Machine Learning Engineer Nanodegree

Capstone Project Proposal

By: Alvaro Azabal
Date: 18th June, 2020
Project Name: Starbucks Capstone Project

Domain Background

As the largest coffeehouse retailer in the world, Starbucks possesses vast amounts of data. Some of this data refers to targeted advertising. The business insights that can be obtained through proper analysis of targeted advertising are extremely valuable. This data contains information about which customers receive special offers, which customers actually use these offers, how much do they spend, etc. Machine learning offers a large variety of techniques that could be applied to this data to gain useful information about the targeted campaigns. One of these useful techniques is clustering. Using this technique one could group customers in different clusters depending on their spend behaviour, thus predicting how these customers will behave in the future when receiving new offers. Another possible machine learning technique that could be used is a classification algorithm to predict if a customer will respond when receiving an offer or not. There are several classification techniques (decision trees, neural networks, etc) that will be explored in the project.

Problem Statement

As briefly discussed in the previous section, this project will aim to solve two main problems: using demographic, transaction and offer data, first it will aim to group customers in relevant clusters, via clustering techniques such as K-Means Clustering; then a classifier will be built to help Starbucks predict if a customer will positively respond (ie: will make a purchase) when receiving a special offer. This will be a very useful model for future more accurate targeted advertising campaigns. There are a large amount of possible classifiers to use. This project will investigate using deep neural networks for this problem, as well as decision trees.

Datasets & Inputs

The data is contained in three files:

- portfolio.json - containing offer ids and meta data about each offer (duration, type, etc.)
- profile.json - demographic data for each customer

- transcript.json - records for transactions, offers received, offers viewed, and offers completed

Here is the schema and explanation of each variable in the files:

portfolio.json

- id (string) - offer id
- offer_type (string) - type of offer ie BOGO, discount, informational
- difficulty (int) - minimum required spend to complete an offer
- reward (int) - reward given for completing an offer
- duration (int) - time for offer to be open, in days
- channels (list of strings)

profile.json

- age (int) - age of the customer
- became_member_on (int) - date when customer created an app account
- gender (str) - gender of the customer (note some entries contain 'O' for other rather than M or F)
- id (str) - customer id
- income (float) - customer's income

transcript.json

- event (str) - record description (ie transaction, offer received, offer viewed, etc.)
- person (str) - customer id
- time (int) - time in hours since start of test. The data begins at time t=0
- value - (dict of strings) - either an offer id or transaction amount depending on the record

Solution Statement

A clustering algorithm will be created to group customers in different clusters depending on their consuming behaviour. To do so, first Principal Component Analysis will be done for dimensional reduction, as the input data has a large number of features, which makes clustering a more unstable problem.

A binary classification model will be built to predict if a customer will buy a product when receiving a special offer.

Benchmark Model

There is no literature related to this particular problem. Therefore, the final selected model will be benchmarked against similar models used for this same application. This will be done by testing several different machine learning algorithms (using GridSearch cross-validation or similar methods) and then evaluating these results against each other.

Evaluation Metrics

Evaluating the clustering model will pose a difficult challenge given that this is an unsupervised learning problem, thus there is no ground truth to evaluate the model performance. Several techniques such as the elbow method or a silhouette analysis will be used to obtain the ideal number *k* of clusters. Then these clusters can be analysed by looking at the centroid locations, the number of customers in each cluster and the average value of the features in each cluster. If the centroids are evenly distributed, the number of customers in each cluster is more or less the same and the average value of the features in each cluster is distinguishable from the others, then we can assume that the clustering was effective.

To evaluate the binary classification model, the dataset will be divided in three sets (training, validation and test). The accuracy, precision, recall and f1-score will be used to evaluate how good the model is. To ensure that the model is learning properly and not overfitting, a validation dataset will be used, making sure that the evaluation metrics mentioned above are in the same order of magnitude for both the training and validation datasets. Once the model has been trained, it will be evaluated using the test dataset, using the evaluation metrics already discussed.

Project Design

Data Preprocessing

The first step will be to load and explore the data. Next, any missing data will have to be removed or predicted. Then, the data will have to be wrangled so that it is in the proper format for training. Some of the features are classes with labels, rather than numeric, so they will be one-hot-encoded.

Feature Engineering

As the number of features is quite large, PCA will be performed on the data to reduce the number of dimensions before clustering the data. First one will have to explore the ideal number of components to reduce to in order to capture as much variance as possible with as little number of components as possible. Then the fitted PCA will be used to transform the data.

Clustering

A clustering algorithm will then be fitted on the reduced data. To identify the ideal number of clusters the elbow method and a silhouette analysis will be done. Different clustering methods and number of clusters could be investigated.

Classification

Train - Test - Validation Split

For any classification algorithm, first the data will be split in the three datasets required.

Model Architecture and Hyperparameter Tuning

Different classifiers will be explored (decision trees, neural networks, etc) and a cross-validation method (such as GridSearch-CV) can be used to select the best model for the problem and the most tuned hyperparameters.

Evaluation

Using the test dataset the selected model can be evaluated. As discussed, this is a binary classification problem so the accuracy of the model will be evaluated, but also the precision, recall and f-1 score.