Machine Learning Engineer Nanodegree

Capstone Proposal

- Domain Background:

Starbucks provide an app for customers to make online purchases so in willing to increase sales and customers satisfaction so once every few days, Starbucks sends out an offer to users of the mobile app. An offer can be merely an advertisement for a drink or an actual offer such as a discount or BOGO (buy one get one free).

- Problem Statement:

Starbucks wants to find the best offer to send to customers based on demographics information and past transactions that user made, so we can do that if we can predict if the user will complete the offer or not and as the output is determined so it will be supervised learning problem and as output is completed or not completed will be a classification problem

- Datasets and Inputs:

- **portfolio.json** : containing offer ids and meta data about each offer
- **profile.json** : demographic data for each customer
- **transcript.json** : records for transactions, offers received, viewed, and completed

1- portfolio.json size = (10 rows , 6 features)

- 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)

2- profile.json size = (17000 records,5 features)

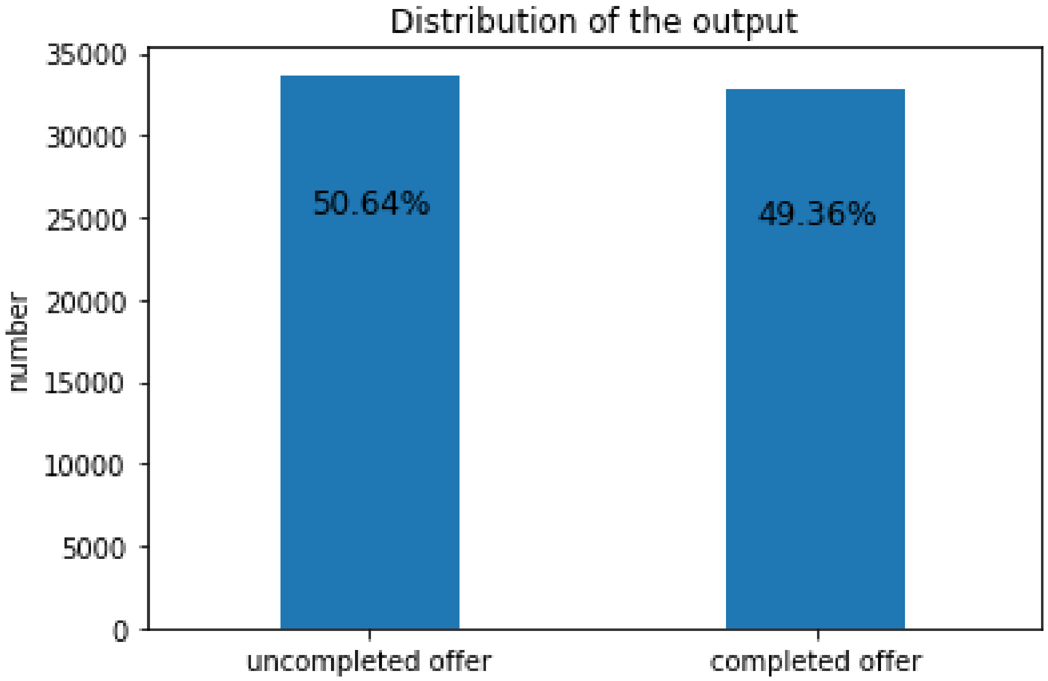
- age (int) - age of the customer
- became_member_on (int) - date when customer created an app account
- gender (str) - gender of the customer
- id (str) - customer id
- income (float) - customer's income

3- transcript.json size = (306534 records,4 features)

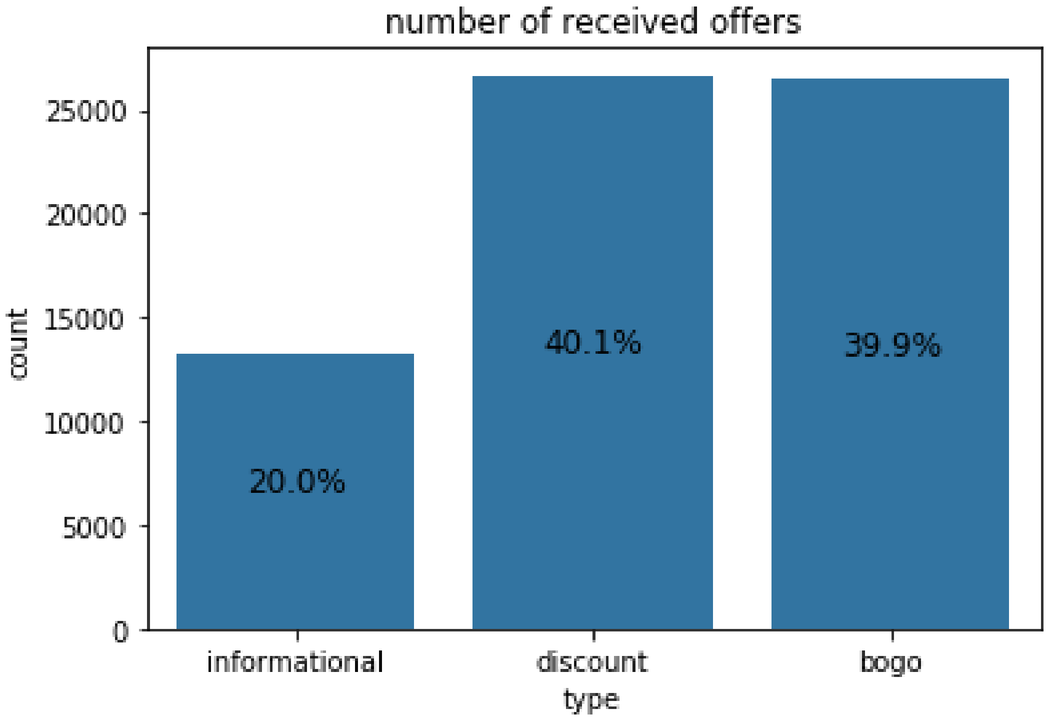
- event (str) - record description (ie transaction, offer received, offer viewed, etc.)
- person (str) - customer id
- time (int) - time in hours since start of test. The data begins at time t=0
- value - (dict of strings) - either an offer id or transaction amount depending on the record

- Inputs:

The main data will be transcript.json and I will join all another data on it, the input that I will use will be all records that related to offer receive so we will short our data to only these records and for every user I will aggregate all transaction he/she made before he/she received the offer.



Types of offers:



- Solution Statement:

Mu solution will be building a machine learning model that can predict if the user will complete the offer or not, by using multiple algorithm as logistic regression, SVM, gradient boosting, XGBoost and after tuning them, I will compare the errors between them and if it is there a small overlapping between them I will build a stack model training on the predicted probability of used models (if model allows probability prediction if not will use the normal output (1,0))

- Benchmark Model:

I will use logistic regression as a benchmark model with regularization L1 and it will be used during features engineering step do make sure that every feature will be added to the model will be helpful and due to L1 it will be multiple features with coefficient equal zero that will help me to understand if the features is helpless or not and finally coefficient of features will help me in features selection when I tune the others model

- Evaluation Metrics:

The distribution of output is balanced as the number of uncompleted offers is 336771 and completed offer is 32824 so we use accuracy as evaluation metrics with F1 to make sure that the difference between recall and precision is small

- Project Design:

1- data cleaning by fixing all quality issues and tidiness problems
2- data preprocessing by extracting all records related to offer receive and aggregate all transactions that made before he/she receives this offer so for every user that will be records as number of offers that he/she receive and for every record will contains all information that relates to the send offer as type, time, reward and difficulty and all information related before that transaction as number of offer received, number of completed offers, number of transaction that contain buying product, sum of all money user paid and many other features like that.
3- data analysis
4- feature engineering: from the aggregated features will try to know the rate of completed offer with respect to the offers user received and rated of received offers with respect to time and difference in time between the last time user received an offer and the current one..etc.
5- building multiple models using logistic regression as baseline model and random forest and gradient boosting and SVM and XGBoosting and I will compare the performance by measuring accuracy for each one with other model and baseline model.

Useful resources:

1- https://machinelearningmastery.com/how-to-get-baseline-results-and-why-they-matter/
2- https://developer.ibm.com/technologies/artificial-intelligence/articles/stack-machine-learning-models-get-better-results/