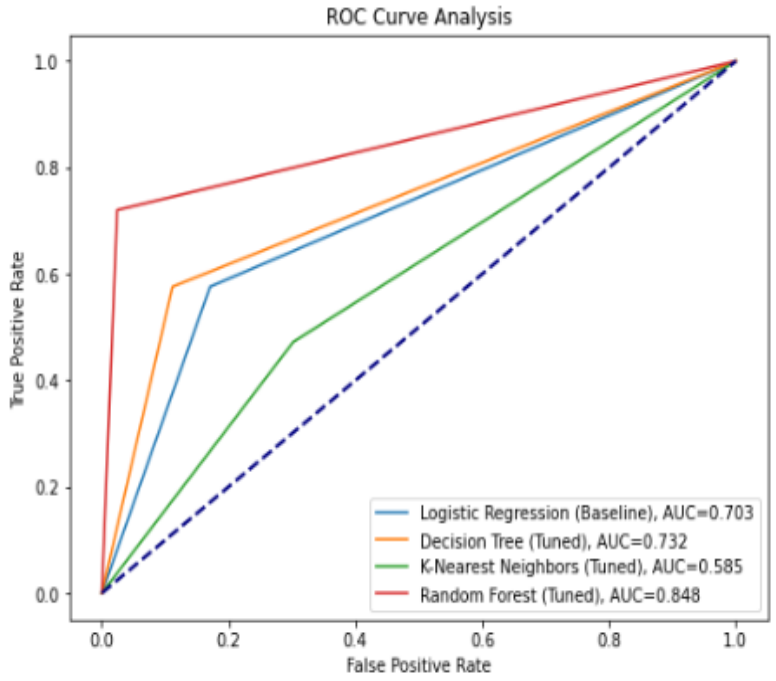
BEFORE MODEL TUNING



Models sorted by AUC in descending order:

Decision Tree
Random Forest
Logistic Regression (Baseline)
K-Nearest Neighbors

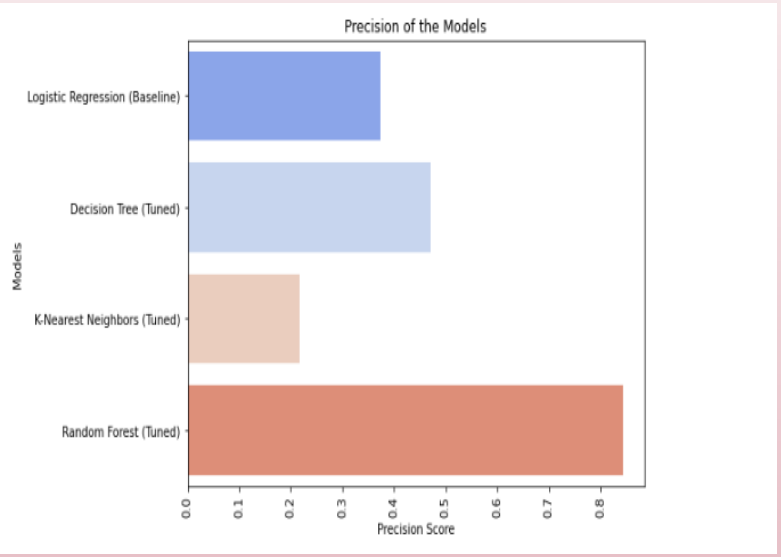
AFTER MODEL TUNING

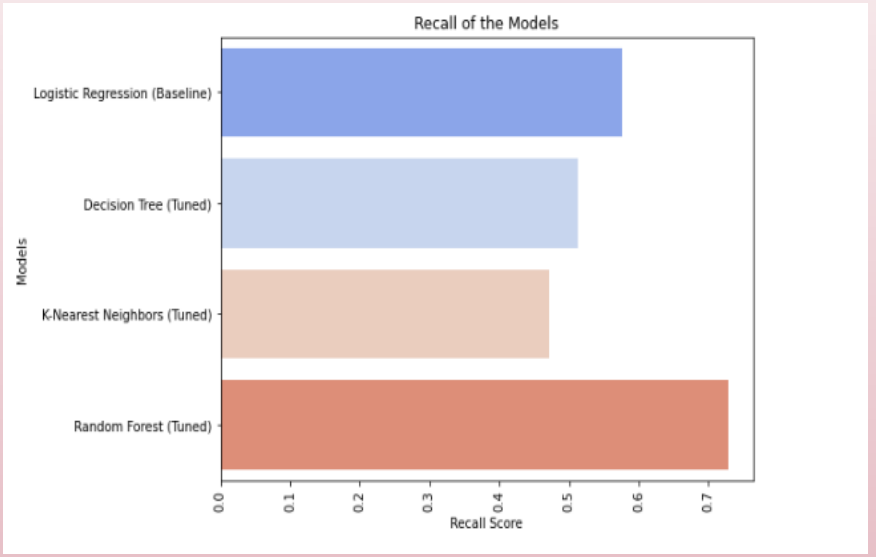


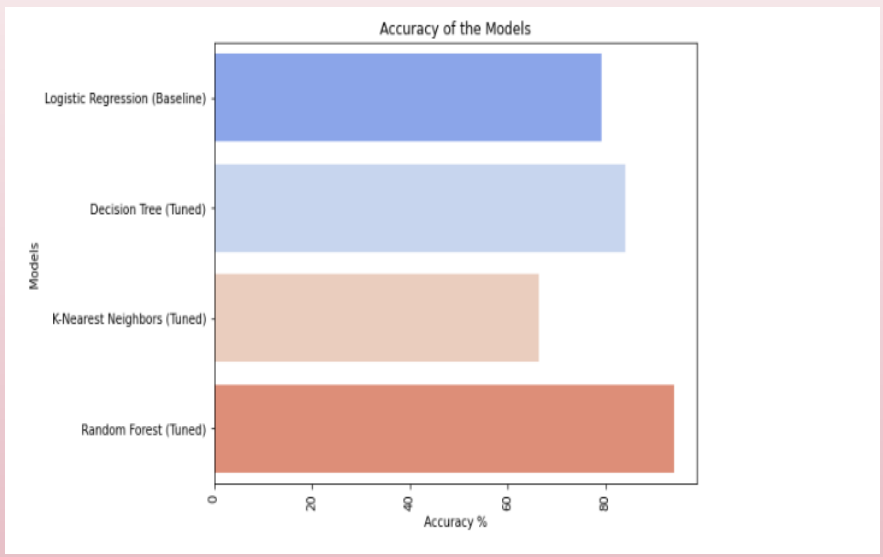
Models sorted by AUC in descending order:

Random Forest (Tuned)
Decision Tree (Tuned)
Logistic Regression (Baseline)
K-Nearest Neighbors (Tuned)

Model comparison using k-fold cross validation







- **Precision of the models**
- **Recall of the models**
- **Accuracy of the models**

SELECTING THE BEST MODEL:

- **The Random Forest model is the most suitable choice due to its strong performance in terms of accuracy, recall, and precision. It achieved an accuracy of 94% and a recall of 0.73, precision of 0.85 indicating its ability to accurately classify instances and achieve a balance between precision and recall.**

FINDINGS:

After performing sequential forward selection(SFS), the following features were selected:

number vmail messages, total day charge, total eve charge, total night calls, total night charge, total intl charge, customer service calls, voice mail plan_no, voice mail plan_yes, area code_408

RECOMMENDATIONS

- Optimize Daytime and Evening Charges
- Address Customer Service Calls
- Ensure Fair Nighttime Charges
- Enhance Daytime Customer Satisfaction
- Tailor Strategies to Each State
- Optimize Voicemail Plans
- Analyze Churn Patterns in Different Areas
- Adapt Marketing Strategies
- Monitor Customer Satisfaction
- Leverage Predictive Models

CONCLUSION

- **In the analysis I used different machine learning models to predict if customers might stop using SyriaTel services. Testing many models like Logistic Regression, Random Forest, Decision Tree and K-Nearest Neighbors and comparing their performance . I used measures like accuracy, F1 score, recall, and precision to check how good they are . Out of all the models tested, the Random Forest classifier (after adjusting it) did the best job at figuring out if a customer might stop using the services(churn).**

THANK YOU