Credit Card Segmentation

Sanket Mote

11 February 2020

Table of Contents

Chapter 1

Introduction

1.1 Problem Statement

We have a credit card customers data wherein based on various variables we have to derive certain Key Performance Indicators (KPI) such as monthly average purchase and cash advance amount, purchases by type (one-off, instalments), average amount per purchase and cash advance transaction, limit usage (balance to credit limit ratio), payments to minimum payments ratio, etc. We also need to segment these customers based on our KPIs.

1.2 Data

We have dataset which has 18 variables and close to 8,950 observations using which we will derive the Key Performance Indicators and clusters based on customer transactions. Given below is a sample data set which we will be using for the clustering or segmentation of customers:

	CUST_ID	BALANCE	BALANCE_FREQUENCY	PURCHASES	ONEOFF_PURCHASES	INSTALLMENTS_PURCHASES	CASH_ADVANCE	PURCHASES_FREQUEI
0	C10001	40.900749	0.818182	95.40	0.00	95.4	0.000000	0.166
1	C10002	3202.467416	0.909091	0.00	0.00	0.0	6442.945483	0.000
2	C10003	2495.148862	1.000000	773.17	773.17	0.0	0.000000	1.000
3	C10004	1666.670542	0.636364	1499.00	1499.00	0.0	205.788017	0.083
4	C10005	817.714335	1.000000	16.00	16.00	0.0	0.000000	0.083

Table 1.1: Dataset (Observations: 1-5)

Chapter 2

Methodology

2.1 Pre Processing

Before starting clustering and derivation of KPIs, we need to check the data and understand various data definitions. Also, we need to check whether data is in proper format before processing the data, if data is not in proper format, we need to clean the data in order to increase the efficiency of the model. To start the process, we have basic steps which needs to be completed such as data cleaning, visualization using graph or plots, standardizing if data is not uniformly distributed these steps are a part of Exploratory Data Analysis. Below are the KPIs

1. **Monthly Average Purchase:** Purchases divided by Tenure will give us monthly average purchase value for each customer.
2. **Monthly Cash Advance Amount:** Cash advance divided by Tenure will give us monthly cash advance amount.
3. **Customer Purchase Habits:** We will divide the customers into groups based on their purchase habits such as one off purchase, installment purchase, both or none.
4. **Average Amount Per Purchase:** Purchases divided by purchases transactions will give us average amount per purchase of each customer.
5. **Average Cash Advance Transaction:** Cash advance divided cash advance transactions will give us average cash advance transaction for each customer.
6. **Limit Usage (Balance to Credit Limit Ratio):** Balance divided by credit limit will give us the limit usage at customer level.
7. **Payments to Minimum Payments Ratio:** Payments divided by minimum payments will give us the required ratio

2.1.1 Missing Value Analysis

We check whether dataset contains any missing value. If it contains missing value, we need to impute missing observations in a variable if it has at least 70% of data. We impute the missing data with the help of mean, median in case of continuous variable and mode in case of categorical variable. We can also use K Nearest Neighbor (KNN) in order to impute missing values. In our dataset we have missing values in Credit Limit and Minimum payments variables. We found that mean method is imputing closest value so we will proceed with Mean for Missing value Imputation. We can proceed to the next step.

2.1.2 Outlier Analysis

There are many variables in our dataset which contains outliers in it. After removing outliers using inter quartile range, we found that approximately 70% of data is getting dropped therefore we will consider log transformation to minimize the outlier effect. Since in removing outlier method we might loss some important data.

2.2 Modelling

2.2.1 Principal Component Analysis

We will start our modelling part by implementing Principal Component Analysis (PCA). We use PCA when we want to reduce the number of variables, ensure variables in the dataset are independent of each other. In PCA we are going to calculate matrix that summarizes how all variables relate to each another. Post this we will break this matrix into two separate components direction and magnitude. By projecting our data into a smaller space, we are reducing the dimensionality of our feature space but because we have transformed our data in these different “directions”, and we have made sure to keep all original variables in our model. From Correlation heatmap some of the variables have similar information and we need to remove them before generating clusters in order to decrease the complexity and biasness towards certain feature.

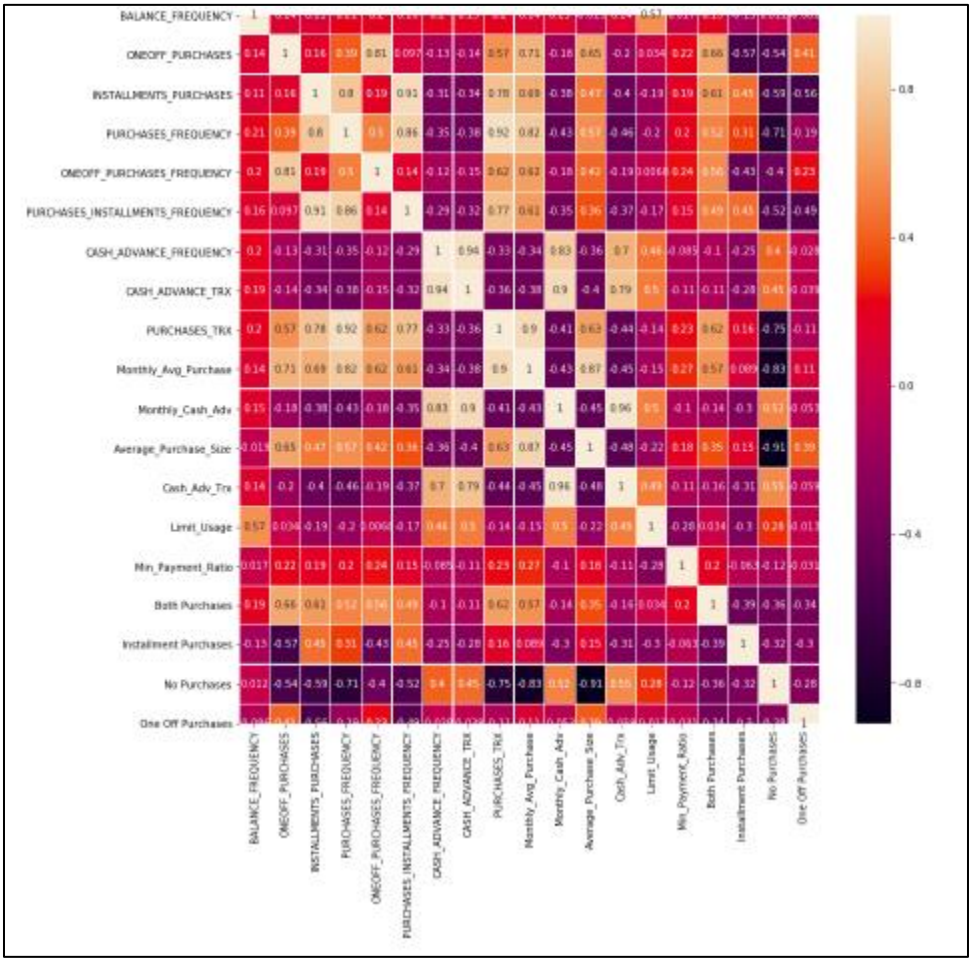

Fig 2.2.1 Correlation Heatmap

2.2.2 K-Means Clustering

In PCA once we reduce variables we proceed with clustering. Clustering is one of the most common exploratory data analysis technique used to get an intuition about the structure of the data. It can be defined as the task of identifying subgroups in the data such that data points in the same subgroup (cluster) are very similar while data points in different clusters are very different. Clustering analysis can be done based on features where we try to find subgroups of samples based on features or based on samples where we try to find subgroups of features based on samples. Since we had made four categories of customers above let's start with four clusters. Below is cluster formation plot.

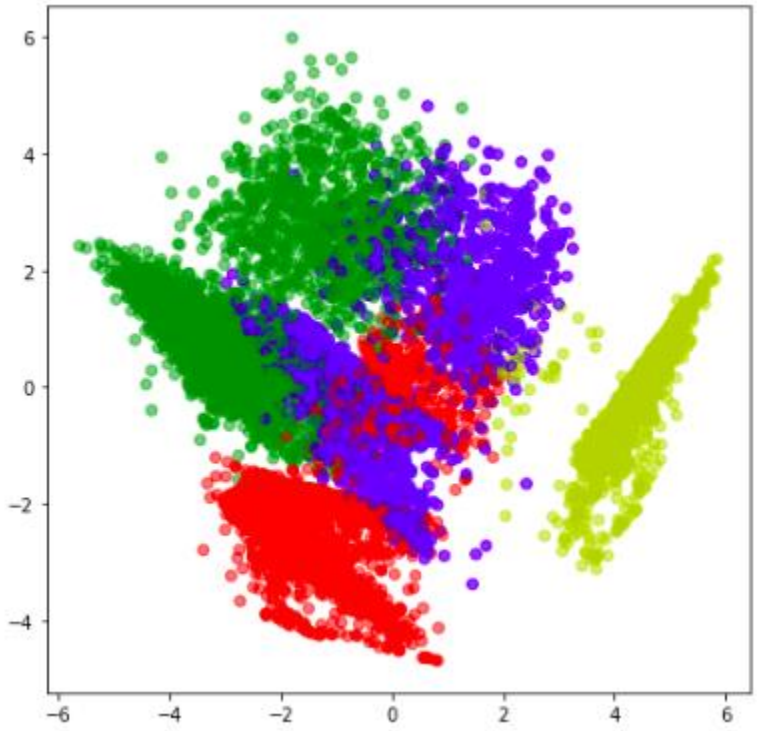

Fig 2.3.1 4 Clusters Scatter Plot

Below is the summary and Insights of clusters formed:

Cluster 0:

In first cluster customers are utilizing lowest credit limit amongst other clusters and they are mostly transacting in installment purchases.

Cluster 1:

This set of customers are mostly doing one off purchase followed with cash advance transactions.

Cluster 2:

This set of customers are doing maximum average purchases and are utilizing maximum credit limit. Basically, this are our high spending customers and are involved in both type of transactions one off and installments.

Cluster 3:

These customers are making cash advance transactions mostly and are not transacting much in other methods we can give them low interest offers in order increase our revenue in future.

5 Cluster:

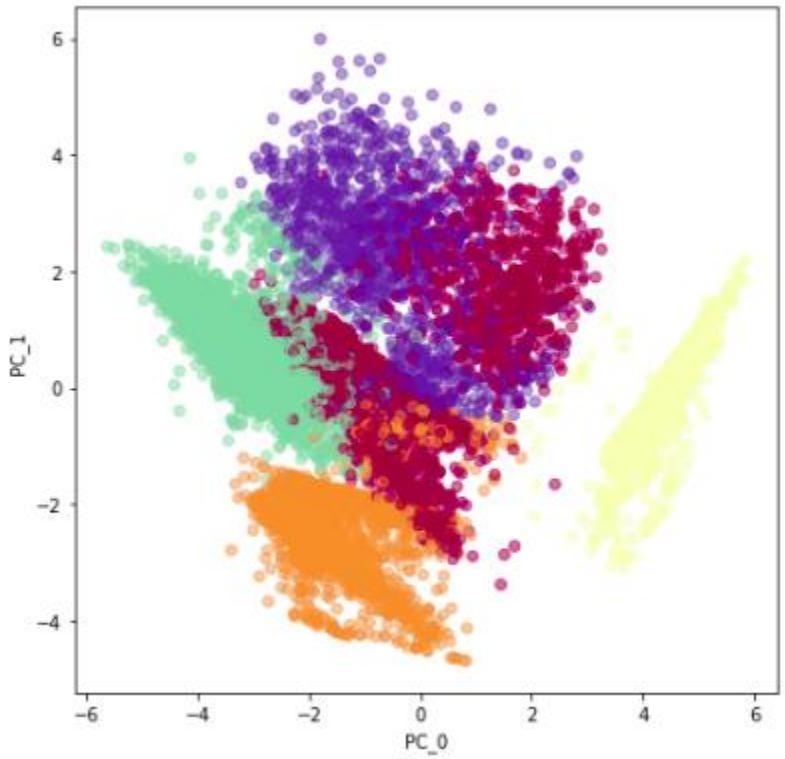

Fig 2.3.2 5 Clusters Scatter Plot

Below is the summary and Insights of clusters formed:

Cluster 0:

Customers are mostly making One off purchase and no installment purchases.

Cluster 1:

This set of customers are making installment purchases and making minimum payments towards the complete payment in billing cycle.

Cluster 2:

These customers are doing Cash advance transactions and are not transacting much in other categories.

Cluster 3:

These customers are highest in doing average monthly transactions and are using maximum credit limit among other groups.

Cluster 4:

This set of customers are doing both transactions and cash advances we can say this cluster might be a combination of cluster 2 and cluster 3.

Chapter 3

Conclusion

3.1 Model Evaluation

After generating predictions from few models as above we need to finalize and select the best performing and efficient model. Mostly models are compared based on below criteria:

- Predictive Performance
- Interpretability
- Computational Efficiency

In our clustering problem we will be using Calinski harabaz and Silhouette score to validate our clustering output.

3.2 Calinski Harabaz

Calinski Harabaz is used to compare clustering solutions obtained on the same data, solutions which differ either by the number of clusters or by the clustering method used. There is no "acceptable" cut-off value. We can simply compare CH values by eye. The higher the value, the "better" is the solution. If on the line-plot of CH values, there appears that one solution gives a peak or at least an abrupt elbow we choose it. If, on the contrary, the line is smooth - horizontal or ascending or descending - then there is no reason to prefer one solution to others. The math formula to measure is

$$\frac{SS_B}{SS_W} \times \frac{N-k}{k-1}$$

Where k is the number of clusters, and N is the total number of observations (data points), SSW is the overall within-cluster variance (equivalent to the total within sum of squares calculated above), SSB is the overall between-cluster variance.

3.3 Silhouette Score

Silhouette score is a way to measure how close each point in a cluster is to the points in its neighboring clusters. It is a neat way to find out the optimum value for k during k-means clustering. Silhouette values lies in the range of [-1, 1]. A value of +1 indicates that the sample is far away from its neighboring cluster and very close to the cluster its assigned. Similarly, value of -1 indicates that the point is close to its neighboring cluster than to the cluster its assigned. And, a value of 0 means it is at the boundary of the distance between the two cluster. Value of +1 is idea and -1 is least preferred. Hence, higher the value better is the cluster configuration.

Thus, the silhouette s(i) can be calculated as

$$s(i) = \frac{(b(i) - a(i))}{\max(b(i), a(i))}$$

The best advantage of using S.A. score for finding the best number of clusters is that you use it for un-labelled data set. This is usually the case when running k-means.

3.4 Conclusion

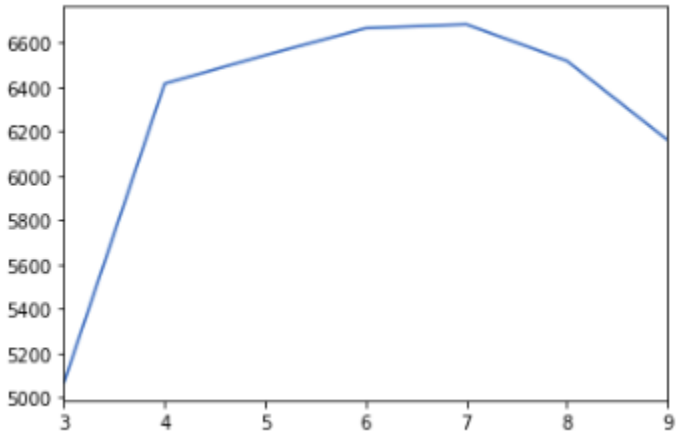

Fig 3.4.1. Calinski Harabaz *Curve*

From above Calinski Harabaz curve we can say that 4 clusters are the optimum number of clusters for our dataset. We will finalize the output of Cluster 4 as final on our credit card segmentation dataset.

Cluster 0: In first cluster customers are utilizing lowest credit limit amongst other clusters and they are mostly transacting in installment purchases.

Cluster 1: This set of customers are mostly doing one off purchase followed with cash advance transactions.

Cluster 2: This set of customers are doing maximum average purchases and are utilizing maximum credit limit. Basically, this are our high spending customers and are involved in both type of transactions one off and installments.

Cluster 3: These customers are making cash advance transactions mostly and are not transacting much in other methods we can give them low interest offers in order increase our revenue in future.

References

Edwisor study materials

Machine Learning (in Python & R) dummies by John Paul Mueller & Luca Massaron

R and Python official documents

https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a

https://towardsdatascience.com/a-one-stop-shop-for-principal-component-analysis-5582fb7e0a9c

https://ethen8181.github.io/machine-learning/clustering_old/clustering/clustering.html