Capstone Project

Mohammed Adel

Machine Learning Engineer Nanodegree

January 22th, 2020

Definition

Project Overview

Companies sends offers to their customers all the time whether it was discounts, information about new products or limited time free offers. These offers are distributed to customer using various methods, through emails, SMS messages, company website, social media promotions, etc. But how active the customers are to these offers? Are the offers make an impact on company's sales? How many customers complete these offers? In this project Starbucks customer and offers data are provided to see offers impact on the customers.

Problem Statement

Using Starbuck provided dataset about customers and the offer they received and how much offers completed by the customers, but how active the customers are to these offers? What are the characteristics of customers who complete received offers? How many customers complete these offers? The goal of this project is to achieve the following:

- Have an insight about customers characteristics of those who complete received offers.
- Segregate completed offers by customers who received and did not view the offers and those who did.

- Build machine learning model that uses decision tree binary classifier to predict if an offer or group of offers given customer information will be completed or not.

Metrics

F-measure used to determine the accuracy of the model:

$$\text{Accuracy} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}.$$

Evaluating model accuracy using f-measure will make sure that the model does not classify those of wrong user characteristic group as customers who their habit is completing an offer as that will provide confusion in having an insight about the customers. F-measure will help in allocate if there are any false positives in predicted labels and if not, that will support the analysis and modification of the data set was beneficial in finding correct characteristics of customer and offer types that assure the offer will be competed.

Analysis

Data Exploration

Starbucks data is contained in three tables:

- Portfolio - containing offer ids and meta data about each offer (duration, type, etc.)
- Profile - demographic data for each customer
- Transcript - records for transactions, offers received, offers viewed, and offers completed

Here is the schema and explanation of each table field in provided tables:

Portfolio

- 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)

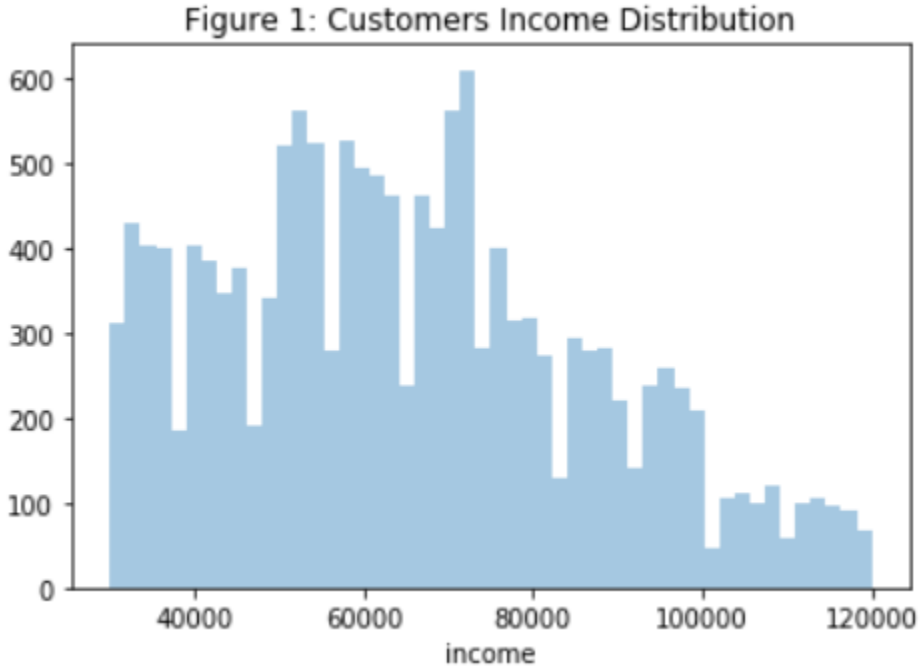
Profile

- 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

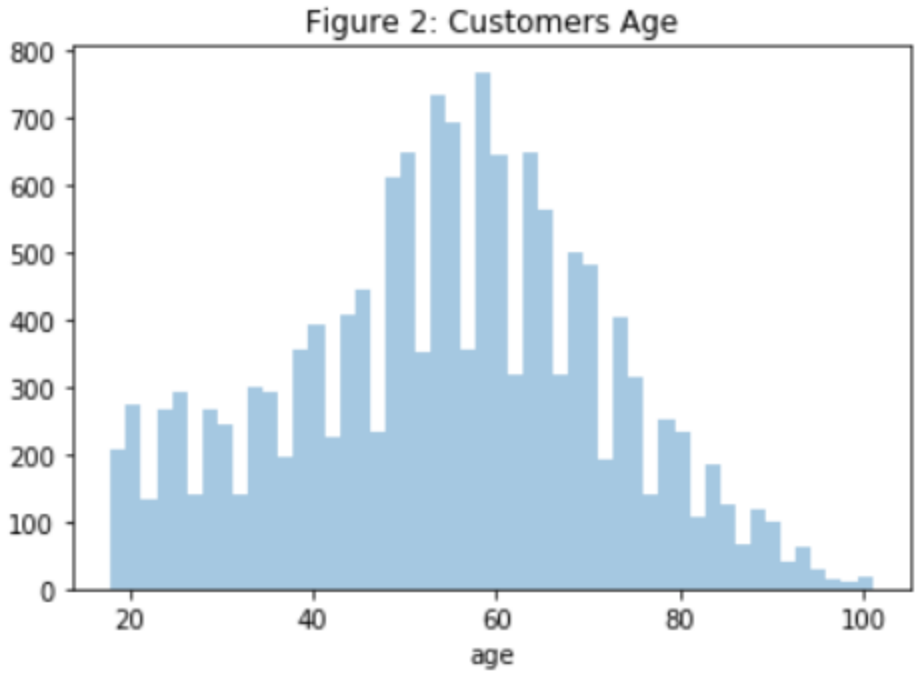
Transcript

- 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

Most of customers income are below or on the income mean (figure 1), which makes sense, people who are interested in these kinds of offers are mostly those who have an average or low income, in contrast people who have higher than average income may not be as interested.



Customers age distribution shows a normal distribution (figure 2), which maybe mean that age is not a significant factor in determining the characteristics of customers who will complete an offer.



Algorithms and Techniques

Algorithm used for the classifier are decision tree, it is a common algorithm performs great on binary classification problem, it is also fast and efficient and easy to visualize and know how the algorithm made its heuristics on the data given.

Decision tree algorithm works in iterated fashion, in each iteration a split decision is made based on some criterion, and each iteration shorten the distance to the leaf node where final decision is made, and prediction produced for the given example.

Benchmark Model

To create a benchmark for the classifier I create a logistic regression model to compare its result with my DT classifier, the LR model had classified the examples with 68% accuracy, against classifier, # classifier has accurately classifier the examples with F-measure preprocessed the data given and extracted features for the decision tree classifier, features are age, gender, income, duration, offer type and reward, the classifier will predict if an offer will be completed or not based on given offer and customer details.

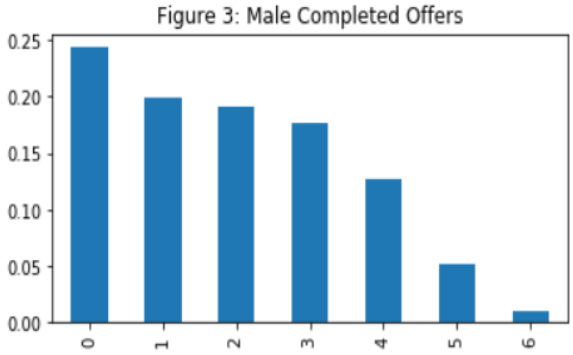
Methodology

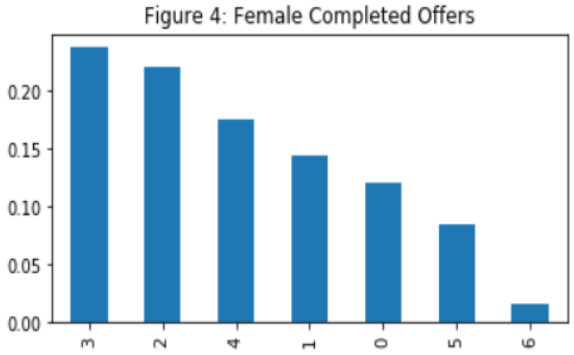
Data Preprocessing

I merged the three tables, first left joined profile (id) and transaction (person), then resultant table (id) with portfolio (id) to have one data-frame containing all customers, offers, and operations data.

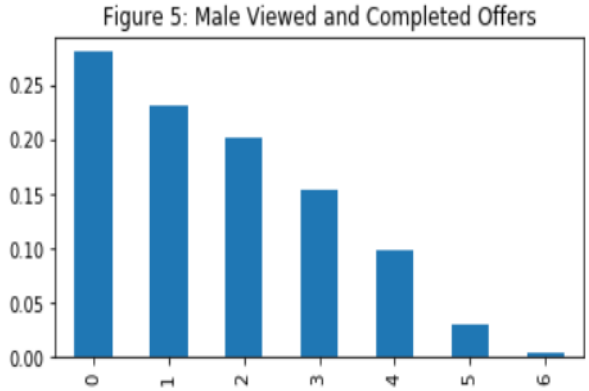
After that I separated completed offers by users who completed and viewed certain offer and those who completed and did not view the offer, the reason why if a customer completed an offer without seeing the offer, the reason that customer completed the offer was note because of Starbucks offer message, which will not be in use for our case, as the goal is to see how the offers impacts customers in completing a received offer.

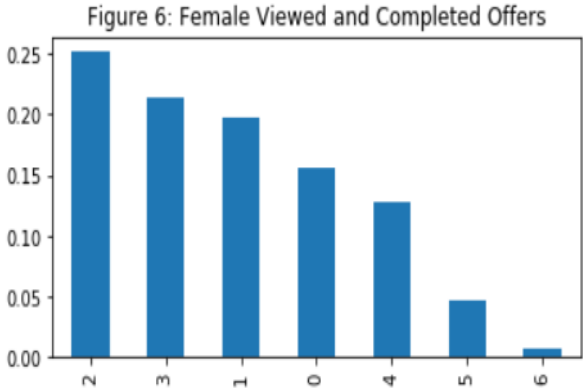
For example, figure [3 4] shows female and male completed offers without checking if a customer viewed the offer, and about 24% of men did not complete any offer, whereas women who did not complete an offer are about 12%.





On the other hand, figure [5 6] shows completed offer based on customer gender of those who completed and viewed the offer, the percentage increases for men and women who did not complete an offer compared to figure [3 4] by about 2% for men and 4% for woman, which is somehow significant given that we have important to allocate to provide accuracy to the data and trained ML model

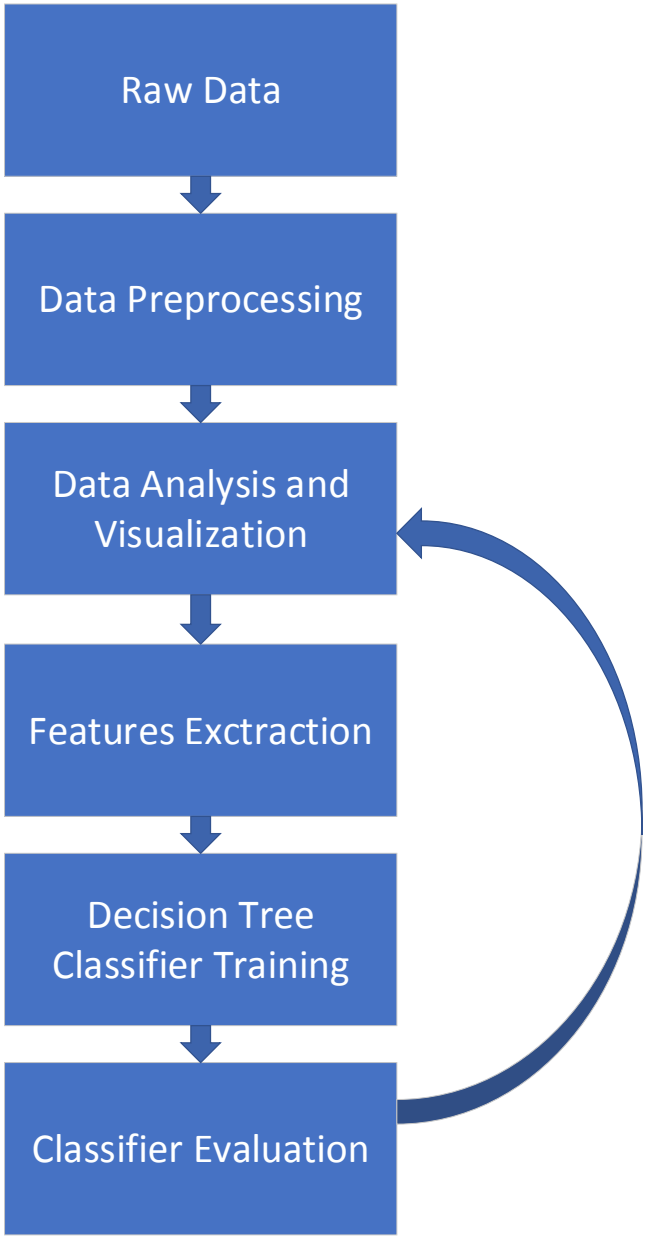




Implementation

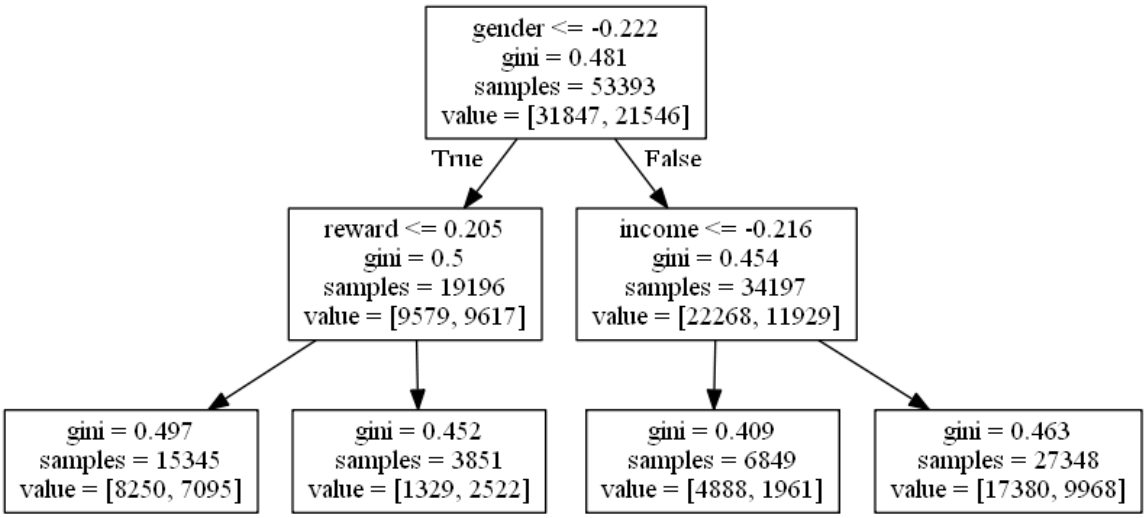
After preprocessing received raw data, analysis of the data has been made to see the factors and characteristic the determine if an example record will complete an offer, figure 7 summarize the whole process and workflow.

Figure 7 is the workflow and process went through during developing the solution.



Classifier training was done using decision tree algorithm, classifier fitted the preprocessed data and produced the following tree (figure 8) to be used for prediction (the DT classifier is set to max depth of 2 as is not visible to visualize the whole tree).

Figure 8 shows the first couple splits made by the fitter DT classifier



Refinement

Classifier was improved by including more example, some example had multiple empty fields, but other fields was there which would be waste of information if it was excluded, so, following techniques were used to fill missing fields:

- Mean normalization for empty fields in “income” column.
- Mean normalization for empty fields in “gender” column, after applying label encoding technique.
- Mean normalization for empty fields in “age” column.

Result

Model Evaluation and Validation

The DT classifier was evaluated using f-measure mentioned previously.

Accuracy = $$2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}.$$

model final accuracy evaluation is 76%, given the input size of 76277 examples 70% for training 30% for testing and evaluation.

Justification

It is not easy to build heuristics from given data, because as mentioned above different customers from different gender, income, and age group have contributed almost evenly in completing received offer, which makes it hard to draw conclusion of which offer will be completed from a certain customer with certain characteristics and offer details, in my opinion I find 76% an acceptable accuracy for the given data.

Benchmark model discussed previously has 68% accuracy, whereas solution model has 76% accuracy, which decently better than the benchmark model.

Conclusion

Free-form Visualization

Figure [3 4] shows female and male completed offers without checking if a customer viewed the offer, and about 24% of men did not complete any offer, whereas women who did not complete an offer are about 12%.

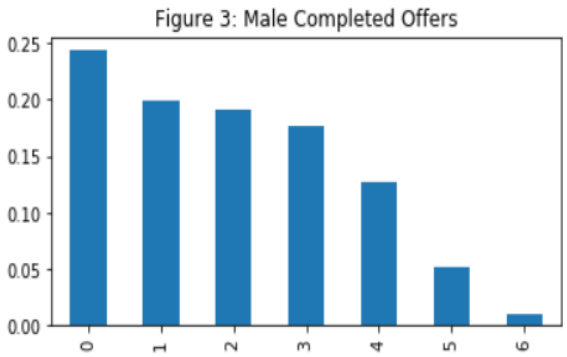


Figure 3: Male Completed Offers

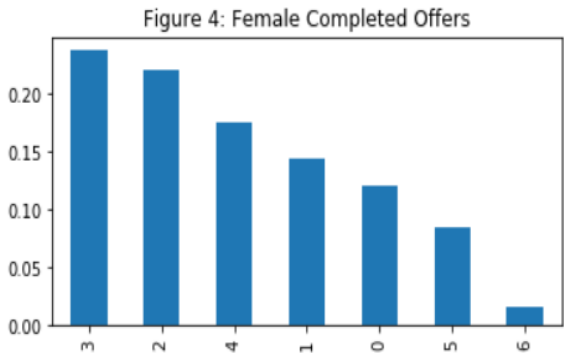


Figure 4: Female Completed Offers

On the other hand, figure [5 6] shows completed offer based on customer gender of those who completed and viewed the offer, the percentage increases for men and women who did not complete an offer compared to figure [3 4] by about 2% for men and 4% for woman, which is somehow significant given that we have important to allocate to provide accuracy to the data and trained ML model

Figure 5: Male Viewed and Completed Offers

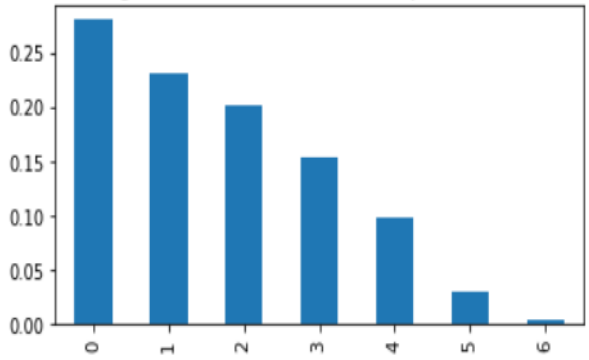
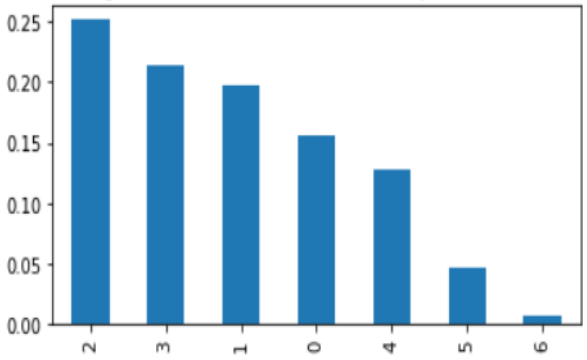


Figure 6: Female Viewed and Completed Offers



Reflection

Project scope was determining the combination of characteristics of customers and offers that determine if an offer will be completed given an example. Data was distributed in different files and to be collected and explored before preprocessing. It was tricky to find a significant factor that greatly impact the decision of an offer being completed or not. Following points summarize main reflections on the project:

- Columns in the dataset was general enough to know the main characteristics of a customer or an offer.
- One of the difficult parts was determining factors that impacted the decision of an offer being completed or not.
- Separating examples of completed offer by a customer who did not view the offer and those who did was a bit tricky, it was time consuming and code took a lot of time to run.
- Model built does a decent job in deciding if a customer will complete an offer, given offer and customer general details, but not necessarily generalized to be used in a production environment, it needs improvements which will be mentioned in improvements section.

Improvement

Decision tree algorithm is perfect for this project, it runs fast given a huge data size, and perfect for binary classification, but the model could be improved by the following:

- Get more profile, completed offers data to help the classifier determine the characteristics of an example input completing an offer.
- Consider only viewed and completed offers, not including completed offers by customers who did not view the offer in the dataset, to save computational time filtering the dataset.