Capstone Project | Mohd Zaki Aftab

Machine Learning Engineer Nanodegree | 31st October 2020

Starbucks Project

Domain Background:

The task is to combine transaction, demographic and offer data to determine which demographic groups respond best to which offer type. This data set is a simplified version of the real Starbucks app because the underlying simulator only has one product whereas Starbucks sells dozens of products.

Problem Statement:

The objective of the venture is to manufacture an AI model that can be utilized inside web application to handle genuine world, clients. I'd be focusing on creating an optimal model that would yield correct predictions on the dataset, that is to check if a particular offer is suitable for a customer or not.

Datasets and 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:

I'll be creating and evaluating models on the cleaned data, evaluate the performance and select the correct model. Owing to the complexity of the dataset one can expect the test accuracy of the model to be around 60%. This model is basically about sending offers to people, so this amount of accuracy is fair enough. The data would be cleaned, and the trends would be taken into consideration, after that one can make sure that a specific dataset is created by combining the three datasets. The dataset be tested on various classifiers and the accuracy to be reported. Finally, one can deploy the final model to give out predictions.

Project Design:

Stage 1: Import the vital dataset and libraries, Pre-measure the information and make train, test and approval dataset.

Stage 2: Clean the data find useful observations.

Stage 3: Analyze and create a dataset which can be used to test and train the data

Stage 4: Test and train the data on each model

Stage 5: Check the accuracy and select the most accurate model.

Stage 6: Deploy a final predictor that would help in predicting which offer type is suitable.

References:

1. Udacity Starbucks Dataset
2. https://machinelearningmastery.com/gradient-boosting-machine-ensemble-in-python/
3. Extra Tree Classifier: https://www.geeksforgeeks.org/ml-extra-tree-classifier-for-feature-selection/#:~:text=Extremely%20Randomized%20Trees%20Classifier(Extra,to%20output%20it%27s%20classification%20result.
4. Sklearn Documentation: https://scikit-learn.org/stable/