Udacity Machine Learning Nanodegree 2020 Capstone Proposal

Customer Retention in Telecommunications

Michael George

March 2020

Domain Background

Customer churn occurs when a customer (player, subscriber, user, etc.) ceases his or her relationship with a company.

The ability to predict that a customer is at a high risk of churning, while there is still time to do something about it, represents a huge additional potential revenue source for every business.

The full cost of a churning customer includes both lost revenue, marketing costs involved with replacing those customers with new ones and that the costs of initially acquiring that customer may not have already been covered by the customer's spending to date. It is always more difficult and expensive to acquire a new customer than it is to retain a current paying customer.

Businesses that fail to address churn suffer further debilitating consequences in reduced attractiveness to investors and doubts about their future viability. It's essential to measure, monitor, and reduce churn. Reducing churn is a key business goal of every business.

In order to succeed at retaining customers, the business must be able to

(a) predict in advance which customers will churn; and

(b) know which actions will have the greatest retention on each particular customer.

So the main objectives of this project will be to use Machine Learning techniques to classify 'churning' customers and their attributes for a given business dataset.

Problem Statement

Every business has customers that cease doing business with them. Failing to deal with churning customers has major consequences on a business, so the ability to predict and inhibit these customers from leaving is a must.

The business goal of this exercise is to apply Machine Learning techniques to:

1. Analyze customer specific data to understand who could be the next potential customers to leave the business.
2. Find what attributes contribute to the higher churn rate of customers and what could be some of the solutions to address this.

Datasets and Input

For this project we will use a dataset called Telco Customer Churn dataset from Kaggle [1] provided as a CSV file.

This dataset contains a total of 7043 rows. Each row is unique for a customer and is identified using customerID. The target column for classification is 'Churn'.

The dataset contain total 21 columns whose details are listed below:

- 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)

Assumptions made :

1. The sample data is correct representation of the entire population and is randomly selected
2. The columns in the dataset are exhaustive list of features that determine churn rate

Solution Statement

The proposed solution is to apply supervised, binary classification models to identify churning and non churning customers. These models will then be used to examine 'churning' features.

First we will analyse the data to find who are the most common churning customers based on their categorical features (age, gender, etc). We will then dive deeper and analyse the common attributes of these churning customers based on features like contract length, payment and the services they buy. By the end we will correlate these to find feature importance.

Second, we will use classifiers from the SKLearn library to predict customer churn and the features that account for churn.

Lastly we'll see if we can improve further on this model by deploying it to AWS SageMaker to use one of their built in algorithms

Benchmark Model

For most customer churn models an accuracy of at least 70% is required, although we will aim to get a lot higher than this.

Evaluation Metrics

The evaluation metric for this problem will be accuracy score.

When tuning the model we will factor in precision and recall to limit the number of false negatives and positives.

Project Design

Notebook 1

Data Preprocessing:

- Identify different data types
- Clean the data

Data Analysis:

- Visualise churning customers
- Visualise categorical features
- Plot the feature importance as a heat map

Notebook 2

- Model Training and Evaluation on 10 SKlearn algorithms.
- Model Training and Evaluation on an ensemble classifier of the 3 best classifiers.
- Tuning of the best classifier
- Visualisation of the most important features.

Notebook 3

- Model Training and Deployment to AWS S3 of a LinearLeaner classifier.
- Evaluation and tuning of the hyperparameters of the model
- Visualisation of the most important features.
- Recommendations for lowering churn

References

[1] Telco Customer Churn dataset https://www.kaggle.com/blastchar/telco-customer-churn