Machine Learning Engineer Nanodegree

Capstone Project

Noreen Abd ALLAH
October 9, 2018

I. Definition

Project Overview

Telecom industry is a highly competitive industry in today's world, there can be several service providers or companies that are highly competing to keep their customers and prevent them from churning to another company. Every Telecom company now have several varying services that are slightly different from other company's services making it easy for customers to go from one company to another.

with growing pressure from competition, it's becoming extremely important to predict customer behavior to improve customer's retention programs by proactively identifying customers that are about to churn in order to take preventive measures to retain these customers by making targeted retention strategies and offers.

churn rate is the measure of number of subscribers who leave a service provider during a given time period. In order for the Telecom company to maintain and increase it's growth rate; the number of it's new customers must exceed it's churn rate.

I have found the dataset on kaggle datasets and it was initially obtained from IBM sample datasets. The dataset contains relevant customer data to predict their behavior and retain customers that have a tendency to churn from the company.

Problem Statement

The growth and revenue of Telecom companies are hugely affected by it's customers churn rate, when a customer leaves a company it causes a huge loss for that company as the costs for initially acquiring that customer to subscribe for the companies' services may not have been recovered yet.

My solution for this problem is to build a machine learning model that can predict what class does the customer belong to; whether that customer will churn or not, given his/her relevant information.

Datasets and Inputs

I have found the dataset on kaggle datasets and it was initially obtained from IBM sample datasets. The dataset contains relevant customer data to predict their behavior and retain customers that have a tendency to churn from the company.

The raw dataset contains 7043 rows (customers) and 21 columns (features).

The "Churn" column is our target.

The data set includes information about:

- Customers who left within the last month – the column is called Churn .
- Services that each customer has signed up for – phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies.
- Customer account information – how long they've been a customer, contract, payment method, paperless billing, monthly charges, and total charges.
- Demographic info about customers – gender, age range, and if they have partners and dependents.

The independent features of the dataset will be used to predict the dependent target variable (Churn). The output of the model will be whether a particular customer will churn or not.

The dataset will be split into training set that the model will learn from and test set to test the model performance and check whether it generalizes will or overfit.

The training set will be further split into folds for validation during hyper parameter tuning.

Columns:

- **customerID:** Customer ID
- **gender:** Customer gender (female, male)
- **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)

- **tenure:** Number 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)
- **InternetService:** Customer'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)
- **Contract:** The 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)

Solution Statement

My proposed solution for this problem will be to build a machine learning model that can more accurately predict customer churn than other non machine learning strategies.

It will predict what class does the customer belong to, whther that customer will churn or not, given his/her relevant information.

Benchmark Model

My Benchmark model will be a plain vanilla Logistic regression model which can be used as a binary classification model and it is the most commonly reported data science method used at work for all industries according to kaggle's The State of Data Science & Machine Learning survey so I thought it would be interesting to outperform it and beat it's score.

Evaluation Metrics

I will choose Type I error as my main evaluation metric that I will choose the right model based on it. Type I error (False Negatives) is failing to identify a customer who has a high tendency to churn so, business wise this is the least desirable error as failing to identify the customers who are going to churn will highly affect the revenue of the company. The model the I will choose will be the one with the least type I Error.

Other metrics that I will use will be recall, F score (sklearn report), accuracy and ROC/AUC curve for the final model.

Project Design

1. Exploratory Data Exploration (EDA)

   First I will start by some data exploration and statistical measures to gain insight of what the data is all about and the structure of the dataset, find if there are missing values, explore the data types of columns and so on.

2. Visual Data Exploration

   Second I will analyze the dataset by some visual data exploration in order to gain valuable insights and a deeper understanding of the data that I'm working on.

3. Preprocessing

   In this step I will do some feature engineering for the features that needs some preprocessing like imputing missing values, scaling, labeling and encoding, adding other features, maybe reducing the dimensionality of the dataset.

4. Splitting Data

   Splitting the dataset into training and testing data.

5. Building Benchmark Model

   Building the Benchmark model which is a logistic regression model and evaluating it.

6. Building Models

   In this step I will Build the models that which will be aplied to this problem.

   The models that I will test would be:

   (a) Random Forest
   (b) AdaBoost
   (c) XGBoost

   These models will be tested without any parameter tuning.

7. Evaluation and Benchmarking

   (a) Evaluating the proposed models.
   (b) choosing the best one.
   (c) comparing it to the Benchmark model.

8. Validation

   Hyper parameter tuning and validation for the selected model using Grid Search Cross Validation and then Evaluate the final results.

References

[1]IBM Guide to Sample Data Sets

[2]Kaggle Telco Customer Churn dataset

[3]Efficient ways for Customer Churn Analysis in Telecom Sector

[4]Building Predictive Models for Customer Churn in Telecom

[5]Churn rate

[6]Kaggle The State of Data Science and Machine Learning

[7]Proactively Connecting With Connected Customers and Addressing Telecom Churn