Machine Learning Engineer Nanodegree
Capstone Proposal

David Popoola
April 22nd, 2020

Xente Purchase Prediction (from Zindi Competition)

Domain Background

Xente is an e-payments, e-commerce, and financial services company in Uganda offering various products and services that can be paid for using Mobile Money (Airtel Money, MTN Mobile Money), Bank Card (Visa Card, Master Card), Xente wallet and on credit (Pay Later). Some of the products consumers can buy include airtime, data bundles, event tickets, movie tickets, bus tickets, and more. They can also pay water and electricity bills and TV subscription services.

Xente is a Ugandan e-commerce startup that makes it easy for consumers to make payments, get loans, and shop using simply a mobile phone.

Problem Statement

Create a machine learning model to predict what and when individuals will purchase next, based on their purchase history.

This is a prediction problem and can be framed in the context of binary classification. Inputs are records of purchases made by different customers on different days for different products.

The goal is to analyse this data and predict if a customer will purchase a certain product on a certain day.

I will be visualizing input, observing and making intuitive opinions on the data given.

Then I plan to use different supervised learning techniques to train the dataset, and gain better models.

Datasets and Inputs

The datasets are provided by Xente on Zindi competition website. They are free to download.

Input Data fields

- PID – the unique id given to a particular product.
- Date – Previous dates of purchases made by customers
- Account ID – Unique ID given to customers on the platform

- Prediction - the target variable, set to 1 if a particular customer will purchase a particular product on a certain day, and 0 otherwise.

Solution Statement

The solution will be predictions of either 'will purchase' or 'will not purchase' in the test dataset. First I will do some visualizations on the train datasets to get some understanding and intuition of records and purchase patterns. I will then move on to begin training different models on the data.

For training models I will compare logistic regression, decision trees, nearest-neighbors, SVM, XGBoost and CatBoost since this is a classification problem. Finally I will select the best model for this problem and fine tune parameters to get best accuracy.

Benchmark Model

For this problem, the benchmark model selected will be the random forest models. I will try to beat its performance with other algorithms.

Evaluation Metrics

Prediction results are evaluated on the F1 score between the predicted values and actual purchases values ('1' for purchased and '0' for not purchased). According to Zindi competition webpage.

Since this is a Zindi competition project, I will take the leaderboard score as my evaluation.

Project Design

Before even start training models, I will analyse the data, make visualizations and see what the shape is and and try to make intuitive patterns from the data. Then I will do necessary feature engineering based on the input given in the train datasets. This is a problem with features and labels, so its better managed with supervised learning models since we know what we are predicting(label). Due to few features, Principal Component Analysis, clustering techniques may not be of great advantage.

To train models, I plan to choose 3-4 different models to compare. Because this is a classification problem, a few approaches that come to mind are regression, decision trees, SVM, KNN, and random forest. Using cross-validation I can find which model performs best, and then use that one to tweak different hyper parameters.

The final F1 score will be calculated against the test data set provided by Zindi.

Reference

1. https://zindi.africa/hackathons/umojahack-2-xente-purchase-prediction-challenge
2. https://www.xente.co/