SYRIATEL TELECOM CUSTOMER CHURN PREDICTION

Predicting Likelihood for customer churn using classification Models

Overview.

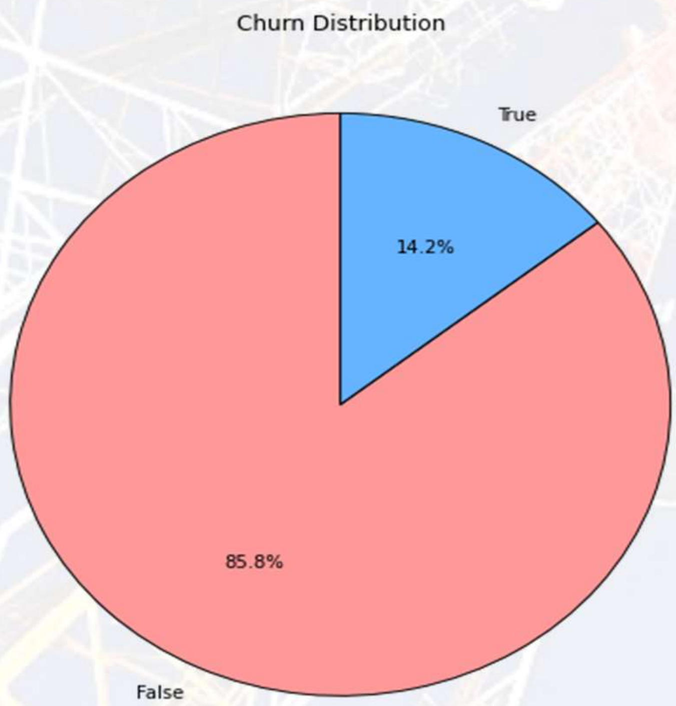
- The main objective for this presentation is to reduce the rate of customer churn in our company
- The goals are to:

1. Come up with a classification model to predict customer churn
2. Take actionable measures based on the predictions made.

Business and Data Understanding

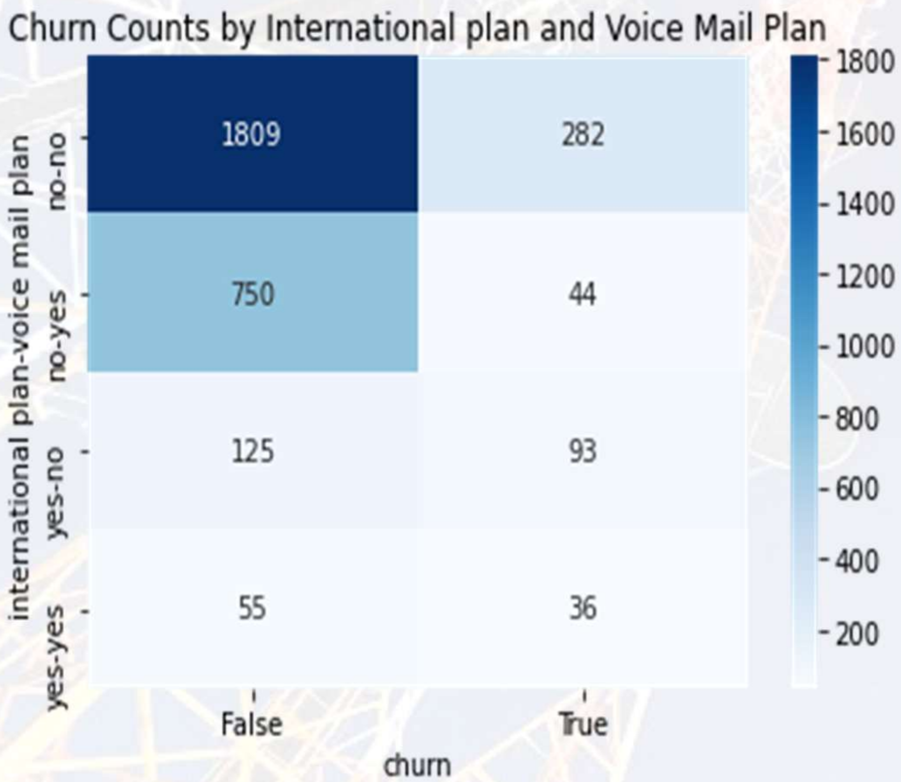
- Customer churn for syriatel is becoming a challenge as it is impacting the company’s revenue.
- The predictions aims to classify customers into Non-churning customers and churning customers.
- The data used in this prediction is the syriatel Customer Churn data sourced from kaggle.
- The 4 Most important features in the data for this classification problem are; Total night minute, Total night charge, Total evening minute and Total evening charge.

Target Variable Distribution



- The figure shows that 14.2% of the customers are likely to churn while 85.8% are likely to stay.
- This therefore shows there is a class imbalance which needs to be addressed.

Churn Counts by International plan and Voice mail plan



- Customers with no international plan and no voice mail plan are less likely to leave.
- Customers with no International plan and with a voice mail plan are more likely to leave the company.

Modeling

- Data preprocessing, here the data was prepared for modeling by using various preprocessing approaches:

1. OneHotEncoding categorical columns – these are columns that have Yes or No, 1 or 0 and true or False entries.
2. Scaling, using MinMaxScaler – This ensures the data is into a standard scale before fitting a model.

- Modeling Metrics – These are various types of models used for prediction.

1. Logistic Regression
2. Decision tree

- Hyper parameter Tuning – These are methods to improve Model performance. They include; SMOTE(Synthetic Minority Over-sampling Technique), maximum tree depth, minimum sample split, maximum feature and minimum sample leafs.

Model Evaluation and Interpretation

1. Logistic Regression with No SMOTE.

- Precision = 0.85: Out of all instances that were predicted as class 0, 85% were actually class 0.
- -Recall = 0.99: Out of all actual class 0 instances, 99% were correctly identified by the model.
- -F1-score = 0.91: The harmonic mean of precision and recall. It's a balanced measure that considers both false positives and false negatives.
- -Precision = 0.53: Out of all instances predicted as class 1, 53% were actually class 1.
- -Recall = 0.09: Out of all actual class 1 instances, only 9% were correctly identified by the model.
- -F1-score = 0.15: This is quite low, indicating that the model has trouble balancing precision and recall for class 1. -Accuracy = 0.84

Model Evaluation and Interpretation

2. Logistic Regression with SMOTE.

- Precision = 0.94: Out of all instances that were predicted as class 0, 94% were actually class 0.
- -Recall = 0.79: Out of all actual class 0 instances, 79% were correctly identified by the model.
- -F1-score = 0.86: The harmonic mean of precision and recall. It's a balanced measure that considers both false positives and false negatives.
- -Precision = 0.39: Out of all instances predicted as class 1, 39% were actually class 1.
- -Recall = 0.72: Out of all actual class 1 instances, only 72% were correctly identified by the model.
- -F1-score = 0.51: This has improved from 0.15, indicating that we have sorted out the model's trouble in balancing precision and recall for class 1.
- -Accuracy = 0.78: The overall accuracy is 78%, meaning the model correctly predicted the class in 78% of all instances.

Model Evaluation and Interpretation

3. Decision Tree Model:

- Precision = 0.94: Out of all instances that were predicted as class 0, 94% were actually class 0.
- -Recall = 0.94: Out of all actual class 0 instances, 94% were correctly identified by the model.
- -F1-score = 0.94: The harmonic mean of precision and recall. It's a balanced measure that considers both false positives and false negatives.
- -Precision = 0.67: Out of all instances predicted as class 1, 55% were actually class 1.
- -Recall = 0.67: Out of all actual class 1 instances, only 23% were correctly identified by the model.
- -F1-score = 0.67: This is quite low, indicating that the model has trouble balancing precision and recall for class 1.
- -Accuracy = 0.89: The overall accuracy is 89%, meaning the model correctly predicted the class in 89% of all instances.

Recommendations

- Logistic Regression before applying SMOTE is good at identifying customers that are unlikely to churn (class 0) but less effective for churners(class 1)
- Logistic Regresssion after applying SMOTE is a better model for predicting Churners(class 1). This is because recall for class 1 has improved, This is the best for identifying at-risk customers
- Decision tree model has difficulty predicting churners although it provides high accuracy, it performs poorly in identifying churners
- ***The Logistic regression (After Smote)*** is the most effective model for identifying high-risk churners.

Next Steps

- Monitoring Model Performance
- Regular update for model features and retraining models based on how customers will be behaving as time goes by.

THANK YOU!