Project Deliverable

11/20/2019

Cohort B Team 7

Kunpeng Huang, Yoki Liu, Lyufan Pan, Yunlei Zhou, Jiayuan Zou, Sherry Zuo

Describe Questions:

The problem that motivates us is the segmentation of customers that could be very necessary and powerful to define marketing strategies. The credit card usage behavior of customers with 17 behavioral features seems to be a perfect dataset for us to explore the customer segmentation based on their purchasing behaviors with credit cards.

Clean the dataset

a. Check and replace missing values in the dataset
   Replace 313 NAs in *minimum_payments* with colMeans()
b. Adjust different value types
   Remove character variable *cust_id*, so the remaining 17 variables are all numeric.
c. Chose meaningful variables for analysis
d. Detect abnormal values and manipulate

Review of Data - EDA

(0 = not very frequently, 1 = very frequently)

In the cleaned dataset, there are 17 numeric variables and 8950 observations. From the plot and statistic table of variables, we found that the *balance* of majority users are below 5000 with an average balance of $1564.47; most customers are active and have a *balance frequency* around 1; For the *purchase frequency* with credit cards, the graph shows us that the purchasing polarization. Most credit cards record is concentrated on either not very frequently or very frequently group.

Visualization

a. Correlation Plot
   Based on the correlation plot colors, we think there are more than 4 clusters in our dataset.
b. Box Plot
   Focus on variables balance, balance_frequency, purchases, purchases_frequency, oneoff_purchases, oneoff_purchases_frequency, installments_purchases.
c. Histogram
d. Scatter Plot

Algorithms

summary(lm(credit_limit~. , cc))

We consider the variable credit_limit could be as y since this variable where the information is the bank offered

Baseline clustering

- Silhouette score-choose cluster is 13(7 and 9 is also very high)

- WSS-choose cluster is 2, 4, 7, 9
  Compare those cluster plots, we think k=9 is the best since each cluster has a similar size in the baseline clustering.

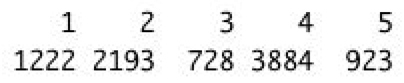
Dimension Reduction (PCA model)

Compare **Eigenvalue** and **Cumulative Variance**

Based on the eigenvalue, we want to choose eigenvalue>1, so we could choose Dimension with 5; however, since we want cumulative variance too small, we choose eigenvalue>0.7, so we choose **Dimension with 8** which also has 85% of cumulative variance.

Clustering for PCA model

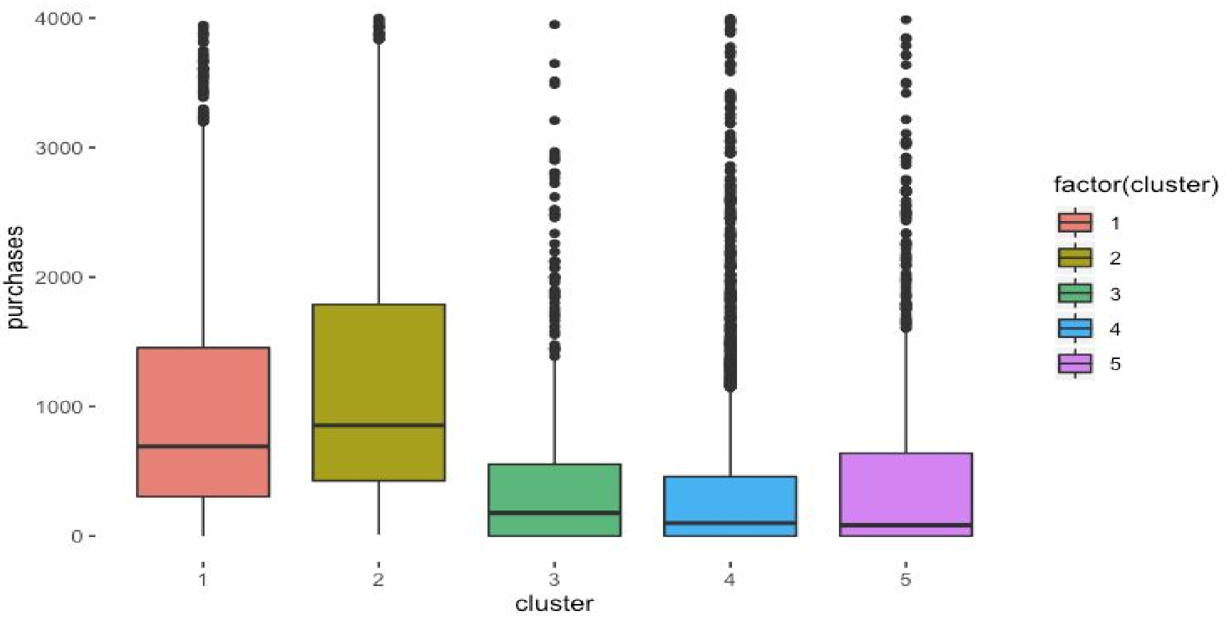
- Silhouette score-choose cluster is 2
- WSS-choose cluster is 2, 5, 7, 9
  Since we don't want the size of the cluster too small or larger and try to average the size of clusters, so we think the best k is 5 for the PCA model.
  Our best model overall:



```
   1    2    3    4    5
1222 2193  728 3884  923
```

Results business related

Add clustering back to the original dataset, think about the characteristics of each cluster. So far, we find cluster2 has relatively large purchases_frequency, which has the range of [0.8,1.0]. Those people could be those who are dependant on credit card using. Cluster5 has relatively large credit_limit and balance. These clusters of people could be those who have larger incomes. More characteristics of each cluster would be found later.


purchases
cluster
factor(cluster)