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 offer 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 my strategy for solving this problem I will prepper the data so first start to cleaning the data by removing missing and duplication or imputing the missing value if it is not effect on result also encode some of features and before training my model I will be combining the data together and train the model I would analyze the data to extract some information from that data.

So, based on the model results I am going to evaluate the accuracy of the model by checking the F-score of models F-measure (the $\beta=1$ case of the more general measure), is a weighted harmonic mean of Recall & Precision (R & P). There are several motivations for this choice of mean. In particular, the harmonic mean is commonly appropriate when averaging rates or frequencies.
So the I used the F score because I need to know how much accuracy could the model give for label variable or for result that I looking 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.

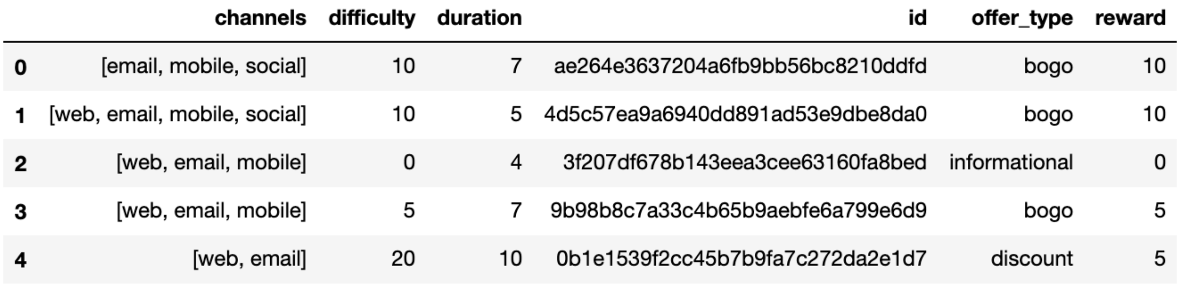
Analysis

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

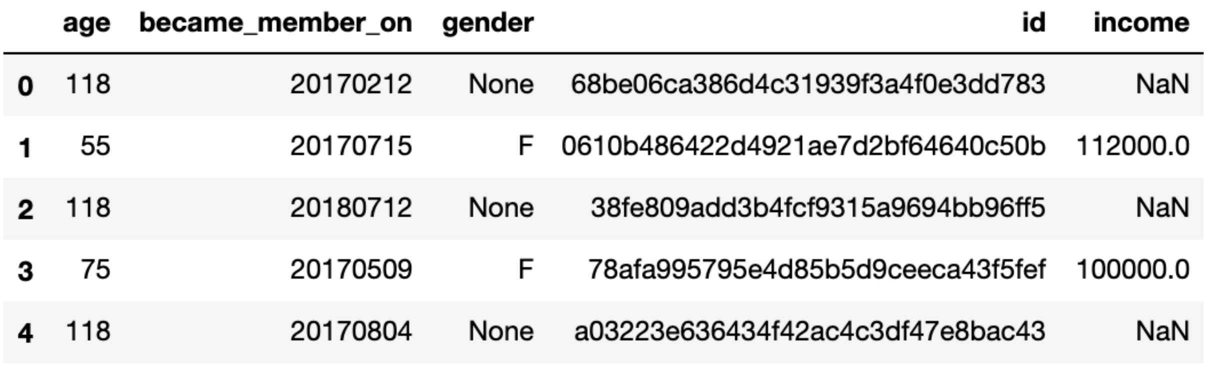
- id (string) - offer id
- offer_type (string) - type of offer ie BOGO, discount, informational
- difficulty (int) - minimum required spend to complete an offer
- reward (int) - reward given for completing an offer
- duration (int) - time for offer to be open, in days
- channels (list of strings)



	channels	difficulty	duration	id	offer_type	reward
0	[email, mobile, social]	10	7	ae264e3637204a6fb9bb56bc8210ddfd	bogo	10
1	[web, email, mobile, social]	10	5	4d5c57ea9a6940dd891ad53e9dbe8da0	bogo	10
2	[web, email, mobile]	0	4	3f207df678b143eea3cee63160fa8bed	informational	0
3	[web, email, mobile]	5	7	9b98b8c7a33c4b65b9aebfe6a799e6d9	bogo	5
4	[web, email]	20	10	0b1e1539f2cc45b7b9fa7c272da2e1d7	discount	5

profile.json

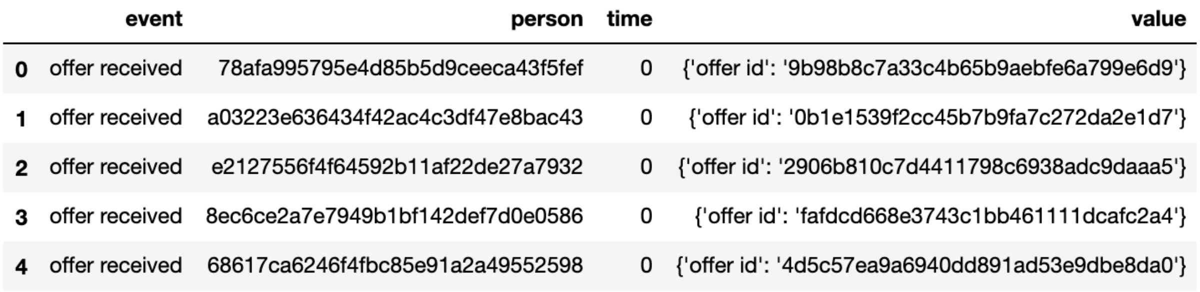
- age (int) - age of the customer
- became_member_on (int) - date when customer created an app account
- gender (str) - gender of the customer (note some entries contain 'O' for other rather than M or F)
- id (str) - customer id
- income (float) - customer's income



	age	became_member_on	gender	id	income
0	118	20170212	None	68be06ca386d4c31939f3a4f0e3dd783	NaN
1	55	20170715	F	0610b486422d4921ae7d2bf64640c50b	112000.0
2	118	20180712	None	38fe809add3b4fcf9315a9694bb96ff5	NaN
3	75	20170509	F	78afa995795e4d85b5d9ceeca43f5fef	100000.0
4	118	20170804	None	a03223e636434f42ac4c3df47e8bac43	NaN

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

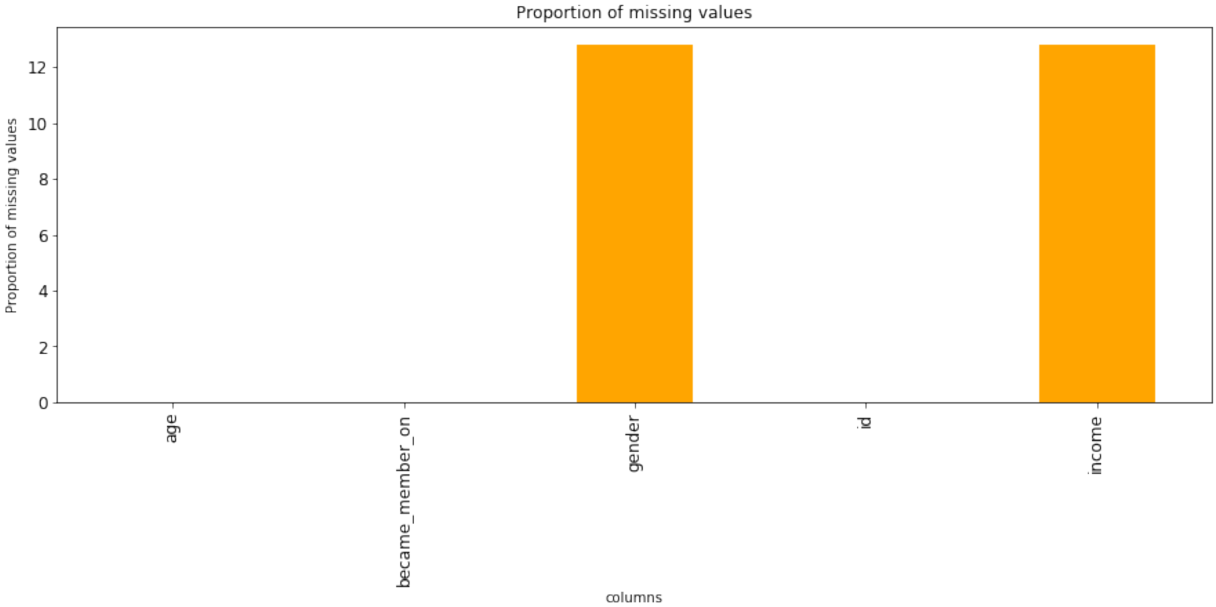


	event	person	time	value
0	offer received	78afa995795e4d85b5d9ceeca43f5fef	0	{'offer id': '9b98b8c7a33c4b65b9aebfe6a799e6d9'}
1	offer received	a03223e636434f42ac4c3df47e8bac43	0	{'offer id': '0b1e1539f2cc45b7b9fa7c272da2e1d7'}
2	offer received	e2127556f4f64592b11af22de27a7932	0	{'offer id': '2906b810c7d4411798c6938adc9daaa5'}
3	offer received	8ec6ce2a7e7949b1bf142def7d0e0586	0	{'offer id': 'fafdcd668e3743c1bb461111dcafc2a4'}
4	offer received	68617ca6246f4fbc85e91a2a49552598	0	{'offer id': '4d5c57ea9a6940dd891ad53e9dbe8da0'}

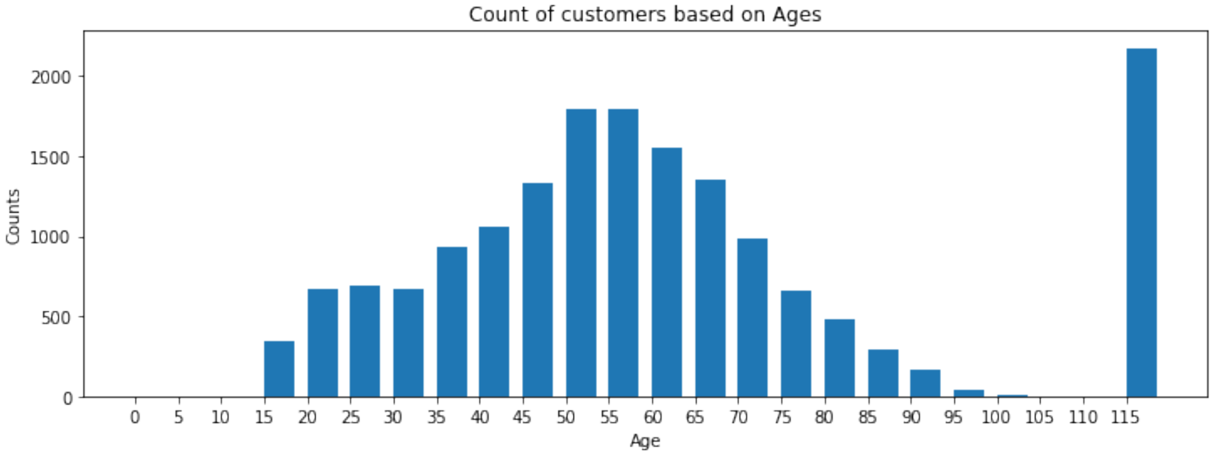
Data Exploration

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

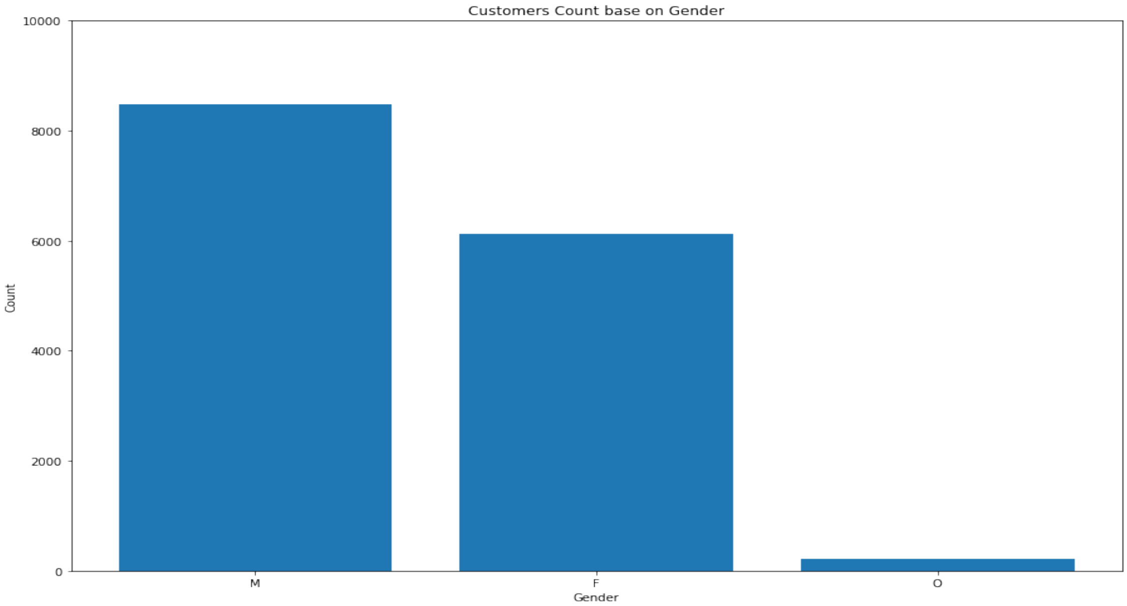


we can see there is customers above 115 years more than 2000 customer and it is highest then other ages which is not make sense so we can see this is defiantly outlier so we must delete or replace it with zeros or null for those customers.

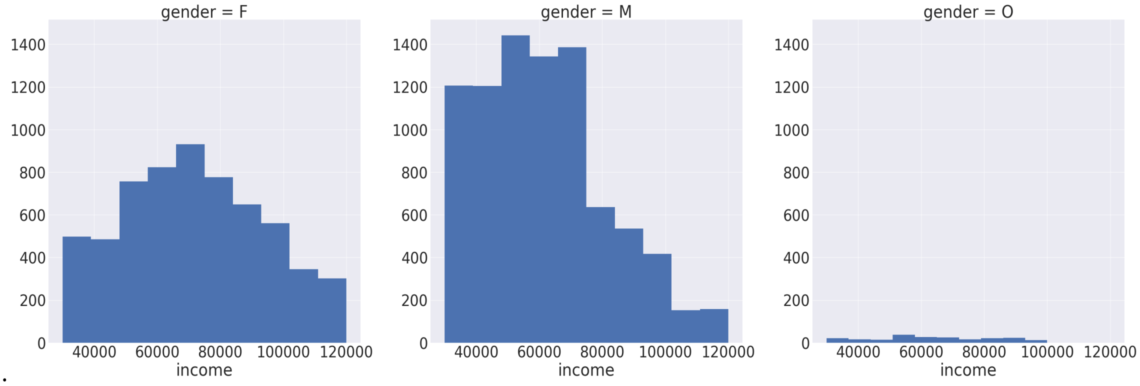
- Transcript not contain missing data after explore the file.

Exploratory Visualization

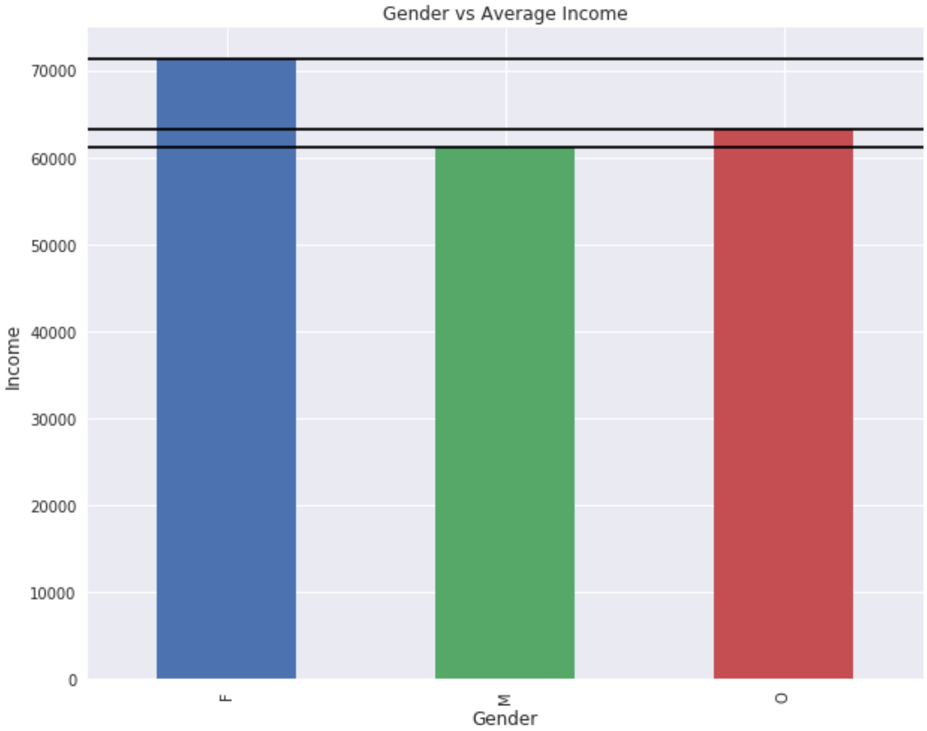
The plot below shows comparison male and female customers and is shows the male customers are larger than female customer.



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



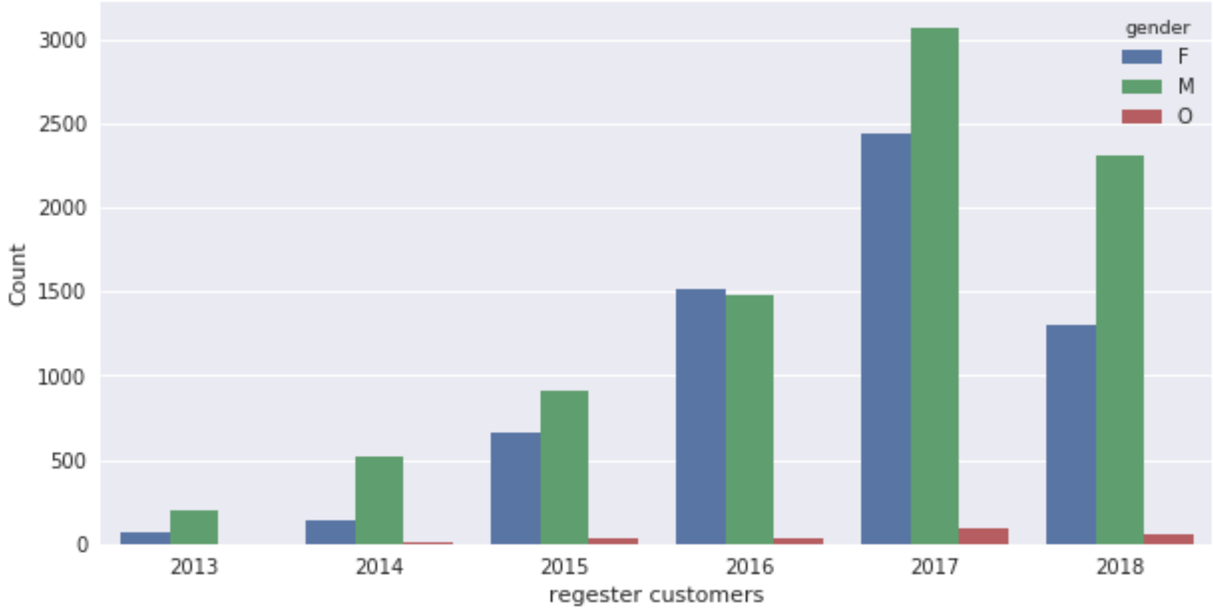
.



It can be seen that:

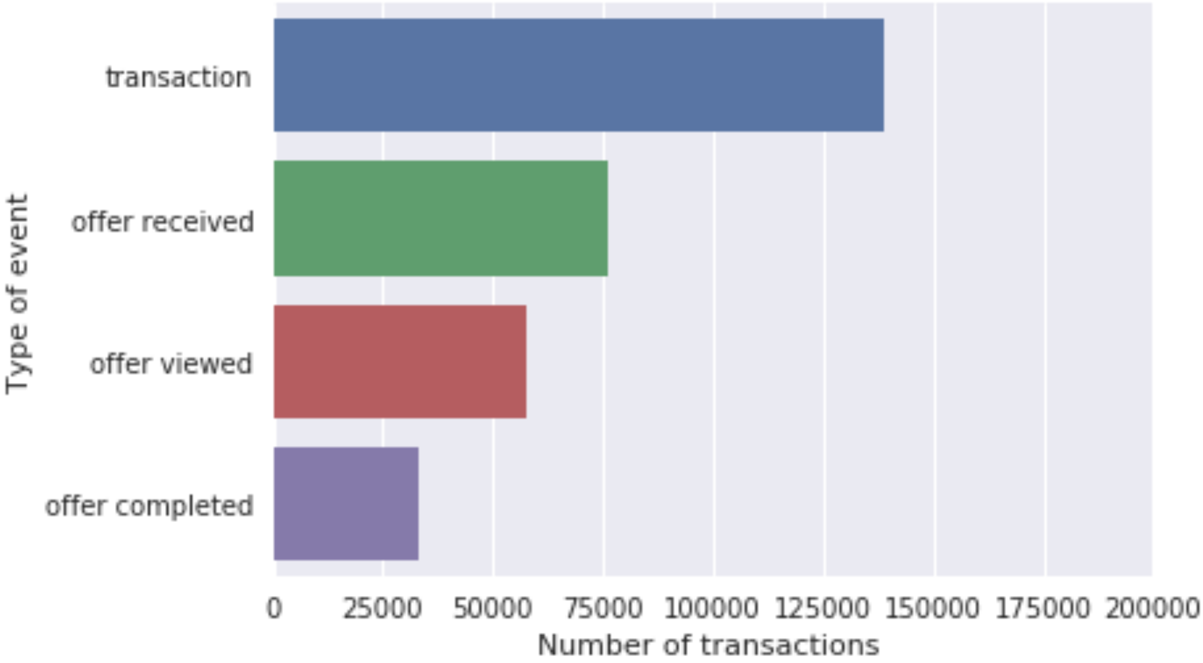
- The average of income for female is over 70000 while the male less than 60000

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

1- We start by peppering the data by removing missing data and duplication so to find the missing data

- I use some of visualizations that can be help me to find out the missing values using matblot library.
- I find some data entry mistakes regarding age variable so replace it with null values.
- Remove the missing data of gender and income variable.
- Covert and spilt the channel column to many columns and encode the values based on type of channel for this column 0 if it is not value and 1 if it is.
- Convert the became_member_on column to date it was string to make new column call year.
- In transcript data frame I extract the data from value column and spilt the value in separate column using panda library functions.
- Spilt or pivot 'event' column to many columns
- Remove duplication of data transcript data frame
- Drop all any column that not helping to training model or not useful in model section
- Finally start to combining all data files together based on person id also create new column call total to calculate the duration of the offers

2- Answering some questions after combining data by exploring and visualization data.

3- Import the library for chosen models

```
from sklearn.ensemble import GradientBoostingClassifier, RandomForestClassifier
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score, f1_score
from sklearn.metrics import fbeta_score, make_scorer
from sklearn.model_selection import GridSearchCV, RandomizedSearchCV
```

4- Use useful columns in final dataset and spilt training set and test set and start training model

- Random Forest Classifier
- Logistic Regression

5- Evalute the model by finding accuracy score to measure the performance of model and F is the harmonic mean of recall and precision, with a higher score as a better model.

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

Refinement

After we training the models I try to make some model is better so I was using my local machine rather than Jupiter notebook workstation included in with project because some of models was take long time to training them like the random forest model I tray to enhance the model by increase the max_depth and that make the model is better than first training the f- score was increasing dramatically witch it was give more confidant with results lately. Also, I try to enhance logistic regression by increase some property but I could get better the final result of that.

Improvement

After we training the models I try to make some model is better so I was using my local machine rather than Jupiter notebook workstation included in with project because some of models was take long time to training them like the random forest model I tray to enhance the model by increase the max_depth and that make the model is better than first training the f- score was increasing dramatically witch it was give more confidant with results lately. Also, I try to enhance logistic regression by increase some property but I could get better the final result of that.

Results

As we See we apply the the Random Forest and Logistic regression on combining dataset for starbuckes company and it was contain some data about clients and offers our goal was build model to predict if the customer will respond to the offer or not so we start training our data using our models and evaluates the model we find the accuracy of random forest was good and the accuracy and performance are better compared to logistic regression also in I use classification_report to compare the models and it was model robust enough to be trusted regarding the result of evaluation.

In first section I was wandring and have some questions like what is the most gender of Starbucks clients? And we find the male was more then other gender also whom have heist average of income based on gender ? and the answer was Female and I asked about 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.

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 this result is representing the scores of datasets without applying algorithms.

Obviously, the model was the most accurate Random Forest and I was improving the model by increasing depth of tree

I was surprised with result of models and each time I try to improve the model I got different results some time good result and some time getting bad result the most difficult it was running time of models some time spend more time as usual which is depressing also I found little difficult to calculating the time for each offer.

Finally to improving the model as we learn we could adding more data to dataset and this could improve the model and if we training more data there is chance to increase the accuracy beside as I mention early tuning the parameter of the algorithm these parameters majorly influence the outcome of learning process.