Machine Learning Engineer Nanodegree
Capstone Proposal

Hao Yun

November 22, 2022

1 Domain Background

Many companies have their reward programs which are designed to encourage people to pay for their products or services. Customers are interested in getting rewards. This also benefits companies in attracting and retaining customers. A variety of rewards are provided in the market. People may respond differently to certain type of rewards. Comparing with giving the whole population the same reward, providing variant rewards at an individual personalized level not only helps companies improve the profitability, but also improves customer experience.

When people decide to join a reward program or a membership, companies will get personal information such as name, age, gender and financial information. The data generated from the transaction records is equally important to a company.

Analyzing demographic data and transactional data helps companies understand cutsomers better. And contributing a statistical model or a machine learning model based on these data can help marketers know customer preferences, evaluate products and services, predict trends in the future and make new marketing strategies.

2 Problem Statement

Starbucks has provided an experimental dataset on kaggle [1]. Two main problems can be solved by analyzing this dataset. One goal is to predict how much a customer will spend at Starbucks during the experiment period. Another is to predict the probability that a customer will respond to a reward, so that Starbucks can come up with new strategies to determine what kind of reward will be more appropriate for a certain customer.

3 Datasets and Inputs

Starbucks dataset contains simulated data that mimics customer behavior on the Starbucks rewards mobile app. And it's a simplified version that the simulator has only one product of Starbucks. The dataset has three separate files including offer information, demographic data and records of transactions during the test month. More details about the variables of each file are available on kaggle.

The objective of this project can be realized by a regression model. Since three separate files are provided, the first step is to combine the files, then select the features that will be used in the model.

The target variables are the total amount that each customer spends at Starbucks and the probability that each customer completes offers. Not all variables in the dataset are appropriate to predict the target variables. And some variables need some processing. Eventually, the dimention of the input data will be 5. The input variables are age, gender, income, the number of times customers received offers and the proportion of offers they have viewed.

4 Solution Statement

The problem requires the predictions of continuous output variables. This is obviously a supervised learning problem, more precisely, a regression problem. The main task is to learn a mapping from multiple input variables to numerical variables. Various kinds of algorithms can be used to solve this problem, such as linear regression, random forest and neural network. After modeling, output values of an amount and a probability will be predicted from a 5-dimention input data. The models can be evaluated by comparing the predicted values and true values.

5 Benchmark Model

Since the dataset is not for competition, there isn't a leaderboard or a similar problem on kaggle. So a multivariate linear regression model will be used as a benchmark model because it's easy to implement and efficient to train. There are many metrics used to evaluate the performance of a regression model like Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE) [2].

To predict the amount and the probability, there might be other models that perform better with a lower error. MAE measures, on average, the absolute values of the differences between predicted values and true values. It will be used to determine whether the model is better compared to the linear regression model.

6 Evaluation Metrics

Mean absolute error (MAE) is a measure of errors between paired observations expressing the same phenomenon and is calculated as the sum of absolute errors divided by the sample size [3]:

$$MAE = \frac{\Sigma_{i=1}^{n} |y_i - \hat{y}_i|}{n} \tag{1}$$

In this project, the main task is to solve a regression problem. MAE is an appropriate metric that can be used to quantify the performance of both the benchmark model and the solution model. A low MAE means that predicted values are close to true values.

7 Project Design

To implement a regression model, the following steps are required:

- Data loading and exploration

  Starbucks dataset contains three separate json files. The first step is to combine the variables into a format that can be used in the model which will be contributed based on customer information. Then identify the category of variables, find the distribution of variables and compare correlation of variables.

- Data cleaning and pre-processing

  There will be missing values and outliers. Typically, outliers can be removed. How to deal with missing values will depend on the situation.

- Feature engineering and data transformation

  Not all variables are useful to predict the target variables, such as offer id, customer id. And there are variables that can not be used directly in the model like offer type and the time that a transaction happens. Some numerical features can be extracted from these variables, for example, the number of times customers received offers or viewed offers.

  Then the processed data will be split into training data and test data. For neural network, data normalisation or standardization is required and very important. However in a linear regression and xgboost, this step is not necessary. In order to compare the metrics of models, data will be normalized or standardized for all models in this project.

- Defining and training models, making improvements on the models

  Three models will be trained using training data. The benchmark model will be a multivariate linear regression. Additionally, xgboost and neural network will be implemented.

  After training the model, make predictions on test data. The performance of models can be improved by tuning parameters.

- Evaluating and comparing model test performance

  The final result will be quantified based on metric values. The last step may be to do some reflections according to the process of this project to consider if it exists a better solution or if there are more strategies to improve the model.

References

[1] Starbucks app customer rewards program data

[2] Choosing right metrics for regression model

[3] Mean absolute error - Wikipedia