Who can get a Starbucks coupon?

Introduction

In this project, I analyzed simulated Starbucks customers' spending behavior with different offers provided. I analyzed customer clustering using demographics data and all data respectively. In specific, I set out to answer the following questions:

1) What are the customer clusters?

2) What are the responses of different clusters to different offers?

3) For customers with incomplete information, what cluster do they belong to?

4) What are the potential business suggestions that could be made?

Data Cleaning and Exploratory Data Analysis

There are two main features about this dataset:

1. There are a substantial amount of missing values where age is 118, income and gender is NA:

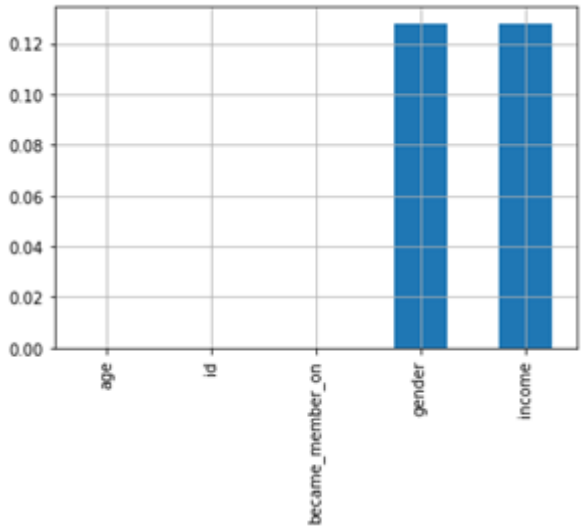


Figure 1. Missing value distribution

2) We can see that age 118 is an anomaly based on the following graphs. Where the original age histogram shows a concentration for over 200 years. By manually checking the observations with age over 100, I find that a great portion of customers' information is problematic: they have null values for income and gender . They also have 118 for their age and None for their genders. Therefore, we don't have much demographic information about them.

I think these customers represent those who shopped but never bother to complete her profile information and we only know when they came to be a member.

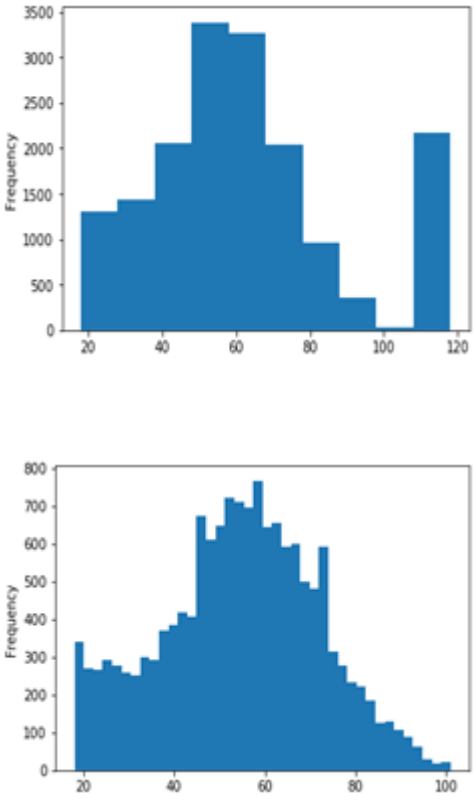


Figure 2. Age distribution for all and those with age not equal to 118

After I take out those aged 118 from the customers, I obtained an age distribution that looks much more normal (graph on the right in Figure 2)

3) Another problem that we have is that we only have customer behavior data one after another and we need to aggregate them into treatment and result of each offer. To tackle this problem, I first group the customer activities by the person, then I took the events where the customer actually viewed and completed offers before the expiration date as successful. As for informational offers, as long as a single transaction happens during its effective period, it's counted as a success.

Modeling Customer Clusters

As a starting point, I applied the K-means algorithm to data not related to offers. And I applied it to customers with valid demographic information. I conducted standard pre-processing like

transforming joining date into the relative length of joining experience, changing the gender variable into a categorical variable, and standardize all variables. The process for choosing the number of clusters is shown in the following graph. I picked K = 6 as the 'elbow' point.

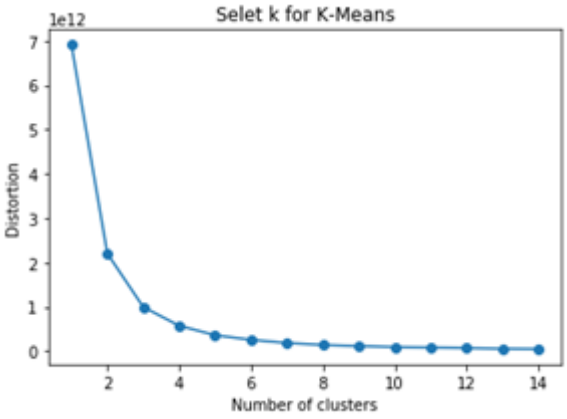


Figure 3. Choosing cluster number

Then I performed K-means and visualized it using t-SNE algorithm:

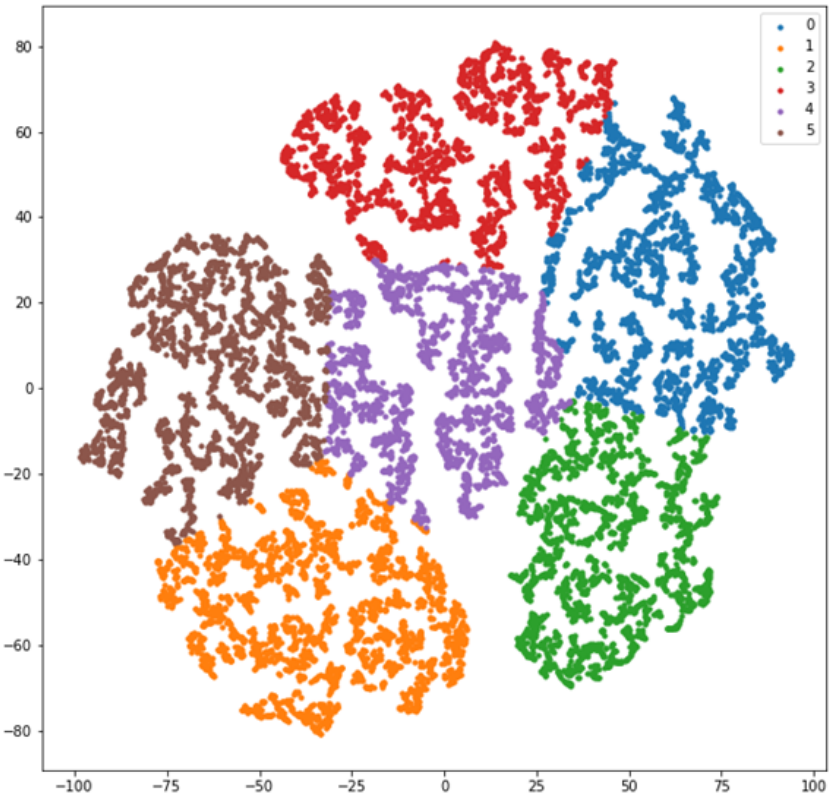


Figure 4. Visualizing K-means using t-SNE (n components = 2)

Then I compared the average features including transaction and offer-related features for different clusters obtained here. The result can be fond in the notebook and not displayed here. They are in general slightly different from each other. However, when we analyze the impact of different types of offers, they show quite different behavior. Which can be seen from Figure 5.

These results could be used for answering business problems like how to distribute the offers. Take cluster 2 for example, we can conclude that BOGO 0/2, discount 0/3, information 0 are more effective and should be used more compared to offers of the same type.

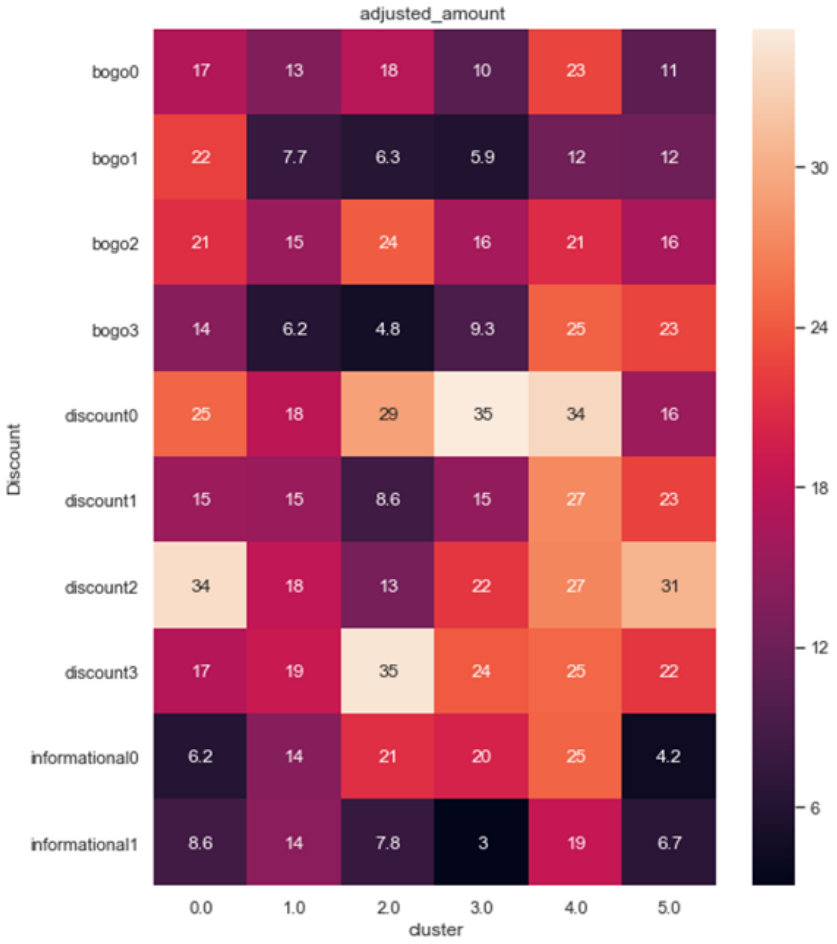


Figure 5. Response to different types of offers: number is the coefficient of OLS regression Coefficient for different groups.

(Notice: demographic variables including age, gender, experience, income are also included as control variables but their coefficients

are not shown here. Further, I adjusted the amount spent by the offer's reward to reflect the true amount spent and earned)

Since in real business, with the accumulation of customers' data, we can use customers' behavior when analyzing their clusters, I further used all features available for the clustering. I performed K-means then t-SNE on all demographics feature and I transformed behavior features into rate of success instead of raw counts to increase comparability. We see that K-means missed some structure inside these data and is probably why it has no obvious elbow point.

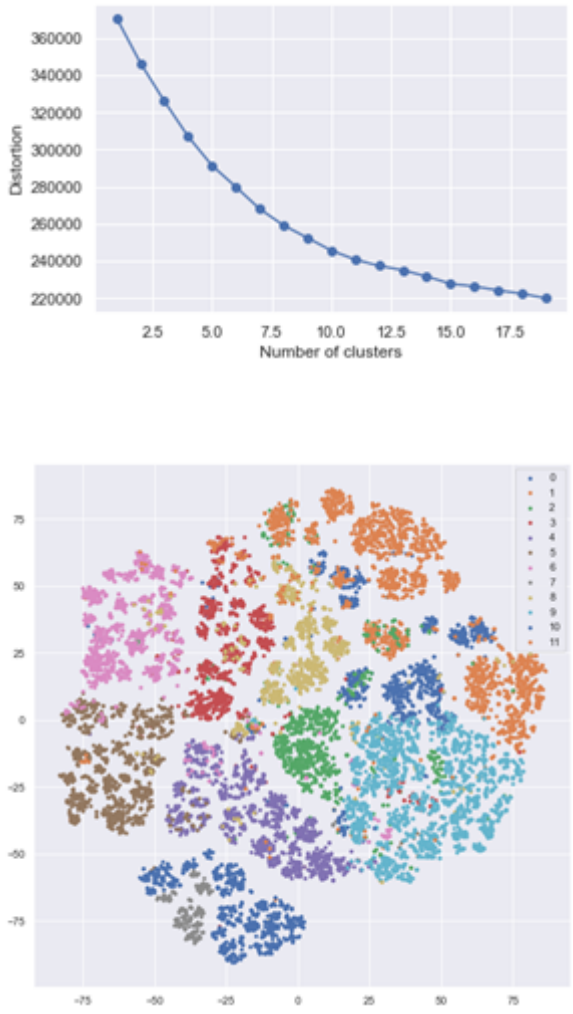


Figure 6. Visualization of K-means clustering using t-SNE

Improvement:

To tackle the problem, I first used t-SNE and then applied K-means on the 2-dimensional data. I acquired the following results:

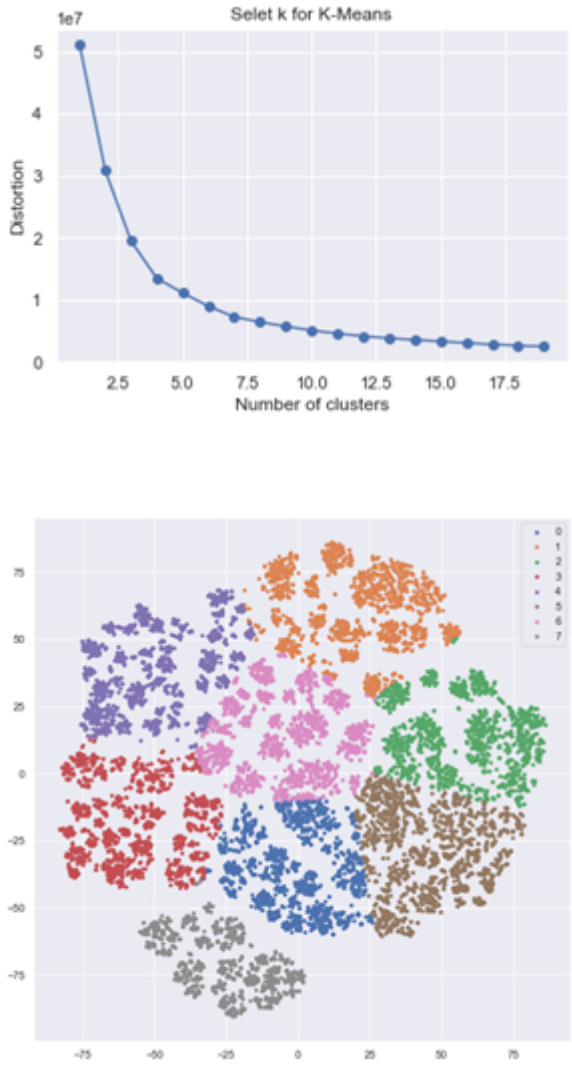


Figure 7. Visualization of K-means clustering after using t-SNE

Clearly, we achieved a better result with fewer clusters. And I build upon this result to analyze the missing value group. In specific, I train a random forest classifier using the labels acquired and the features available for both groups on the no-missing value data. I also used cross-validation to enhance the model's performance.

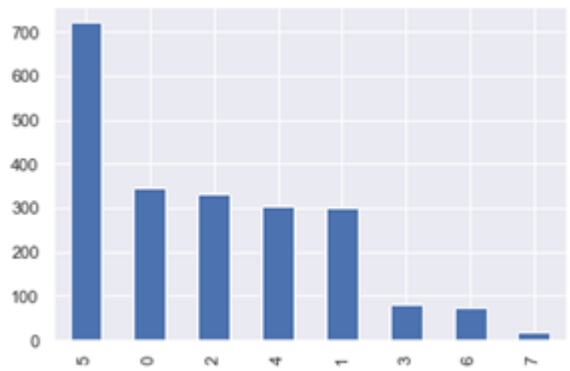
I obtained an F-1 score of over 0.995 (either micro or macro averaged). Then I used the model to predict the labels of missing value group customers. The result is as the following:

We can see that our model discovered some structure hidden beneath that is difficult to see using simply amount and experience.

I further plotted the distribution of different clusters in the two groups of customers. The result show that they are essentially

different:

Figure 8. Missing value clustering using t-SNE and raw features (amount-experience)



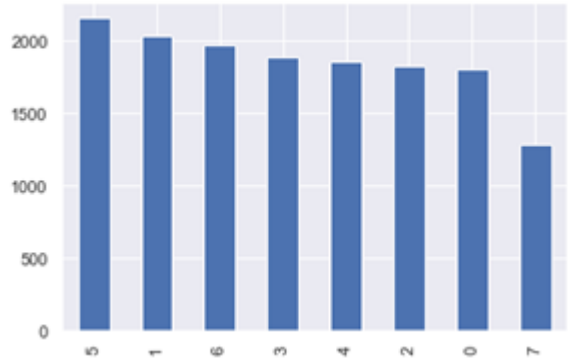


Figure 9. Distribution of clusters: missing value group (up) and no missing value group (down)

Finally, I run a regression for each of these groups to see how they respond to offer types. Since we already used customer behavior data, so, **the result is not causal but a measure.**

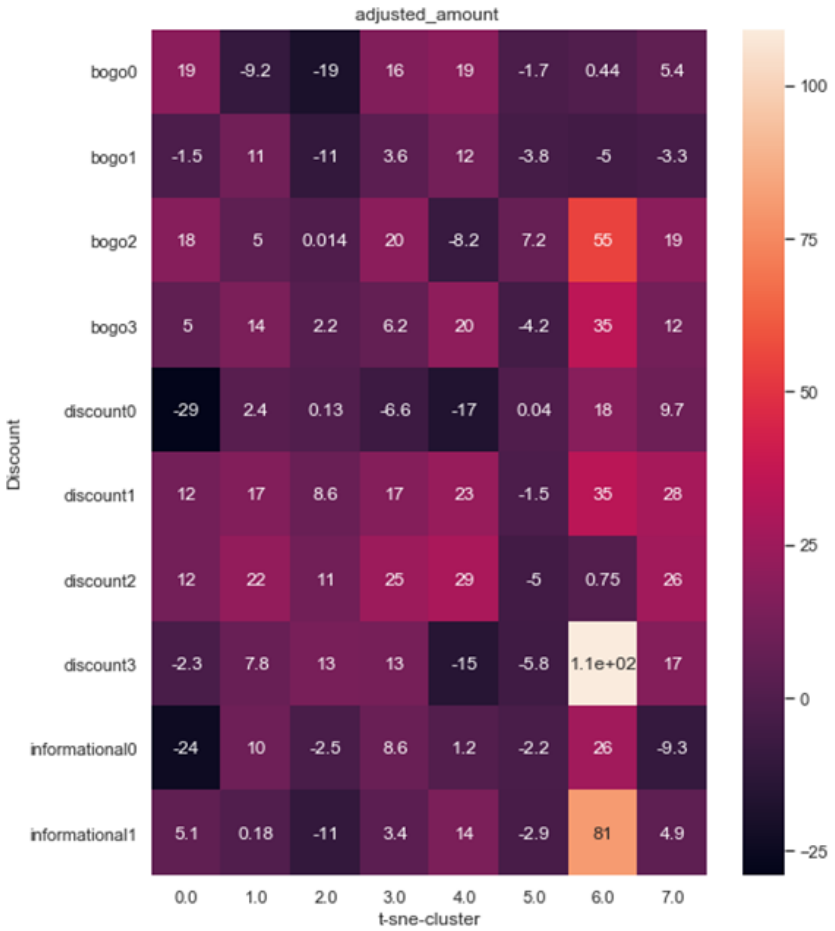


Figure 10. Response to different offer types by clusters.

Conclusion and Reflection

- Conclusion

a. Using demographics data, we can divide customers into 6 clusters. We find that they show different responses to different offer types.

b. We can achieve better results using these response measures.

c. Using customer behavior data and random forest, we can get a more refined clustering of customers and infer cluster structure of missing value group. We find the composition is indeed very different

- Reflection

a. Could perform A/B test in addition to regression if we have more data. Though the grouping itself already controlled the demographics problem among users.

b. Could simulate the business setting: divide the dataset into different time periods and test how much more revenue is generated.