Machine Learning Engineer Nanodegree

Capstone Proposal

By Vishad Bhalodia

Domain Background

Rewards, offers and loyalty programs – these are the cornerstone of every company's marketing strategy in this digital age. Starbucks Rewards Program generates customer loyalty, increased revenue, and data for the company to create meaningful 1:1 relationships and personalized marketing efforts. Starbucks has attributed their rewards program for most of the increase in their revenues over the past 2 years. In this project, we will analyze the offers to see if we can predict whether the user will respond to the offer or not based on their characteristics (spending/age/gender/income).

Problem Statement

I will analyze and answer two questions for this project –

1. Exploratory Analysis - Use the data to identify which groups of people are most responsive to each type of offer and how does it related to their characteristics.
2. Modelling - Build a model that predicts whether a customer will respond to an offer or not (1/0).

The predictions for the second part of the problem will be evaluated on accuracy/F1 score since it is an appropriate metric to evaluate binary classification problems. Based on this metric we will calculate how likely a person is going to respond to this offer making it as 0 or 1.

Datasets and Inputs

The datasets provided by Udacity and Starbucks contain simulated data that mimics customer behaviour on the Starbucks rewards mobile app. The dataset contains hidden details on how people make purchasing decisions and how those decisions are influenced by offers which need to be extracted in a meaningful way.

The capstone instructions provide 3 datasets, along with the instructions –

1. The first dataset is **portfolio,** which contains the list of all available offers that can be shown to the user. Each offer can be one of the three categories – discount, BOGO (buy one get one) or Informational.

   The dataset contains offer ids and their corresponding meta data about each offer:

   - reward: money awarded for the amount spent
   - channels: channels through which offer was delivered
   - difficulty: money to be spent in order to receive the reward
   - duration: time till the offer expires in days
   - offer_type: bogo, discount, informational
   - offer_id

2. The second dataset is **profile**, which includes the details of all users that have used the app. For each profile, the dataset contains personal information in order to better categorize each user. Demographic data for each user contains the following 4 fields/features:

   - gender
   - age
   - became_member_on
   - offer_id
   - income

3. The third dataset is the **transcript** dataset: It is essentially an event log of all actions recorded on the app for offers, which includes user's transactions data as well the time lag between the offer delivery and the offer use by the user.

   - person
   - event: which can be offer received, offer viewed, transaction, offer completed
   - value: contains the amount spent in "transaction and the reward gained from using the offer
   - time: time difference between offer shown and offer use

Solution Statement

The goal is to develop a supervised machine learning model to predict whether the user will respond to the offer or not. For this, I'll use several machine learning models - to evaluate which performs best on the dataset. It is also possible to implement a small, simple neural network model with one hidden layer to see if we can improve accuracy of the prediction.

Benchmark Model

We can use a naive model that assumes all offers were successful as a baseline for evaluating the performance of models that will be used.

Evaluation Metrics

The result that needs to be evaluated is whether we successfully predicted that the user responds to the offer or not. We will use different statistical measures like accuracy score, F1 score to evaluate the model predictions.

Project Design

The workflow for approaching the solution detailed above includes several machine learning techniques, given below -

- **Data loading and exploration**
  Load files and present some data visualization in order to understand the distribution and characteristics of the data, and possibly identify inconsistencies like missing values, data skewness and categorical features with too many categories.

- **Data cleaning and pre-processing**
  Having analyzed the data, we now apply several standard techniques to fix possible issues found. These include (but not limited to) missing imputation, categories encoding, and data standardization.

- **Feature engineering and data transformation**
  Prepare the data to correspond to the problem stated and feed the machine learning algorithms. The transcript dataset must be structured and labelled with an effective offer label for supervised learning.

- **Preparing different models and transform inputs for models**

  After that, we develop the different models that we want to check our predictions on. We create different machine learning models using different algorithms such as Gradient Boosting, Logistic Regression, Support Vector Machine, and Random Forest.

- **Model Prediction and Tuning for Improvement**

  We then evaluate the models on accuracy and F1 score – if we do not see significant improvement relative to our benchmark model, we apply model tuning using GridSearchCV.

- **Conclusion**

  Finally, we present the resulting predictions, along with justification on model performance and future improvements.