SYRIA TEL CUSTOMER CHURN PROJECT

BY KAVATA MUSYOKA

OVERVIEW

- The telecommunication industry is one of the fastest growing industries as consumer communication needs increase and diversify and as new technologies emerge.

- The telecommunications landscape is therefore quickly evolving to meet these needs to and to integrate the new emerging technologies such as cloud computing, decentralized telecom networks, virtualized network services and artificial technologies amongst others.

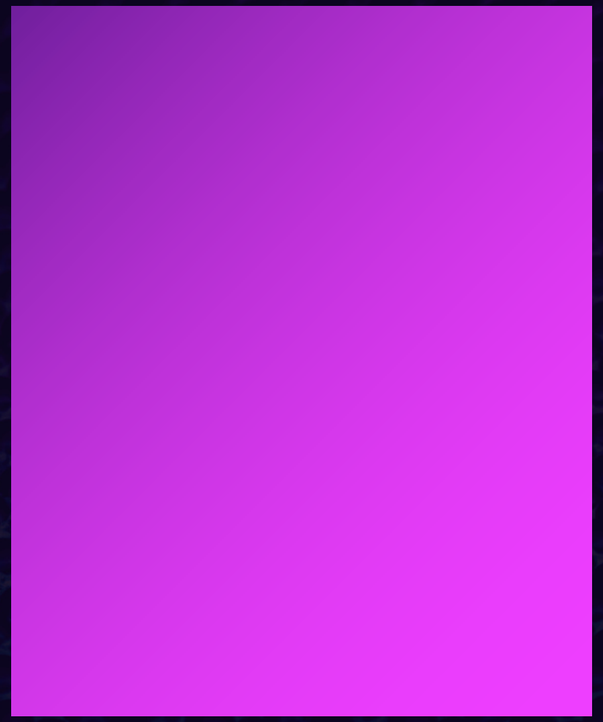
- I am undertaking a project to help SyriaTel to predict customer churn with a view to reduce losses that are associated with customer churn.

- Customer churn refers to the loss of customers or subscribers for various reasons.

PROBLEM STATEMENT

- Customer churning is one of the main killers of business growth.

- Therefore to reduce losses incurred from churning, identify customers who are at risk of churning and take proactive steps to retain them, stabilize their market value and optimize profits, there is need to predict customer churn at SyriaTel



OBJECTIVES

The main objective of this is to build a classification model that can predict whether or not a customer will churn.

To achieve the said main objective, the project will focus on the following specific objectives-

- Conduct exploratory data analysis of the dataset;
- Fit various classification algorithm models to determine the one that can provide the best churn predictions;
- Make predictions using the best prediction model; and
- Check the accuracy of the predicted variables

DATA

- The dataset contains 3333 rows and 21 columns. There are both numerical and categorical columns in the dataset.
- The target variable is “Churn” which indicates whether a customer has churned or not.
- From an analysis of the distribution of the numerical features, they all have normal distribution apart from customer service calls and the number of voicemail messages.

DATA PREPARATION

- Unnecessary columns were dropped in the course of the project namely, phone, area code and state.
- There were no missing values in the data nor were there any duplicates found.
- The class weights were adjusted to account for the imbalance in the target variable.

DATA EXPLORATION

- About 14.5% of the customers have churned in the dataset.
- Customers without an international plan with SyriaTel, end up churning.
- Majority of customers who churned did not have a voicemail plan.
- Texas and New Jersey had most of the customers who churned.

DATA MODELLING

The following classification models were built and used to evaluate the dataset:

- Logistic Regression;
- DecisionTree; and
- the RandomForestClassifier

LOGISTIC REGRESSION REPORT FOR TEST DATA

	precision	recall	f1-score	support
0	0.93	0.69	0.79	709
1	0.28	0.70	0.40	125
accuracy			0.69	834
macro avg	0.61	0.70	0.60	834
weighted avg	0.83	0.69	0.73	834

The logistic regression has a recall value of 0.70, which is a good baseline model. Meaning that the model can identify atleast 70% of the actual positive instances accurately.

DECISION TREE CLASSIFICATION REPORT FOR TEST DATA

	precision	recall	f1-score	support
0	0.96	0.84	0.89	709
1	0.46	0.78	0.58	125
accuracy			0.83	834
macro avg	0.71	0.81	0.74	834
weighted avg	0.88	0.83	0.85	834

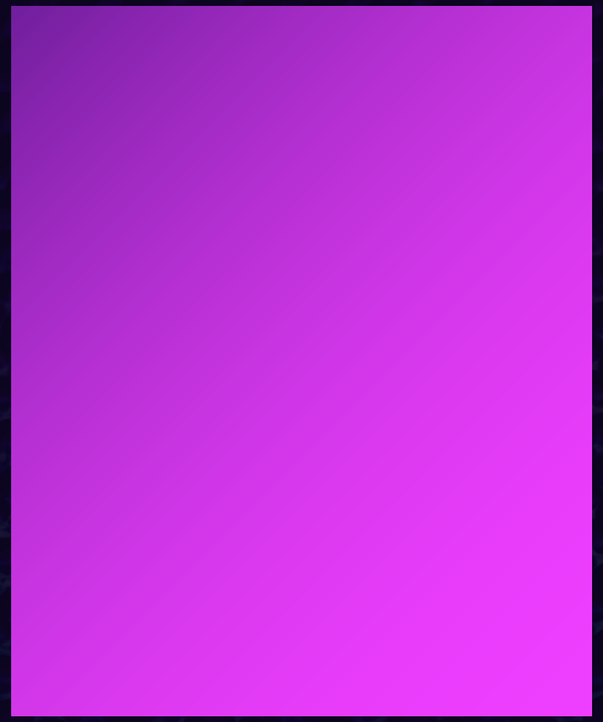
The decision tree model has a recall score of 0.78, meaning that it can identify 78% of the actual positives instances accurately. Further meaning that its making predictions accurately more than inaccurately.

RANDOM FOREST CLASSIFICATION REPORT FOR TEST DATA

	precision	recall	f1-score	support
0	0.96	0.95	0.96	709
1	0.73	0.78	0.76	125
accuracy			0.92	834
macro avg	0.85	0.87	0.86	834
weighted avg	0.93	0.92	0.93	834

The random forest classifier model has a recall score of 0.78 which is similar to the decision tree classifier model, meaning that both models can accurately identify the positive accurancy by 78%.

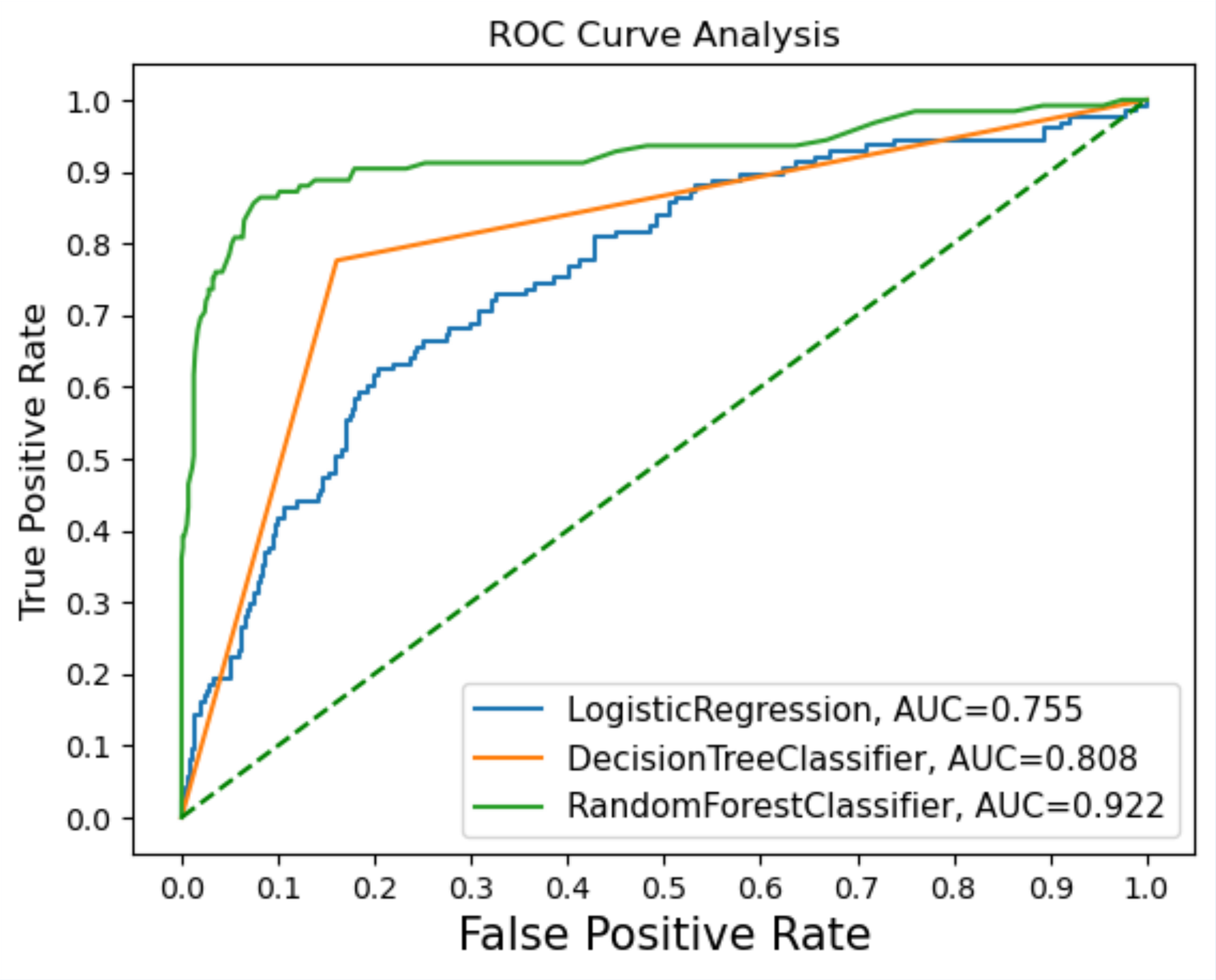
MODEL EVALUATION



The dataset contains 3333 rows and 20 columns. There are both numerical and categorical columns in the dataset. The target variable is “Churn” which indicates whether a customer has churned or not.

About 14.5% of the customers have churned in the dataset.

MODEL EVALUATION



The ROC curve analysis shows that the RandomForestClassifier has the best performance of 0.922 followed by the DecisionTreeClassifier with 0.808 and the last is the LogisticRegression with 0.755.

RECOMMENDATIONS

Based on the findings, I have the following recommendations to SyriaTel:

- The number of customer service calls was identified as one of the most influential predictors of churn hence the company should invest in excellent customer services and prompt redress of issues.
- The company should provide incentives for customers to have international plans.
- The company should also offer attractive voicemail plans to its customers as those who churned did not have a voice mail message plan.

NEXT STEPS & CONCLUSION

Exploring advanced techniques like ensemble methods, Xgboost and Adaboost should be explored to improve further the prediction performance.

Monitoring and evaluation should continue to capture any new data that might mitigate against customer churning.

The project built a classification model to predict customer churn for SyriaTel. Various models were built and the best one was the Random Forest classifier with a recall score of 78% and AUC level of 0.92 once tuned.

By taking the recommendations into consideration, the company can improve customer retention and reduce customer churn.

THANK YOU