Prediction Model for SyriaTel customer churn

Business Understanding

Customer churn is one of the most important metrics for a growing business to evaluate as it is much less expensive to retain existing customers than it is to acquire new customers. Customers in the telecom industry can choose from a variety of service providers and actively switch from one to the next.The technical progress and the increasing number of operators has raised the level of competition. Companies are working hard to survive in this competitive market depending on multiple strategies. This becomes a problem, as Telecom companies usually incur huge costs to attract subscribers.Since it is costly to lose customers, the goal is to use this data to build a classifier that can predict which customers will stop dealing with them for another provider, and identify how the company can avoid the loss of those customers.

Research Question

Due to the direct effect on the revenues of the companies, especially in the telecom field, SyriaTel is seeking to develop means to predict potential customers to churn. Therefore, finding factors that increase customer churn is important to take necessary actions to reduce this churn.

Objectives

To build a ML model that predicts the customers that are most likely to churn with an acceptably high accuracy.

-To compare different ML models predictions to achieve highest accuracy. .

-Identify customers that are likely to churn.

-Advice the Company on the best strategy.

Data Understanding

Data Source

We are going to perform Exploratory Data Analysis as part of the steps towards building and deploying a churn prediction model, based on the dataset from Kaggle which is an online community of data scientists and machine learning practitioners.

Data Description

Our data was in csv format and contained data grouped in columns of dependent and independent variables. The SyriaTel.csv dataset contained 21 columns and 3333 observations (rows). The 21 columns contained 16 numerical features , 2 categorical features ,1 bool(churn) and 3 object features.

Data Preparation

Loading the data

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

Reading and checking the data

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 our dataset did not have duplicates nor missing values.However it had outliers.

Cleaning the data

Our dataset was pretty much clean .The outliers were kept because they were essential for training the model. Outliers are only removed after performing model diagnostics and it has been established that they are irrelevant for the study, otherwise they are kept.

External data source validation

The data set was measured against a reliable external data source (report) to ensure that it was in line with what it should be and checked for any additional issues with the data set.

Database Marketing Institute published a report that established Wireless companies today measure voluntary churn by a monthly figure, such as 1.9 percent or 2.1 percent. This is the average number of customers who quit their service per month. Annual churn rates for telecommunications companies average between 10 percent and 67.

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(customers that tend to leave or stay) using data visualization tools. The questions answered and variable relationships established include:

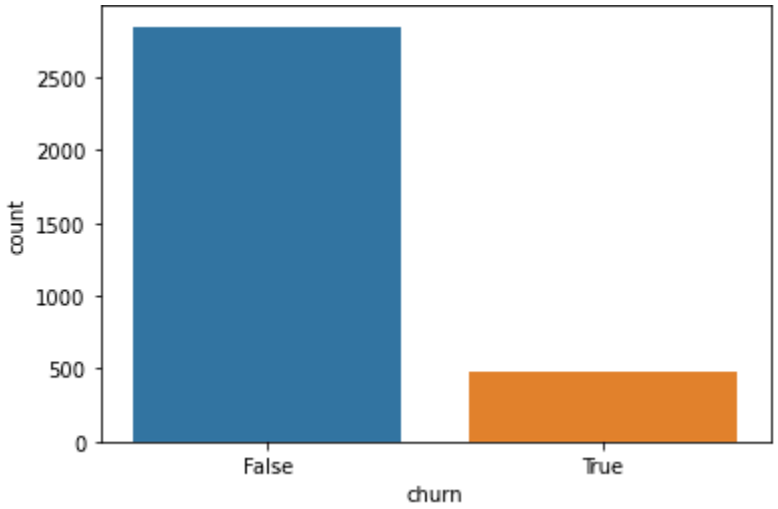
1.What is the total percentage of churn?
2..How much a client is charged for all phone call categories (international, evening, night, and day)?
3.What are the charges for day calls?
4.What is the relationship between customer service calls and customer churn?
5..Find out what attributes have the highest correlation with churn
6.Which customers are likely to churn

Modeling

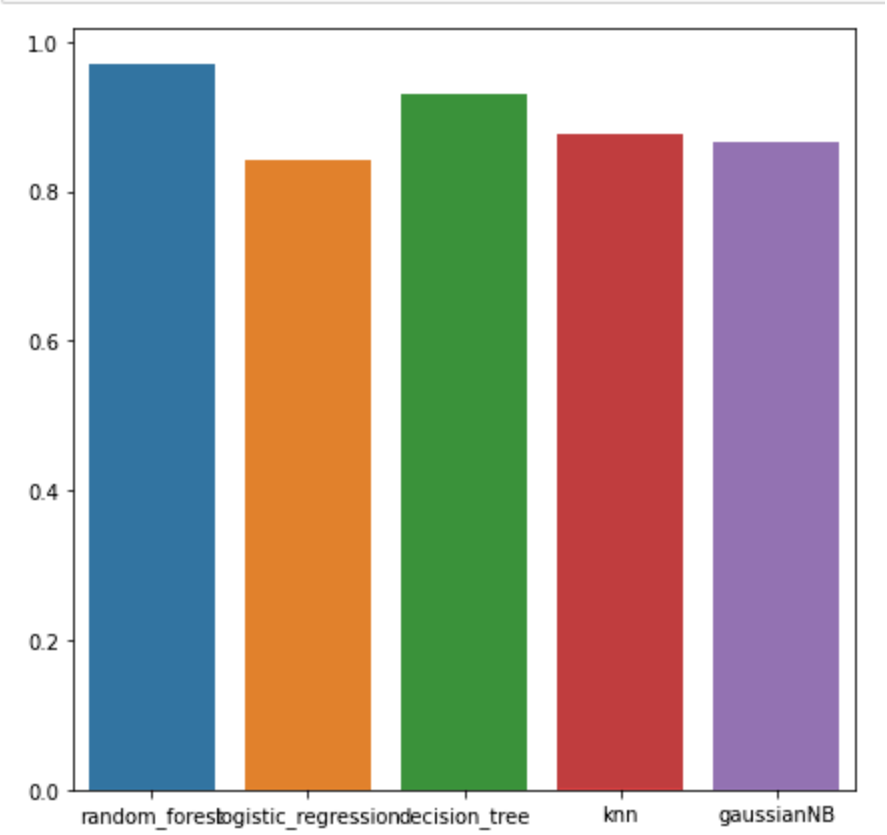
Data modeling commenced by checking the correlation between the predictor and target variables. This resulted in weak correlations for all predictor variables and a further analysis needed to be and was conducted.

Since the data set consisted of both categorical or numerical variables, label encoder was used to convert the data into a form that can be used .The data was on different scales which would have hindered the effectiveness of models.MinMaxScaler() was used to correct this.

Our target variable Churn was not normally distributed.his could have led to the minority class, churned clients, not being properly represented in the model. The problem was addressed using SMOTE to ensure that the minority class is well accounted for.



Without performing hyperparameter tuning,built and evaluated several classification models then picked the top three for further tuning.



I then employed hyper parameter tuning to improve model performance.
Random Forest model performed the best andi went ahead to instantiate a final model with the best parameters.

Conclusion

Customers who called customer service more than 3 times tend to leave.

-CustomerS who had international plan churn at a higher rate than the customers who do not have

Recommendation

Customer service should follow-up with the customers who call 3 times and offer promotions or discounts like a free month.

Syriatel should revisit its international plan and adjust the pricing.

Syriatel should offer a free voicemail plan for everyone.