Starbucks Capstone Project

Faisal Alsaeed

Data Science Nanodegree

Sep 4th, 2019

Definition

Project Overview

Starbucks collects its customers' data, and they are using it to knowing permanent customers, and they providing a special offers to their customers for certain products as well as marketing products better by reducing advertising costs and marketing less-selling products.

The data set contain some of the data that mimic customer behavior like the offer that sent to customer is he looked at the offer or not, did he receive it and how the customer is responding to the offers. In this project, I will be combining the data together and clean and analyze the data to extract some information from that data.

Starbucks has sets of data contain information about customers and the purchase transactions and they are want to understand their customers to provide them offers based on transactions and regularity of customers so they have some Questions like as:

- Forecasting the membership based on gender
- Average of customer age
- Analyzing the percent of customers base on year of register as a member
- most channels effect on customers
- Number of transactions based on age
- Number of transactions based on events

Problem Statement Metrics

The goal of the project is to build a model that predicts if the offer will be accepted by the customer or not. So the way is combining the two files together (the transcript file which contains all transactions and customers behavior) the second file is (profile file that contains demographics data about customers like age, income) and the last file is (a portfolio which contains the offers data like points, offer type, reward, and channels).

So, based on the model results I am going to evaluate the accuracy of the model by checking the F-score of model for a successful transaction that seen by the customer so if the accuracy of the model is low that is mean the model not good and we will not use it otherwise if the accuracy is high and it satisfied we will use the model on that data.

Analysis

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

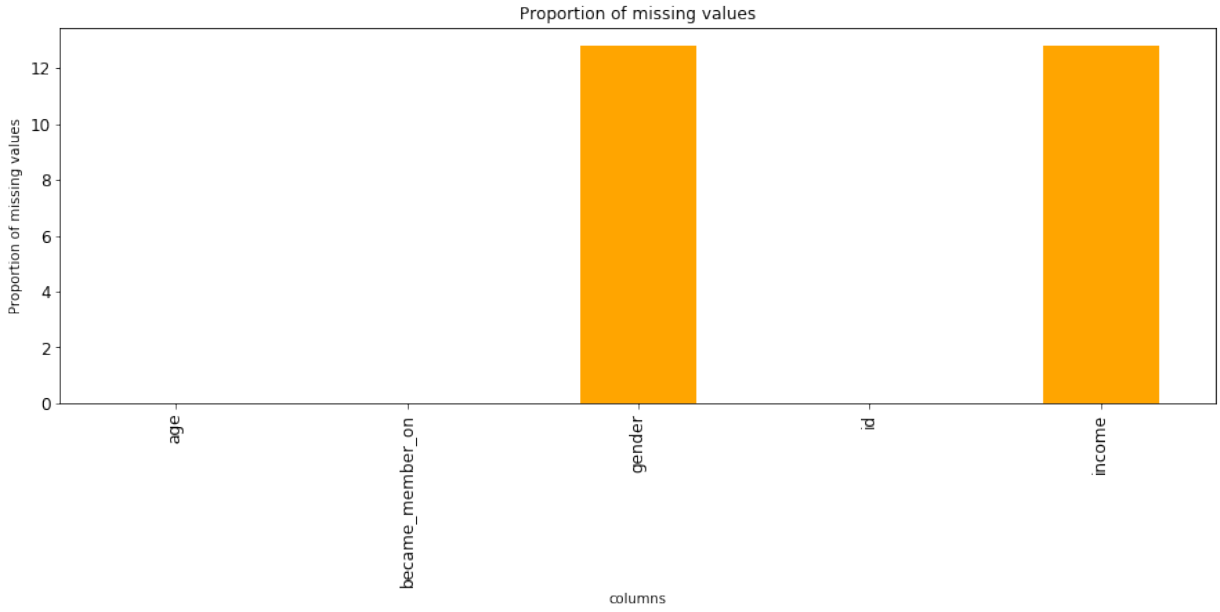
transcript.json

- event (str) - record description (ie transaction, offer received, offer viewed, etc.)
- person (str) - customer id
- time (int) - time in hours since 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

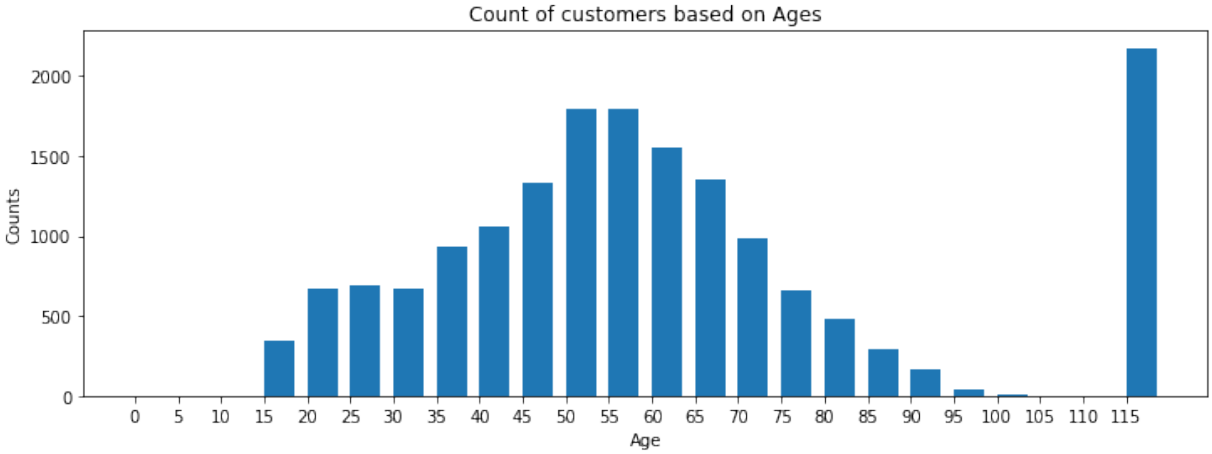
Data Exploration

We start explore the data from json files and check if we have any missing or outliers data:

- The portfolio file and from visulizion is not contain any missing data so no need to fix.
- The profile file and after we explore we find outliers and missing data
  Below the most column have missing data



And after we explore age we can see the age is over 115 years and it is not make sense the very old people visit Starbucks store a lot it could be bad data entry

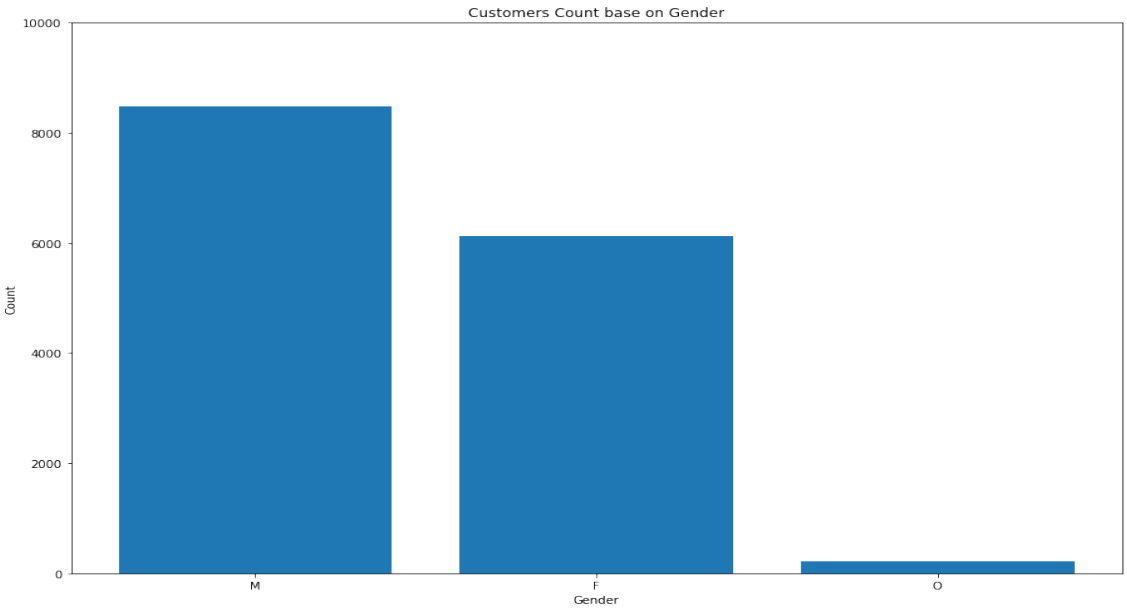


we can see there is customers above 115 years more than 2000 customer and it is highest then other ages which is not make sense so we can see this is defiantly outlier so we must delete or replace it with zeros or null for those customers.

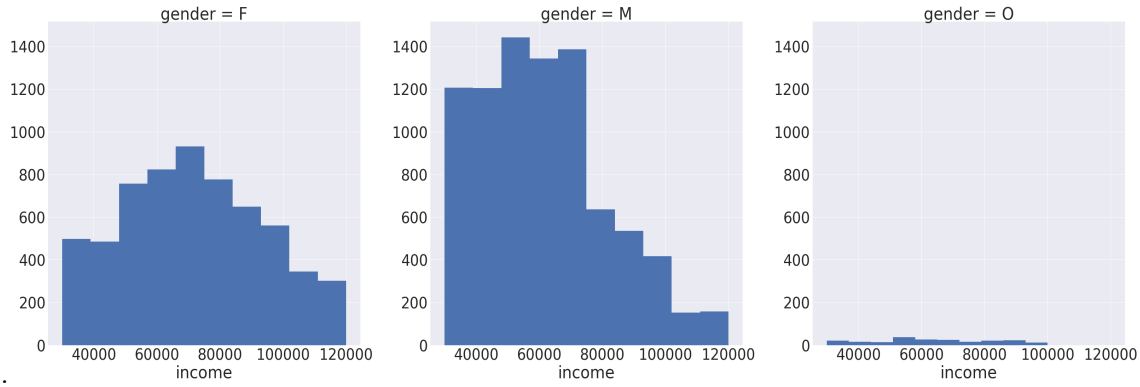
- Transcript not contain missing data after explore the file.

Exploratory Visualization

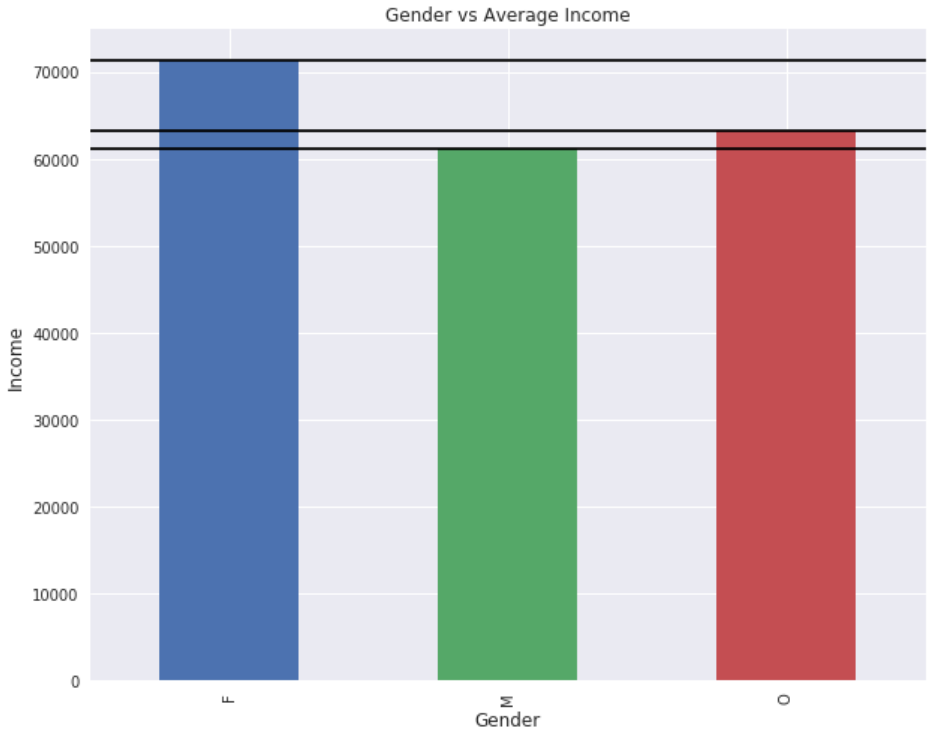
The plot below shows comparison male and female customers and is shows the male customers are larger than female customer.



A Next plot is answer the question (Based on income of customers who have highest income?)



.



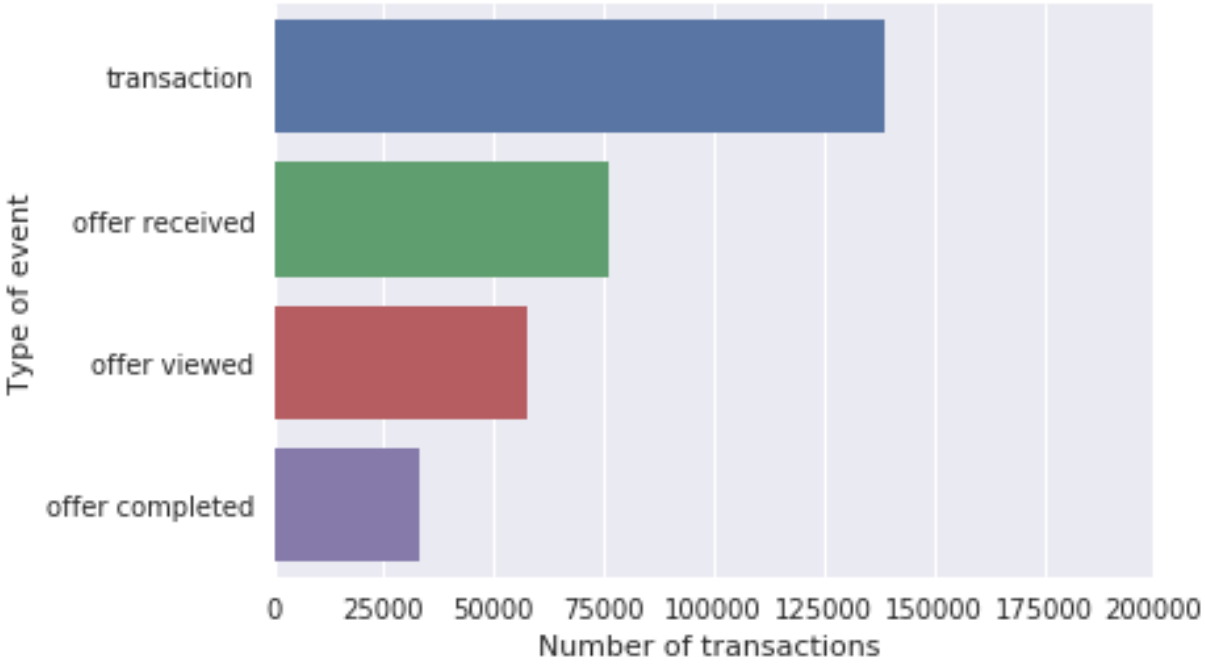
It can be seen that:

- ❖ The average of income for female is over 70000 while the male less than 60000

The following plot shows the register customers based on gender over years

We can see the number of customers are increased dramatically over years while the last years was decreed and most of years we can see the male customers are usually more than female except 2016

The plot below shows the event that send it to customers and measure the counts between the events we can see most event are transactions and the complete event is almost 50 % of received even.



Algorithms and Techniques

The classifiers used Logistic Regression, Naive Forecast and Random Forest , I merge all dataset together and training data set to check a successful transaction and check the accuracy of each model based on score and F- score,

Conclusion

My goal of the project was to analyze customer data and answer a number of questions that I was wondering, such as knowing the average age of customers, the number of customers over the years and comparing the sex of customers as well as knowing who are the customers who will respond to their offers from Starbucks so I built three The models were (Logistic Regression, Naive Forecast and Random Forest) and I train the data that was merged together after cleaning the data and removing the missing data and repeated data and check the outliers of data, then I evaluated the models and find out the most accurate model among the three models and next is the results of these models:

1. Random Forest

accuracy: 0.976 f1-score: 0.093

2. Logistic Regression Mode

accuracy: 0.975 f1-score: 0.000

3. Naive Forecast Models

accuracy: 0.025 f1-score: 0.049 Obviously, the model was the most accurate Random Forest and I was improve the model by increasing depth of tree