Capstone Project

Machine Learning Engineer
Nanodegree

Divya Dayashankar Jaiswal
October 30th, 2020

Domain Background

Machine Learning (ML) is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. ML focuses on the development of computer programs that can access data and use it to learn for themselves. It has become an important part of our future technology. Where the companies will be implementing the ML algorithms on the huge data, they have.
Starbucks is one of flagship worldwide company having coffeehouse chain with tremendous database of users. Having the transactional data showing user purchases with timestamp and the amount of money spent on the purchase. This transactional data also has a record for each offer that user receives as well as a record for when a user actually views the offer. There are also records for when a user completes an offer.
Therefore, using this data, my model will determine which demographic groups respond best to which offer type.

Problem Statement

Starbucks wants to find a way to give customer the special offer in the app. The goal is to determine which demographic groups respond best to which offer type. To analyze data of the users and then to develop an algorithm that associates each customer to the right offer type.

Datasets and Inputs

We have three JSON Files (data dictionary):

- profile.json: Rewards program users (17000 users * 5 fields)
    - gender: (categorical) M, F, O or null
    - age: (numeric) missing value encoded as 118
    - id: (sring/hash)
    - became_member_on: (data) format YYYYMMDD
    - income: (numeric)
- portfolio.json: Offers sent during 30-day test period (10 offers * 6 fields)
    - reward: (numeric) money awarded for the amount spent
    - channels: (list) web, email, mobile, social
    - difficulty: (numeric) money required to be spent to receive reward
    - duration: (numeric) time for offer to be open, in days
    - offer_type: (string) bogo, discount, informational
    - id: (string/hash)
- transcript.json: Event log (306648 events * 4 fields)
    - person: (string/hash)
    - event: (string) offer received, offer
    - value: (dictionary) different values depending on event type
        - offer id: (string/hash) not associated with any "transaction"
        - amount: (numeric) money spent in "transaction"
        - reward: (numeric) money gained from "offer completed"
    - time: (numeric) hours after start of test

Solution Statement

To develop machine learning (ML) model to determine the best offer type for each customer of Starbucks.

- Fetching the data: Convert the data sets mentioned above to CSV files.
- Cleaning the data: Remove duplicates, deal with missing values, normalization, data type conversions, etc…
- Data visualization and analysis: To visualize data with appropriate relationships between variables and then to split data into training and test data sets (80:20 ratio). Hyperparameter tuning, for model efficiency and good improvement.
- Training model: To make a training model and train the model on the train data sets
- Testing the model: Using metrics or combination of metrics to measure the performance of the model on test data set.

Benchmark Model

We will use Logistic regression model as our Benchmark to compare our model's performance to, as it is faster and simple to implement. Also, we will implement the AUC, Precision and Recall metrics to compare other model's result.

Evaluation Metrics

- Area Under the ROC Curve, a measure that calculates the area under the Receiving Operating Characteristic Curve. This particular curve accounts that higher probabilities are associated with true positives and vice-versa.
- Precision/Recall, the percentage of true positives (conversions predicted correctly) on all positives (predicted as conversions)/on all conversions. These two measures contrast each other: tuning the model to grow the precision results in a smaller recall and vice-versa.

Project Design

- Programming Languages, tools and libraries:
    - Amazon Sagemaker Machine Learning Services
    - Amazon Sagemaker XGBoost built in Algorithm
    - sklearn.tree
    - sklearn.neighbors
    - sklearn.ensemble
    - sklearn.linear_model
    - sklearn.utils
    - sklearn.metrics
- Models Used:
    - Amazon Sagemaker XGBoost built in Algorithm
    - Random Forest Regressor
    - Decision Tree Classifier
    - K-neighbors Classifier
    - Logistic Regression
    - CatBoost Model
    - LightGBM Model