SyriaTel Customer Churn Analysis and Modeling

Business Understanding

Across the telecommunication industry, customer churn is one of the most important concerns that directly affect a telecommunication company's business. Hence, companies that are able to successfully predict customer churn will be able to allocate capital and resources more efficiently and thereby improve profitability.

Research Question

This project analyzes customer churn (customers leaving the provider) data from the telecommunications provider SyriaTel. The main objective of this project is to identify what type of customers were churning and develop a model that could predict whether a customer is likely to churn.

.

Objectives Main Objective

To build a model that could predict whether a customer is likely to churn.

Metric of Success

The model will be considered a success when it achieves high accuracy, F1 score, recall, and precision scores

Data Understanding Data Source

Data is downloaded from Kaggle website

Data Description

Data includes:

- The Target:churn

Features/Predictors:

- account length: the number of days the user maintains a phone plan account
- international plan: "yes" if the user has the international plan, otherwise "no"
- voice mail plan: "yes" if the user has the voice mail plan, otherwise "no"
- number of voice mail messages: the number of voice mail messages the user has sent
- total day minutes used: total number of minutes the user has been in calls during the day
- day calls made: total number of calls the user has completed during the day
- total day charge: total amount of money the user was charged by the Telecom company for calls made during the day
- total evening minutes: total number of minutes the user has been in calls during the evening
- total evening calls: total number of calls the user has completed during the evening
- total evening charge: total amount of money the user was charged by the Telecom company for calls made during the evening

- total night minutes: total number of minutes the user has been in calls during the night
- total night calls: total number of calls the user has completed during the night
- total night charge: total amount of money the user was charged by the Telecom company for calls made during the night
- total international minutes used: total number of minutes the user has been in international calls
- total international calls made: total number of international calls the user has made
- total international charge: total amount of money the user was charged by the Telecom company for international calls made
- number customer service calls made: number of customer service calls the user has made

Data Preparation Loading Libraries

Load the libraries necessary for cleaning and analysis

Loading the Data

Load the dataset from the CSV file. The name of the CSV file is “syria_tel_data.csv”.

The shape of the dataset is 3333 by 22 (3333 rows and 21 columns)

Cleaning the Data

Data cleaning is the process of preparing data for analysis by weeding out information that is irrelevant or incorrect. This type of information typically reinforces a false belief, which might have a negative effect on the model or algorithm it is given into.

The steps for the data cleaning process is:

1. Consistency - Ensure there are no duplicates in the data.

2. Uniformity - Ensure the data types for the datasets are accurate.

3. Completeness - Ensure the dataset has no missing values.

4. Dividing the dataframe to categorical and continuous variables

Preprocessing the dataset

1. Converting categorical variables into numerical variables
2. Scaling numerical features
3. Applying SMOTE Technique to Resolve Unbalanced 'churn' Feature

Modelling

Before doing anything else, I conducted a train/test split on the data in order to prevent leakage.

Model 1 - Logistic Regression Classifier

- Logistic regression is a classification algorithm, used when the value of the target variable is categorical in nature.
- It is most commonly used when the data in question has binary output, so when it belongs to one class or another, or is either a 0 or 1.

According to the logistic regression classifier model, total day minutes , total evening minutes and international plan are the top three important features.

Model accuracy is almost 78%, which isn't bad. F1 score is only 49% which means the test will only be accurate half the times it is ran.

Model 2 - Decision Tree Classifier

- Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches

represent the decision rules and each leaf node represents the outcome.
- It is called a decision tree because, similar to a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure.

According to the decision tree classifier, International plan, total International minutes and number main messages are the three most important for the model.

The model accuracy is almost 93% and F1 score is 78%, recall and precition scores are much better than model 1.

Model 3 - Random Forest Classifier

- A forest is comprised of trees. It is said that the more trees it has, the more robust a forest is. Random forests creates decision trees on randomly selected data samples, gets prediction from each tree and selects the best solution by means of voting. It also provides a pretty good indicator of the feature importance.
- In machine learning, hyperparameter optimization or tuning is the problem of choosing a set of optimal hyperparameters for a learning algorithm. A hyperparameter is a parameter whose value is used to control the learning process. By contrast, the values of other parameters (typically node weights) are learned.

According to the Random Forest Classifier, Total Charge, Customer Servise Calls and International Plan have the highest impact on the model.

The model accuracy is almost 98% and F1 score is 91%, recall and precition scores are much better than model 2.

Model 4 -Hyperparameter Tuning of Random Forest Classifier

Cross validation GridSearchCV hyperparameter tuning technique is used.

Models Comparison

The ROC curve plot showed the best performing models which have a curve that hugs the upper left of the graph, which is the tuned random forest classifier in this case.

Modeling Summary

- The accuracy score of 0.979 means that the model correctly predicted the outcome for 97.9% of the observations in the test set.
- The F1 score of 0.92308 is the harmonic mean of precision and recall and is used to balance the precision and recall. The higher the F1 score, the better the balance between precision and recall. In this case, the F1 score of 0.92308 is a good indication that the model has a good balance between precision and recall.
- The recall score of 0.88112 means that the model correctly predicted 88.112% of the positive outcomes in the test set. This score is important when the positive outcome is of great interest and we want to minimize false negatives.
- The precision score of 0.96923 means that when the model predicts positive, it is correct 96.923% of the time. This score is important when false positives are costly and we want to minimize them.

Recommendations

SyriaTel should prioritize implementing strategies aimed at reducing customer churn, as a loss of 14.8% of their customer base has already been experienced. To achieve this goal, the following initiatives should be considered:

- Rate Assessment

It is essential to evaluate the current charging rates and identify any potential areas for improvement, as customers seem to be dissatisfied with high charges, resulting in increased likelihood of churn.

- Customer Service Evaluation

To improve customer experience, it is crucial to assess the quality of customer service offered by the company. With over 50% of customers making more than three service calls churning, additional training for customer service staff and the creation of an internal forum to document common customer issues should be considered.

- International Plan Analysis

To stay ahead of the competition, it is crucial to investigate the viability of offering an international plan to improve retention and customer satisfaction. A thorough market analysis should be conducted to assess the competitiveness of pricing against other providers, with relevant improvements made to the international plan based on the results.

- Voicemail Plan Promotion

The low number of customers subscribed to the voicemail plan suggests that some customers may be unaware of the option, and therefore promoting this plan may help to reduce churn.