SYRIATEL LTD.

DATA-DRIVEN CUSTOMER CHURN PREDICTION

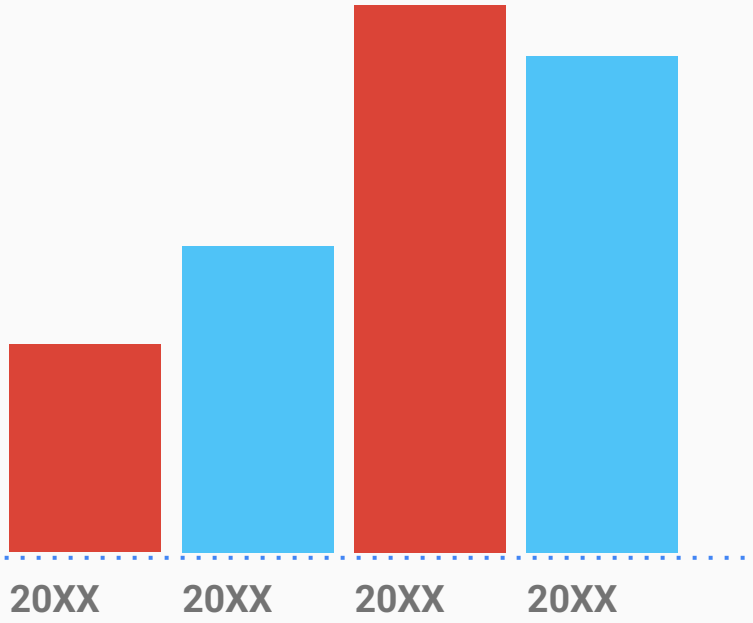
By: Nelson Muia

OVERVIEW:

- **To predict the rate and the drivers of customer churn using Machine Learning models**
- **The analysis will entail a binary classification problem: “Churn and No churn”**
- **Object is to identify subscribers who are most likely to leave and the cause variables**

BUSINESS AND DATA UNDERSTANDING

- High churn rates are a major risk in the telecom industry.
- This can affect a company's profitability and in turn it’s going concern
- Loss of client base by SyriaTec co. ltd can be explained using the Porter’s 5 sources



BUSINESS AND DATA UNDERSTANDING

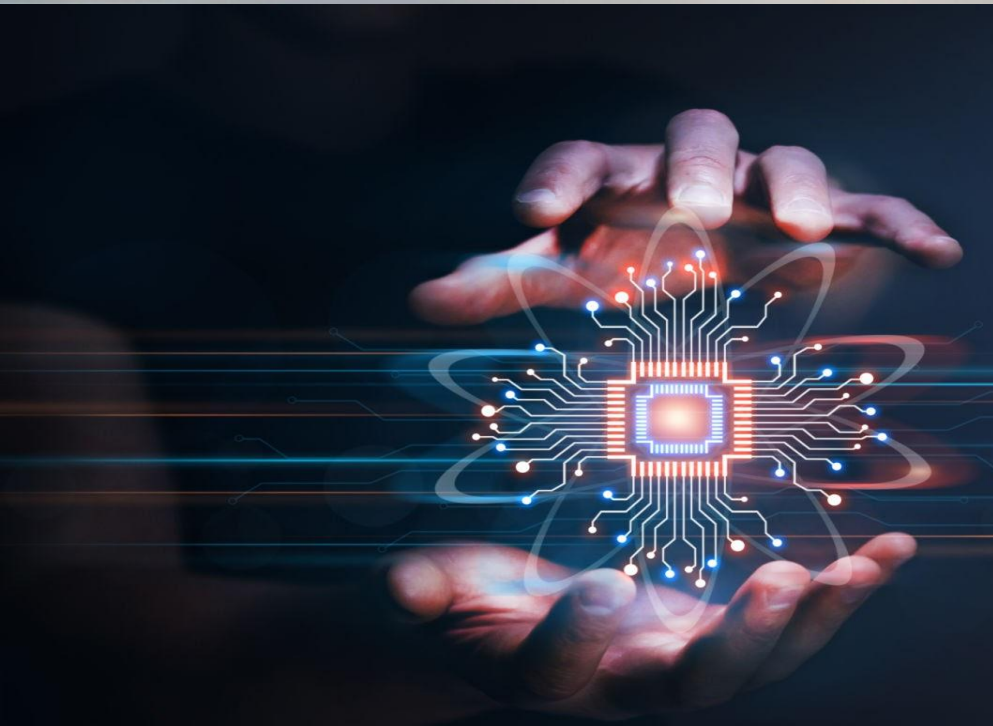
PORTER'S 5 FORCES:

1. Competitive rivalry
2. Threat of new entrants
3. Threat of substitutes
4. Buyer power
5. Supplier power

The solution:

To use Machine Learning models to predict customer behaviour and develop a market-based solutions and strategy

BUSINESS AND DATA UNDERSTANDING



TARGET STAKEHOLDERS

1. Customer Retention team
2. Marketing Department
3. Data analysts and Scientists
4. Senior Management

The Dataset

- **The dataset being used is sourced from Kaggle, specifically the “”SyriaTel Customer Churn” dataset. It can be found at the following link:**

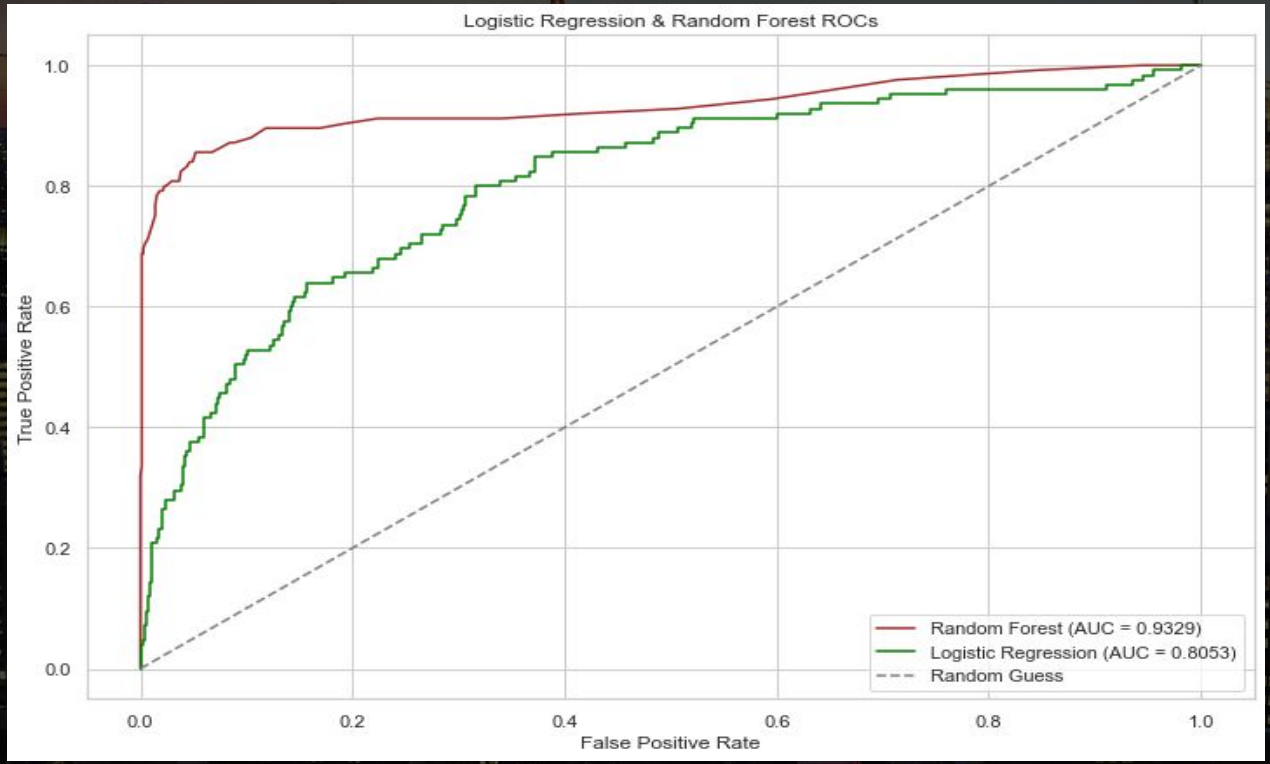
https://www.kaggle.com/datasets/becksddf/churn-in-telecoms-dataset/data

MODEL

The Machine Learning models are:

- Logistics regression - involves a linear model that is used for binary classification
- Random forest classifier - it’s an ensemble model that is used to build many decision trees and combine their outputs.

LOGISTIC REGRESSION & RANDOM FOREST R.O.C'S CURVES



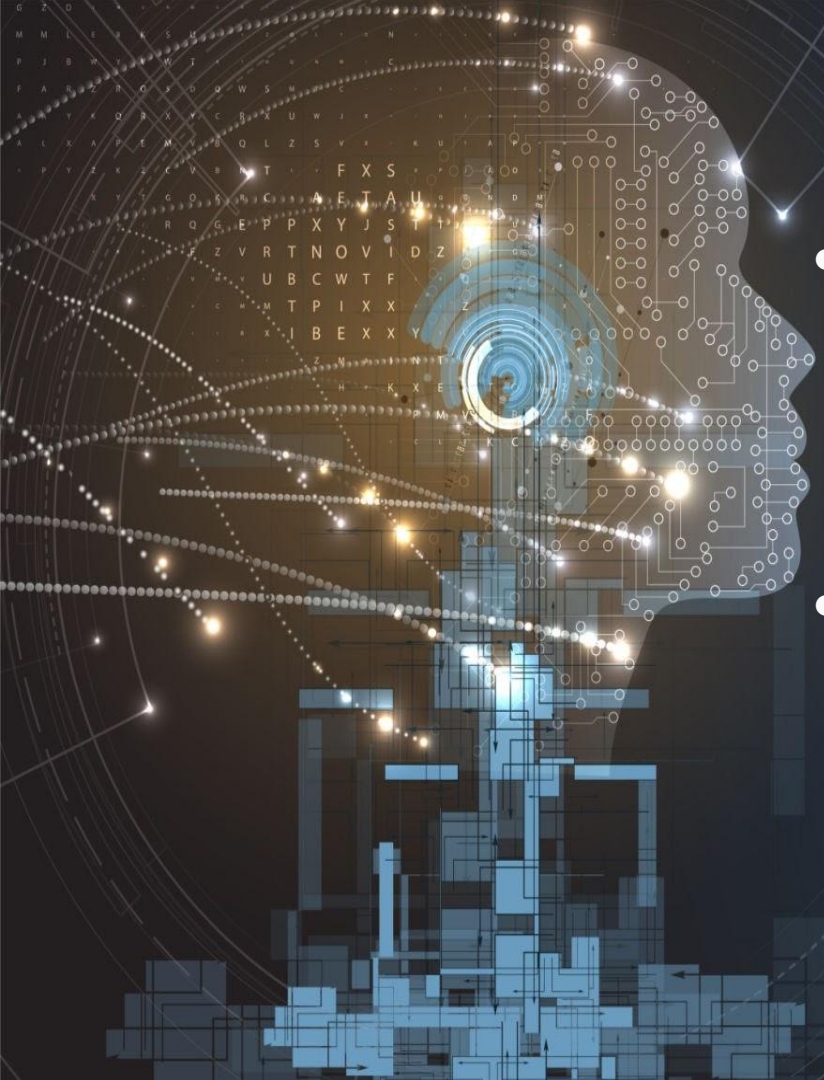
EVALUATION:

- The closer the ROC curve is to the top left, the better the model. Random forest's curve stays above Logistic regressions, indicating stronger performance
- The Random forest model achieves a better balance between recall and false positives, marking it more effective for churn prediction.
- The logistic Regression model performs reasonably but is less accurate at identifying churners compared to Random forest model.

CONCLUSION:

- Logistic regression - performs well in identifying customers who are unlikely to churn, but finds it difficult accurately detecting actual churners. While efficient and easy to interpret, its less effective for imbalance datasets.

- Random forests - they offer stronger and more consistent performance across both churn and non-churn classes. Although it does’nt catch every churn case, it’s significantly higher ROC-AUC score indicates a better class distinguishing ability.



RECOMMENDATIONS

- It is imperative for SyriaTel company to evaluate and improve their international plan offering, as these subscribers are the backbone of the company. They also have a high likelihood of churning.
- The models have also shown us that us a company in mitigate the risk of customer churn and to ensure profitability we should focus a lot on the daytime users, offer attractive prices and respond to customer service calls.