Proposal Mohamed El-Deeb October 23, 2019

Domain Background

With the wide spread of smart phones telco companies are now a more important than ever and have a high number of customer but as the telco market matures and big companies compete in prices, service and quality of service every little advantage is needed for any company and since customers are the most important aspect that drives these companies it is highly important to retain customers in the company hence a churn model to help telco companies predict which customers are in danger of leaving the company is of extreme importance now more than ever.

The primary motivation behind this model is the dire need of businesses to retain existing customers, coupled with the high cost associated with acquiring new ones. A review of the field has revealed a lack of efficient, rule-based Customer Churn Prediction (CCP) approaches in the telecommunication sector in my country.

I am personally interested in this problem as I wish to my join one of the biggest telco companies in my country and since I lack a portfolio to show case my approach for telco related problems this project gives me a perfect opportunity to begin building my portfolio.

Problem Statement

Given a dataset of customer features, an algorithm needs to be developed to classify whether a customer will churn or not and try to determine what are the most contributing features in their churn while also investigating the reason behind the churn of said customers and if possible introduce a solution for different types of churn for the business.

Datasets and Inputs

I will be using the Telco Customer Churn "Focused customer retention programs" Dataset [See here] from Kaggle. This was uploaded for examining customer retention and predicting churn and will be well suited to this study. The data contains more than 7043 row each row represents a customer and 21 features including the target label as a flag.

They contain both numeric and categorical features that represents different customer's attributes that may contribute to customer churn they are as follows:

- customerID - A unique identifier for each cutomer
- gender - Whether the customer is a male or a female
- SeniorCitizen - Whether the customer is a senior citizen or not (1, 0)
- Partner - Whether the customer has a partner or not (Yes, No)
- Dependents - Whether the customer has dependents or not (Yes, No)
- tenureNumber of months the customer has stayed with the company
- PhoneService - Whether the customer has a phone service or not (Yes, No)
- MultipleLines - Whether the customer has multiple lines or not (Yes, No, No phone service)
- InternetServiceCustomer's internet service provider (DSL, Fiber optic, No)
- OnlineSecurity - Whether the customer has online security or not (Yes, No, No internet service)
- OnlineBackup - Whether the customer has online backup or not (Yes, No, No internet service)
- DeviceProtection - Whether the customer has device protection or not (Yes, No, No internet service)
- TechSupport - Whether the customer has tech support or not (Yes, No, No internet service)
- StreamingTV - Whether the customer has streaming TV or not (Yes, No, No internet service)
- StreamingMovies - Whether the customer has streaming movies or not (Yes, No, No internet service)
- ContractThe contract term of the customer (Month-to-month, One year, Two year)
- PaperlessBilling - Whether the customer has paperless billing or not (Yes, No)
- PaymentMethod - The customer's payment method (Electronic check, Mailed check, Bank transfer (automatic), Credit card (automatic))
- MonthlyCharges - The amount charged to the customer monthly
- TotalCharges - The total amount charged to the customer
- Churn - Whether the customer churned or not (Yes or No)

An imbalance can be observed in the target label as the number of lost customers are 1890 rows while the number of non-churn customers are 5174 rows.

Solution Statement

The proposed solution is to create classification model to classify customers whether they are going to churn or not and try to profile the churn types. I am going try different models using methods to tune these models and try to address the imbalance problem in the data using techniques such as SMOTE.

Benchmark Model

Looking on some kernels on kaggle it can be observed that accuracy ranges from 84% to 73% with kernels using relatively simple models as logistic regression (See here) an accuracy of 75% and sensitivity of 75%, other tree based models were used and they perform rather well.

Evaluation Metrics

This study will be evaluated with regards to the model ability to accurately predict the customers that are in danger of churning I will use both F-Score and sensitivity to evaluate my model since the business is mainly concerned with lost customer since the lost customers present a threat to revenue and the cost of regaining a customer is always greater that the cost of keeping one.

Project Design

1. Exploratory Data Analysis
    a. Data Overview
    b. Data Profiling
    c. Correlation Matrix
2. Data preprocessing
    a. Data cleansing if needed
    b. Outlier Detection
    c. Data Scaling
    d. Clustering and profiling customers (.e.g. GaussienMixture and Kmeans)
    e. Investigation customer behavior
    f. Feature importance
3. Modeling
    a. Baseline model
    b. Model selection and model tuning (.e.g. random forest, SVM, XGboost)
    c. Model evaluation