



Data Glacier

Your Deep Learning Partner

Model Deployment

G2M Case Study

21-March-2023

Agenda

Executive Summary

Feature Selection

Model Summary

Dataset

Model Deployment

Summary

Recommendations

Executive Summary

In EDA phase, I found some key factors to invest cab company. It include advantage of place, the number of users, and seasonal trends. For further analysis, I deployed clustering model for customer. By deploying this model, I can detect primary customer that cab company has to focus on.

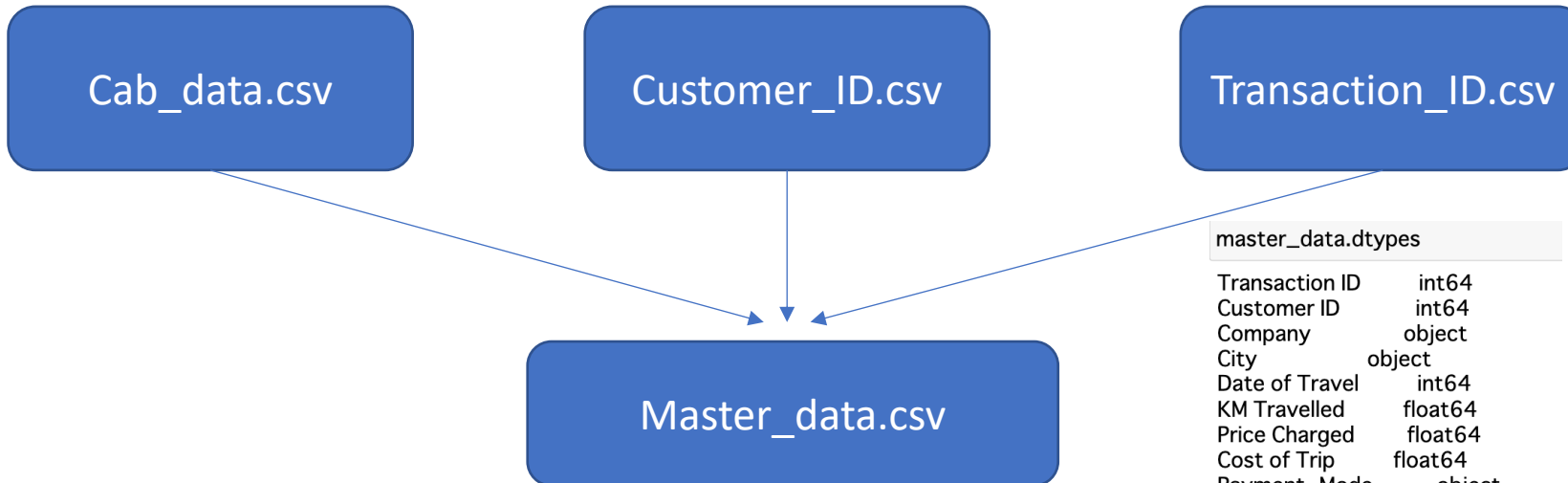
Dataset

Original datasets

- Cab_data.csv
- Customer_ID.csv
- Transaction_ID.csv
- City.csv

Processed Datasets

- Master_data.csv
- City.csv



To conduct analysis, merge three datasets with IDs and named it master data.

master_data.dtypes

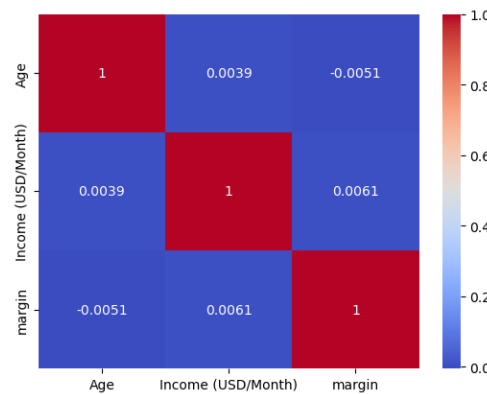
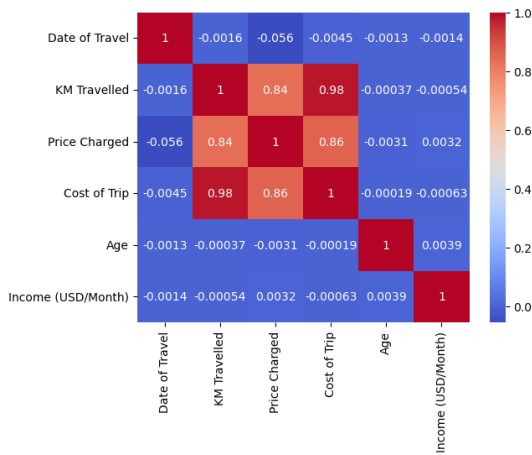
Transaction ID	int64
Customer ID	int64
Company	object
City	object
Date of Travel	int64
KM Travelled	float64
Price Charged	float64
Cost of Trip	float64
Payment_Mode	object
Gender	object
Age	int64
Income (USD/Month)	int64
dtype:	object

Feature Selection

Check correlation with heatmap

	KM Travelled	Price Charged	Cost of Trip	Payment_Mode	Gender	Age	Income (USD/Month)
0	30.45	370.95	313.635	Card	Male	28	10813
1	28.62	358.52	334.854	Card	Male	27	9237
2	9.04	125.20	97.632	Cash	Male	53	11242
3	33.17	377.40	351.602	Cash	Male	23	23327
4	8.73	114.62	97.776	Card	Male	33	8536

	Age	Income (USD/Month)	margin
0	28	10813	57.315
1	27	9237	23.666
2	53	11242	27.568
3	23	23327	25.798
4	33	8536	16.844

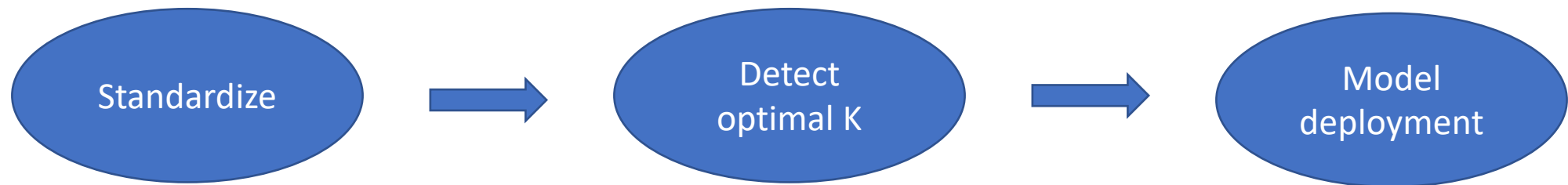


Multicollinearity may violate clustering model. To avoid this, Put Price charged and Cost of Trip together as margin, and remove KM Travelled. Payment Mode, Gender, are not used in this model.

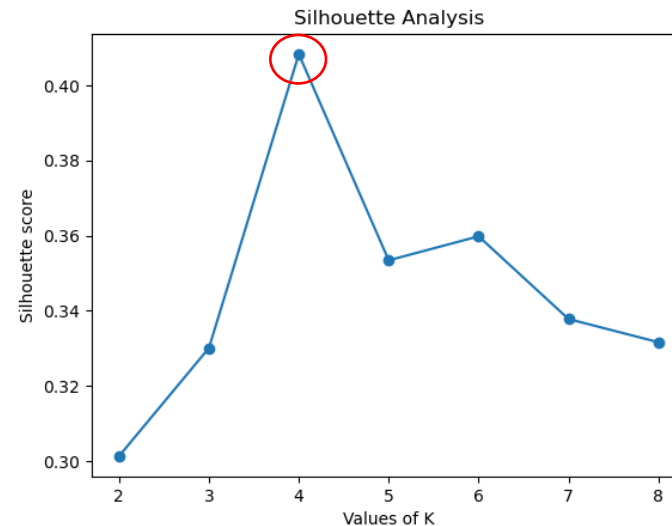
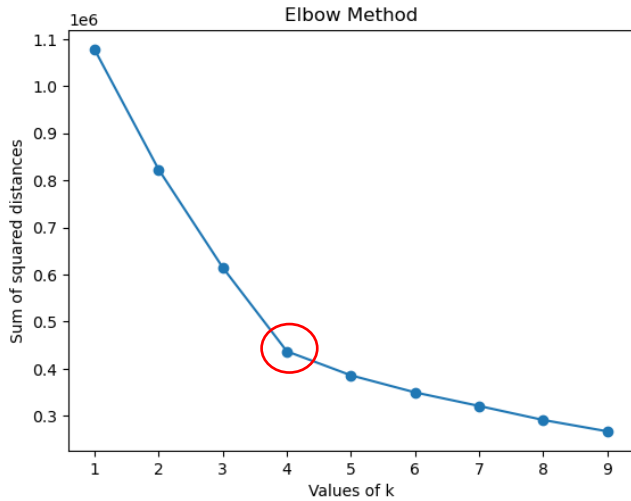
Model Summary

KMeans

- Categorize each data for into K group by calculating distance of each other. Each features need to be standardized to match scale.
- Kmeans need to find optimal K which means how many groups are desirable. I used some methods to detect optimal K.
 1. Elbow method
 2. Shilhouette method



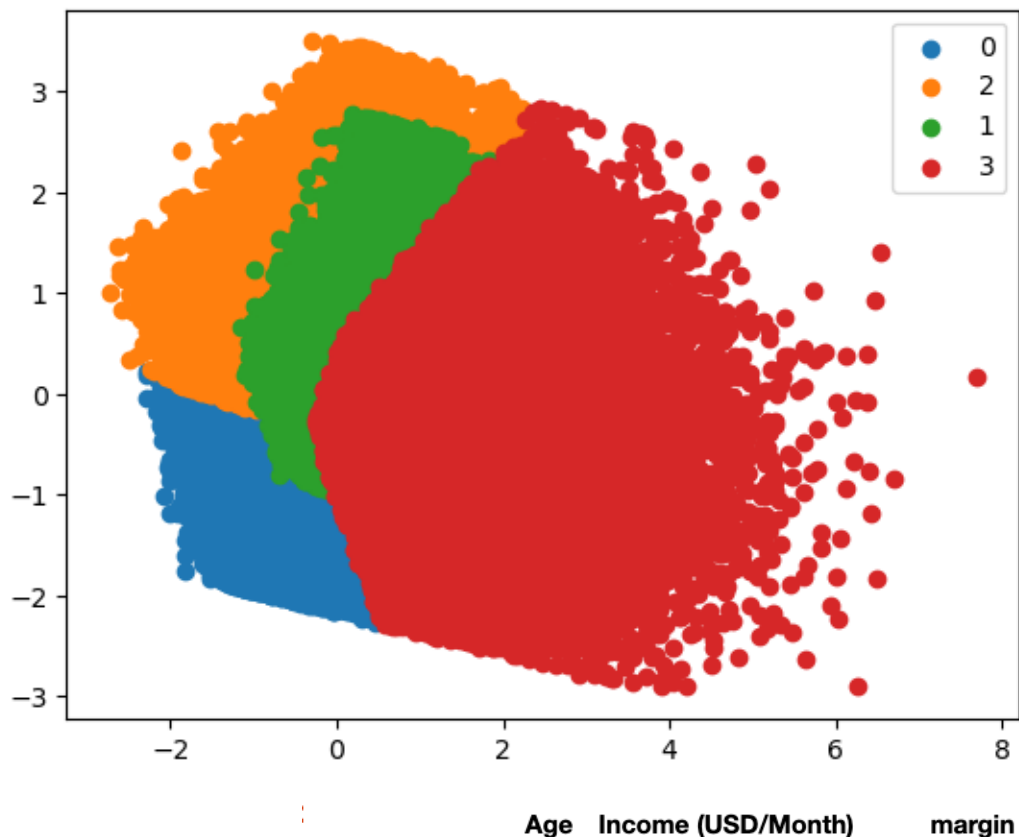
Model Summary



- Elbow method
Optimal K is defined by the point which most reduce sum of squared distances.
Optimal K is 4 according to Elbow method.
- Silhouette method
Silhouette score is between -1 and 1. 1 is best, meaning the K is most optimal. The point that values of K is 4 maximized silhouette score.

4 clusters is optimal in this model.

Model Deployment



cluster			
0	29.499285	8621.161407	85.654907
1	29.880440	22658.053515	88.337003
2	54.842137	14855.727309	98.554194
3	33.902115	14980.631554	479.921335

Customers were separated as follows,

- Cluster 0
Mean of Age is 29. Income and margin are lowest.
- Cluster 1
Mean of Age is same as cluster 0. However, this group has highest income but they don't use cab much.
- Cluster 2
Mean of Age is 54. Income is third highest and margin is second highest.
- Cluster 3
Mean of Age is 33. Income is second highest. They use cab frequently.

Result Summary

- There are two types of customer in mean of 29. One is lowest income, the other is higher income. Both customer types don't use cab much.
- Elderly people use cab more than group of 29 years old.
- The primary customer for cab is middle senior. Cab company can get more than 4 times as much margin as the other groups.

Recommendations

- People which are average 33 years old are primary users of cab so We can get more profit by taking an action to these customer.
- We may need to start campaign for people who are young and have income. By doing so, We can get more profit.

Appendix

Code for this analysis:

<https://github.com/KHUC1998/G2M-market/blob/main/Untitled.ipynb>

programming language:

Python

Used software:

Jupyter notebook Tableau

Thank You