DATA REPORT

Business Understanding

Syriatel, like many other telecommunications companies, faces the challenge of retaining its customers and minimizing churn. By analyzing a dataset containing information on customer characteristics such as account length, international plan, voice mail plan, usage patterns, and number of customer service calls, Syriatel can gain insights into the relationships between these characteristics and churn. This information can be used to implement strategies to minimize churn.

By using statistical techniques such as logistic regression or decision trees, the company can also predict customer churn using the dataset. Accurately predicting churn can improve customer retention and increase customer satisfaction, ultimately leading to improved business performance. The dataset contains information on various aspects of customer behavior, including their state, area code, telephone number, number of voice mail messages, total daytime, evening, and night minutes, calls and charges, total international minutes, calls and charges, and the number of customer service calls made. By analyzing this data, Syriatel can identify the factors that contribute to churn and respond appropriately to retain its customers.

Research Question

The main objective of this project is to build a model that will predict customer churn for Syriatel Company to understand the factors that contribute to customer churn and implement strategies to minimize it. The goal is to improve customer retention and increase customer satisfaction, leading to improved business performance for the telecommunications company.

Objectives

- The primary objective of this project is to reduce customer churn, which is the loss of customers to competitors. By predicting which customers are at risk of leaving
- Identify which features/predictor variables affect the attrition of customers
- Build 3 Classification models and evaluate the best one for classifying and predicting the churn rate

Data Understanding

Data Source

The dataset used for this project was obtained from Kaggle (Churn in Telecom's dataset)

Data Description

Our data was in csv format and contained data grouped in columns of dependent and independent variables. The dataset has 3333 records with 21 columns. These columns include information about the customer's state, account length, area code, phone number, international plan, voice mail plan, and call usage details (minutes and charges) for different time periods (day, evening, night, international). The

last column, "churn", is the target variable indicating whether the customer has churned or not. All the features except the phone number and state have numerical values, with the rest being categorical or binary (international plan, voice mail plan, and churn). The data does not contain any missing values.

Data Preparation

Loading the Data

At the beginning of the process, the necessary libraries were imported and then the bigml_59c28831336c6604c800002a.csv dataset was loaded onto the jupyter notebook using pandas.

Reading and checking the Dataset

The data was read and then checked for anomalies, outliers, missing values and duplicates. This was to determine the next course of action that would ensure the data would be set for use. During this process, it was established that the dataset had no missing values.

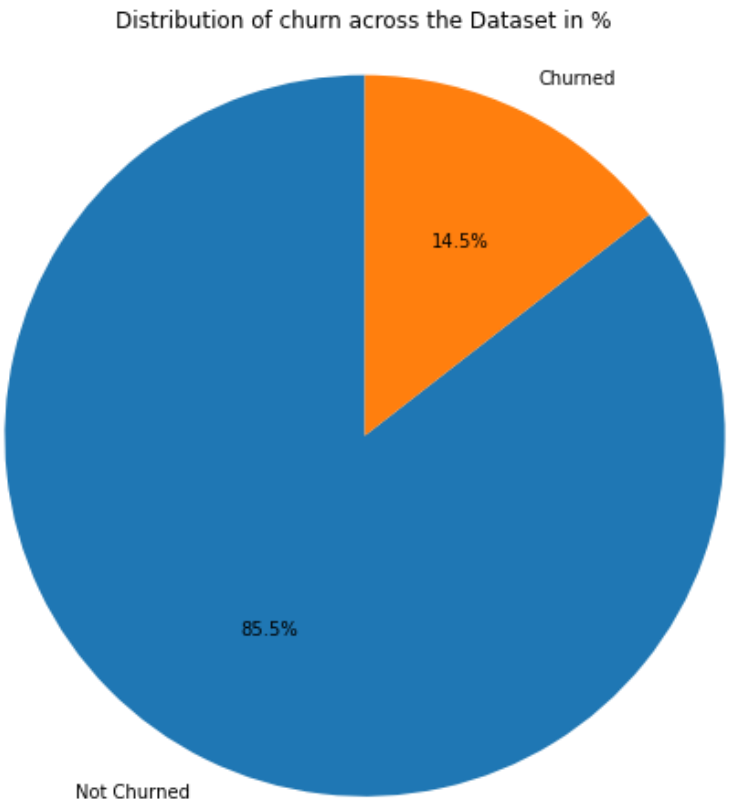
Cleaning the data

The outliers were retained in the dataset as they were considered critical for the model training process. Only after thorough model diagnostics and determination of their ineffectiveness towards the study were the outliers removed. If not, they were maintained in the data.

External Data Source Validation

The data set was measured against a reliable external data source (Analysis Mason, a consulting and Research Company) to ensure that it was in line with what it should be and checked for any additional issues with the data set. Analysis Mason published a report that the intention to churn among telecommunication subscribers in the sub-Saharan African region ranged from 9% to 16% across all operators surveyed. This figure is in line with churn rates in other regions, but lower compared to the neighboring Middle East and Africa (MENA) region, where the intention to churn within 6 months was recorded at an average of 22%.

The study also found that customer service had a significant impact on churn for subscribers in South Africa, where the average churn rate was 17%. This effect was much greater compared to the other countries covered, which may reflect the differences in market maturity in the region. These findings are not very far from that of the our Syriatel dataset as expressed in the chart below.



Exploratory Data Analysis

The data sets were analyzed and trends found by using statistics and visualizations to aid in comprehending the data set. There were several questions that were answered in this step by comparing the predictor variables with the target variable which was churn using data visualization tools. The questions answered and variable relationships established include:

- What are the top 10 States with the highest/lowest churn rate?
- What is the churn rate of customers at SyriaTel ?
- What is the distribution of churned and non-churned customers in the dataset?
- What is the distribution of the numerical columns in the telecom churn prediction dataset and how do they impact the likelihood of churn?

Modeling

Data modeling commenced by checking the correlation between the predictor and target variables. Since the dataset consisted of both categorical and numerical variables, label encoding was used to convert the data into a form that can be used by our various classification models i.e:

- Logistic Regression Classifier model, which is our vanilla/baseline model - works by finding the best line that, separates the two classes in a high-dimensional space.

- Adaboost model - idea is to give more weight to difficult-to-classify points and improve the overall performance of the model.
- Gradient Boosting model -known for its high accuracy and performance
- Random Forest model - tends to reduce overfitting, which is a common issue with decision trees.
- Decision Tree model
- K-Nearest Neighbor model
- Hyper parameter tuning of the Decision Tree model
- Hyper parameter tuning of the Random Forest model

In conclusion, the results of the mean random forest cross-validation score (k=5) on predicting the churn rate showed an accuracy of 0.95, with a weighted average of 0.95. The precision, recall, and f1-score for class 0 (not churned) were 0.97, 0.97, and 0.97, respectively. For class 1 (churned), the precision, recall, and f1-score were 0.82, 0.82, and 0.82, respectively. The macro average and weighted average for precision, recall, and f1-score were 0.89 and 0.95, respectively.

These results indicate that the random forest model performed well in terms of accuracy, with a high weighted average for precision, recall, and f1-score. The model's performance in predicting class 0 (not churned) was slightly better compared to class 1 (churned), with a higher precision, recall, and f1-score. Overall, the random forest model was considered the best model for predicting the churn rate.

Conclusion

Determining the likelihood of telecommunication customers churning can be a challenging task. This predictive model would aid telecommunication companies in accurately predicting churn rates based on various input factors, reducing the risk of losing valuable customers.