Machine Learning Engineer Nanodegree

Capstone Proposal

Peter Shoukry
May 19, 2019

Proposal

Domain Background

Running a profitable business is quite complicated. You need to juggle lots of balls at the sametime. The ability to do so comes down to the ability to make lots of small decisions everyday. Taking out the guesswork out of the decision process can make all the difference between successful and failed businesses.

Entrepreneur at heart I am always interested in understanding how businesses work and that can't be separated from understanding human behviour.

similar research:

- https://thesai.org/Downloads/Volume9No2/Paper_38-Machine_Learning_Techniques_for_Customer_Retention.pdf
- https://www.h2o.ai/wp-content/uploads/2019/02/Case-Studies_PayPal.pdf

Problem Statement

Software as a service (SaaS), specially subscription based services, rely on the existing customer base to cover their expenses and to generally maintain a healthy cashflow. Customer churning can cost a business dearly in;

- Cost of aquiring new customers
- Bad publicity through word of mouth

ref: https://www.salesforce.com/blog/2013/08/customer-service-stats.html

Datasets and Inputs

Because of the great difference between services types, user activity, etc... the model is highly senstive to the dataset that it will train on.

There are two different open data sets that I found:

- https://www.kaggle.com/c/kkbox-churn-prediction-challenge/data (~8GB of SAAS user data)
- https://community.watsonanalytics.com/wp-content/uploads/2015/03/WA_Fn-UseC_-Telco-Customer-Churn.csv

Although the first dataset seems interesting and would give lots of insights into customer churning the size of the dataset makes tackling it a problem on it's own and will require lots of effort in data cleaning, preparation and extraction of a representative sample hence I opted for the second dataset which would help me focus more on the ML aspect of the problem.

Quick Statistics

- The data shape is (7043, 21)
- The number of churning users are 1869 (~26.5%) hence the dataset is unbalanced

Solution Statement

The main objective of the project is to build a model to predict if a customer is going to churn or not. Given data about the user usage of the service, the model can classify into two classes: churning or not-churning. Because of the small size of the selected dataset it is more suitable to use classic ML techniques to deep learning.

Benchmark Model

A simple model that assumes not-churning for every customer will be used as a benchmark model.

Evaluation Metrics

I will use F1 score as the metrics to evaluate the model since the dataset is not balanced.

Project Design

The project will consist of the following steps:

1. Data processing
    - Cleaning
    - Normalization
    - outlier handling
2. Splitting
    - Learning
    - Validation
    - Testing
3. Model trainging and validation
   Will try multiple techniques to identify the best model. My initial thoughts are to use XGboost however I might also try LightGBM.