Machine Learning Engineer Nanodegree Program

Capstone Project Proposal

Starbucks Capstone

Devesh Katiyar
Feb-2020

Domain Background

Machine learning has become an increasingly important part of IT today and usage of ML is increasing daily in almost every domain. However, the power of ML is not utilised to its fullest yet and can be useful for society and market in unimaginable ways. ML can also be useful for various companies in improving their products, sales to offer better customer service and earning a higher growth rate.
STARBUCKS is one of the flagship Worldwide companies which has been established since Mar-1971 and worldwide coffeehouse chain, and has a tremendous database of users. Starbucks offers their free app to make orders online, predict the waiting time and offer better service. That's why analysis on app usage is more crucial to leverage the business and understand the customer's behavior.

Problem Statement

As Mentioned in **STARBUCKS Capstone Challenge,** analyze the Starbucks Customer dataset and build a **model that can make better offers to customers**. Our goal is to analyze the dataset made available by Starbucks about the app usage and offers/orders made by the customer to make a model that can make better offers to the customers so that they redeem the offer made by the model which will finally increase the sales of the Starbucks and help them to reach more new customers.

The problem that is stated in the project is a classification problem and we need to follow a classification approach to solve this problem. As per the problem it is very clear that we need to classify our customer on the basis of various features like age, gender, income and then predict a offer(y) for them on this basis so that they are more likely to redeem the offer.

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

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 the start of the test. The data begins at time t=0
- value - (dict of strings) - either an offer id or transaction amount depending on the record.

The dataset was made available to us by Starbucks and on exploring the dataset we find out that some of features of the dataset included missing data which need to be resolved we can either remove those columns\rows for missing values or we can replace them with mean or the highest value. We will choose the replacement with mean value as this will make our task easier. We also find out that dataset also includes categorical features (web, mobile,

social, email) in social category which need to be solved and turned into numerical features with the help of either LabelEncoding or OneHotEncoding. We chose Here OneHotEncoding for this dataset.

Solution Statement

My strategy is to develop a Machine Learning model to predict which is the type of offer for each customer, such that the offer proposed to the customer turns into a sale. I will develop a model for each offer type and then combine the results to have the best action for each user.

Benchmark Model

We will use different models to get the best out of our dataset.

This will include Decision Tree Classifier, Adaboost Classifier and RandomForest Classifier and the model which will show the best performance will be our final model for the implementation.

On the basis of performance of the model DecisionTree Classifier model was the one that was showing highest accuracy but we RandomForest Classifier instead as RF Classifier has a less probability of overfitting the data while Decision Tree is prone to overfitting and noise in the dataset.

Evaluation Metrics

We will use accuracy score and F-score as the evaluation metrics of our model.

F-score is a metric that is based on the concept of Precision and Recall, hence it is better for our model evaluation.

Project Design

First we load our dataset for exploring by using pandas which was in json format and changed in dataframe to make it ready to be used for our solution.Then we explored the various features if dataframe, while exploring we found out that various features of the dataset contains missing values which are need to be resolved as they will affect the final model. We then solved the problem of categorical variables and turned them into numerical features so that they can be used as the part of the model. Then we dropped the columns for which we have created the dummy variable. Then we visualized

the dataset on the basis of income, age, gender to gain the better insight of the dataset we used boxplot and histograms form seaborn and matplotlib for visualization. Then we normalized our features on the same scale so that each feature equally contributes to the model. Then we trained our algorithms from sklearn i.e DecisionTree, Adaboost, RandomForest Classifier to check which of the performing better on the dataset and would be good for model making we evaluated the algorithms on the basis on accuracy score and F1-score from sklearn.metrics, Later we found out that RandomForest would be a good choice for our problem and we tuned the hyperparameter of the algorithm and finally we made our model on the RandomForestClassifier