Credit Card Customer Segmentation

Shourya Thatha Ravi, Usha Kiran Bellam, Rahul Rajesh Singh

Department of Statistics, Rutgers University, New Brunswick

1 Introduction

Segmenting the customers and profiling the segments created is an essential task for designing marketing strategy for products – new or matured. It has numerous applications in different fields like pharmaceuticals, banking, retail, genomics etc. This project aims to perform customer segmentation on a credit card dataset and profile the segments created. Segmentation and profiling customers is a task for data scientists which provides a good opportunity for the team to learn techniques and experiment with real-world data. The dataset illustrates the usage patterns of around 9000 active credit card users over 6 months. The data contains 18 behavioral variables at the customer level.

2 Review of Dataset

Variable	Function
Customer ID	Identification of Credit Card holder
Balance	Balance amount left in the account to make purchases
Balance frequency	How frequently the Balance is updated
Purchases	Total amount of purchases made from account
One off purchases	Maximum purchase amount done in one-go
Installments purchases	Amount of purchase done in installment
Cash advance	Using credit card to withdraw money
Purchases frequency	How frequently the purchases are being made
One off purchases frequency	How frequently Purchases are happening in one-go
Purchases installments frequency	How frequently purchases in installments are being done
Cash advance frequency	How frequently the cash in advance being paid
Cash advancetrx	Total number of Transactions made with "Cash in Advanced"
Purchases trx	Total number of purchase transactions made
Credit limit	Limit of Credit Card for a user
Payments	Total amount of payments made by a user
Minimum Payments	Minimum amount of payments made by a user
Prc full payment	Percent of full payment paid by a user
Tenure	Tenure of credit card service for a user (in months)

Table 1: Features in the credit card dataset

3 Exploratory Data Analysis

Based on the dataset, it was observed that few variables were not required for clustering. Initially, the Customer ID was dropped from the table. This is because it is an identifier, and it does not provide useful information about the data. An in-depth exploratory data analysis was performed on each of the variables. There were missing values in the variables Credit Limit and Minimum Payments. Each variable in the table was examined, and it was discovered that the variables Cash Advance Transaction, Purchase Transactions and Tenure are categorical and all other variables are continuous. Histograms in Figure 8 were plotted for different variables to understand how the data is distributed.

From Figure 8, it is seen that the distribution of Balance is right skewed since many values are zero on the left and there are wide range of different values on the right. Similarly, it was discovered that the variables purchases, one-off purchases, installment purchases, cash advance, cash advance frequency, cash advance transactions, purchase transactions, credit limit, minimum payments are all right skewed distributions.

The distribution of Balance frequency is left skewed, as can be seen from the Figure 9, with numerous values of 1 on the right and a broad range of other values on the left.

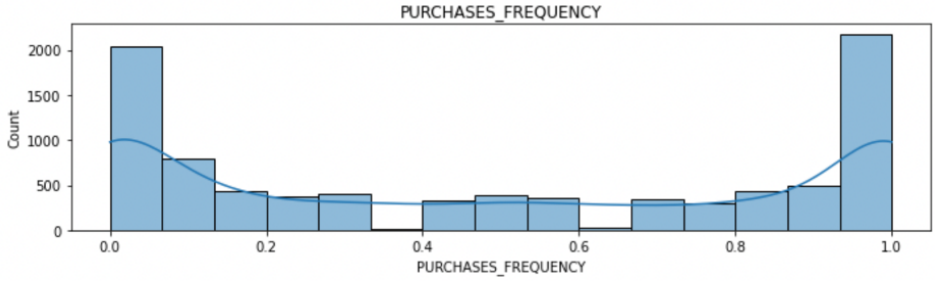

Figure 1: Histogram for Purchase Frequency

From Figure 1, it is observed that the purchases frequency is bimodal. This suggests that there are two distinct groups of customers who use their credit cards frequently or infrequently. Similarly, it was observed that the variable purchases installment frequent is also bimodal.

From Figure 10 Highly correlated features are dropped among the 17 behavioral variables. Columns such as One-Off Purchases Frequency and One-Off Purchases are similar, one of them is removed from the dataset. After selecting the relevant features, 10 variables are used for all the clustering algorithms

3.1 Handling of Missing Values

The use of clustering techniques, which aim to group points based on some similarity criterion, might be complicated by missing values. In the context of clustering, it is usual practice to first impute the missing values before using the clustering method on the whole data. In the variable Credit Limit, there is a single missing value. For this row, it was seen that the other variables are in lower quintiles and in the maximum density range. Hence, credit limit for this corresponding was replaced with the median of the credit limit as it will not cause any problems.

There are 313 missing values in the minimum payments variable. Metrics like mean, standard deviation were calculated (shown below) to see what kind of imputation is to be formulated. Also, a correlation check was performed on the minimum payments across all the variables. None of the variables were highly correlated with minimum payments. Replacing the missing values with methods like Regression or K-Nearest Neighbors will not be ideal. Therefore, the missing values for minimum payments was replaced by median of that column. From Figure 11, it is seen that the standard deviation is larger than the mean. Values of the minimum payments are varied. Minimum payments variable cannot be imputed by univariate imputation methods.

3.2 Outlier Treatment

An outlier is characterized as a noisy observation that does not fit the data's presumed generating model. Outliers are considered observations in clustering that should be eliminated to make clustering more reliable. If the clustered data contain outliers, the quality of the created clusters decreases significantly. Box plots were built on each of the variables to see the spread of data (shown in Fig 12).

From the boxplot the total purchases have many outliers from the actual range of values. To test this quantitatively, a metric was developed to check if the values fall completely outside the range. The upper threshold is mean + 3 * standard deviation, the lower threshold is mean – 3 * standard deviation. If the value of a variable crosses the upper or lower mark, it is capped at the threshold values. By applying this technique, the outliers are treated to fall in the range of the three standard deviations.

4 Modeling Methodologies

4.1 K-Means Clustering

The Go-To method for any clustering-based problem is the classical K Means Clustering, where we begin with "k" random points as centroids between the data-points and form clusters based on a metric (that is minimizing WCSS). This is followed by shifting these centroids based on the mean of the values in the cluster, till we either reach a minimum threshold of shift or exhaust our iterations. This is shown in Figure 13

4.2 Techniques to decide N clusters

Elbow Plot

WCSS or Within-Cluster Sum of Squares, is the sum of squares of the distances of each data point in all clusters to their respective centroids, which is later used as a metric to plot an elbow graph and decide a suitable number of clusters. However, as the number of clusters employed increases, the WCSS decreases since we end up with smaller

clusters which will have less distance on average (overfitting). Thus, we use an Elbow Plot is to select a feasible number of clusters which has a low enough WCSS but not too many clusters as shown in Figure 14.

Silhouette score plot

The silhouette score measures an object's similarity with its own cluster in comparison to other clusters (separation). A high value on the silhouette score implies that the object is well matched to its own cluster and poorly matched to neighboring clusters. The silhouette score has a range of -1 to +1 as shown in Figure 18. The silhouette can be calculated with any distance metric, such as the Euclidean distance or the Manhattan distance.

4.3 K-Means Clustering Results

Using K-means clustering we successfully segregated customers based on factors/variables given in the dataset into 6 distinct groups which would help us derive insightful observations like their spending habits, financial background, ability to pay off installments etc.

As seen in the K Means Clustering approach, we can achieve this to a certain extent, but not without some redundancies in terms of clusters formed. We summarize the values observed in the clusters and find the following about a few clusters as shown in Figure 2

Cluster Number	# customers	BALANCE	PURCHASES	CASH_ADVANCE	PURCHASES_FREQUENCY	CREDIT_LIMIT	PRC_FULL_PAYMENT	TENURE	CASH_ADVANCE_TRX	PURCHASES_TRX	Cluster Name
0	3206	1124	207	575	0.14	3191	3%	12	2	2	Inactive/Indolent customers
1	654	797	377	989	0.41	2344	14%	7	3	5	New customers
2	1081	108	1097	42	0.76	4774	79%	12	0	18	Transactors
3	611	2928	4996	586	0.95	8610	24%	12	2	63	VIP/Prime customers
4	2302	1133	1083	231	0.85	3880	6%	12	1	20	Active customers
5	1096	4508	512	4238	0.29	7967	4%	12	12	7	Revolvers

Figure 2: K-Means Clustering profiling

Cluster 0: (3206) Inactive/Indolent customers - These are customers who rarely use credit card for making purchases. Cluster 0 customers have low purchase frequency (0.14), low credit limit (3191), medium balance (1124), medium cash advance (575) and low PRC FULL PAYMENT 3%)

Cluster 1: (654) New customers - The customers allocated to this cluster had the lowest tenure (7.3 months) and a relatively low balance amount (797), typical of people testing something new.

Cluster 2: (1081) Transactors - These are customers who pay least amount of interest charges and careful with their money. Customers who belonged to Cluster 2 had the lowest amount of cash advance (42), low balance (108), and highest percentage of full payment = 79%.

Cluster 3: (611) VIP/Prime - These are premium customers who have high credit limit (8610), and high percentage of full payment (24%). High number of purchases (63), highest amount spend for purchases (4996). These are your customers who spend and use credit card a lot, which makes them very affluent and hence, potential targets of increased credit limit and purchases.

Cluster 4: (2302) Active customers – These are your active customers with high number of purchases (20), have medium balance (1132), medium credit limit (3880), and low percentage of full payment (6%). Target them for increase credit limit and increase spending habits.

Cluster 5: (1096) Revolvers - These are customers who use credit card as a loan (most lucrative sector). These customers have high balance (4508) and cash advance (4238), low purchase frequency (0.28), high cash advance transactions (12) and low percentage of full payment (4%). This can be ascribed as them revolving their credit due to high cash advances and balances, along with very low full payment percentages. This summarizes the results of K-Means Clustering achieved.

Principal Component Analysis

Moving forward, to visualize how well the formed clusters are separated, we reduced the dimensions to 2 using PCA and showed the customers in the Figure 3. The distribution of certain variables for each cluster is shown in Figure 16 in appendix.

4.4 Agglomerative Clustering

Agglomerative Clustering is a form of Hierarchical Clustering where we start by considering each data point its own individual cluster and work our way up joining clusters based on a distance metric, till we end up with a single cluster as shown in Figure 17. This distance metric can be:

Single Linkage: the closest pair of data points belonging to different clusters

Complete Linkage: the largest distance between two observations of two clusters

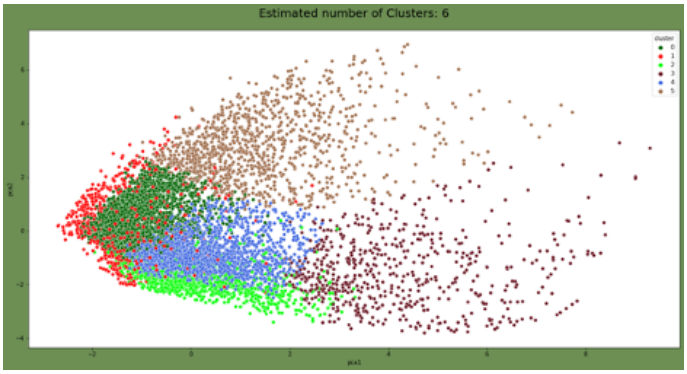

Figure 3: PCA for K-means clustering

Average Linkage: the average distance between all the data points belonging to 2 clusters

Ward's method: Ward's procedure is a variance method which attempts to generate clusters to minimize the within-cluster variance.

In our project, we used Ward's method as the distance metrics as the features that we have in the dataset are all quantitative and Ward's method is said to be the most suitable method for quantitative variables. Silhouette score is plotted for different cluster counts and the number of clusters are estimated as shown in Figure 15.

Though the highest silhouette score is for 11 clusters it may not be practical for credit card company to have 11 different strategies for 11 different customer groups and hence 6 is selected as the suitable number of clusters and Agglomerative clustering is applied onto the model to derive the following results shown in Figure 4.

Cluster Number	# customers	BALANCE	PURCHASES	CASH_ADVANCE	PURCHASES_FREQUENCY	CREDIT_LIMIT	PRC_FULL_PAYMENT	TENURE	CASH_ADVANCE_TRX	PURCHASES_TRX	Cluster Name
0	1559	3767	407	3436	0.23	7174	4%	11	10	6	Revolvers
1	1903	1307	1117	354	0.86	4087	4%	12	1	20	Active Customers
2	843	2408	4308	490	0.94	7945	36%	12	2	57	VIP/Prime Customers
3	3125	905	257	425	0.19	2809	3%	12	2	3	Inactive/Indolent customers
4	559	513	521	513	0.53	2264	27%	7	2	7	New Customers
5	961	77	748	13	0.75	4316	71%	12	0	14	Transactors

Figure 4: Agglomerative Clustering Profiling results

This summarizes the results of Agglomerative Clustering achieved. We can see the clusters formed in the Agglomerative clustering are same as what we observed from K-Means clustering. However, there is some movement from one cluster to other and this movement can be observed from Figure 19. There is net 80% of overlap between the customers falling under same segments for both the Clustering Models. While viewing from the K Means' perspective, one can notice that Agglomerative Clustering has classified most of the customers for the 6 Customer Types correctly or in accordance to the K Means Clusters. With the maximum deviation observed for Active customers group. Moving forward, to visualize how well the formed clusters are separated, we reduced the dimensions to 2 using PCA and showed the customers in Figure 5.

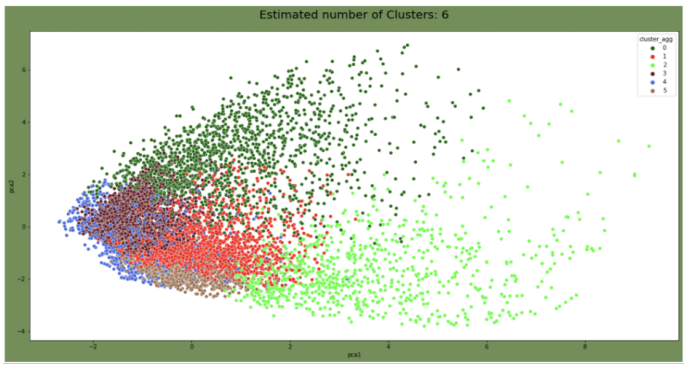

Figure 5: PCA for Agglomerative clustering

4.5 Spectral Clustering

For data which is non-convex in nature, K Means Clustering fails to perform well. That is when Spectral Clustering comes in the picture. The basic idea behind this approach is that data which is not linearly separable in "d" (¡N) dimensions, will almost always be linearly separable in higher dimensions (d¿=N). Thus, data points are mapped onto a higher dimension (N, in this case) by creating Similarity Matrix. This matrix can be constructed by using Pairwise Distances or any Kernel Based Functions. The clusters are segregated in this space by attempting to construct a fitting hyperplane. This step which would usually require computing in N space, is reduced by the implementation of a procedure similar to PCA for the Laplacian Graph formulated from the Similarity Matrix. The construction of a Laplacian Graph can be done in many ways, but over here it is done so by constructing a Normalized Laplacian in Figure 6.

$$L = D^{-1/2}AD^{-1/2}$$

Figure 6: Normalized Laplacian

D is a diagonal matrix whose (i,i)th element is the sum of the Adjacency Matrix's ith row. The Eigen Decomposition of the Laplacian Graph yields projection of the initial data points on these Principal Components (Eigen Vectors), rather than the Principal Components themselves. Thus, k smallest Eigen Values are used to construct a Z matrix which is formed by stacking the corresponding Eigen Vectors for these Eigen Values.
Thus, the ith component of this matrix (Zi) represents the ith data point mapped according to the Principal Components of N space. In a way, Spectral Clustering is similar to Kernel PCA, wherein, the difference lies in construction of Laplacian Graph in Spectral Clustering as opposed to K matrix in Kernel PCA. In the experiment, Eigen Decomposition of the Laplacian Graph yields the following Eigen Values in Figure 20.

As seen in Figure 20, the first 5 Eigen Values have a magnitude of less than 0.1 and thus, their corresponding vectors are used to form the Z matrix. Later, any clustering method can now be applied on this Matrix and results be derived for the same. As mentioned previously, an Elbow Plot is plotted and the number of clusters are estimated as shown in Figure 21.

Thus, 4 is selected as the suitable number of clusters and K Means is applied onto the model to derive the following results shown in Figure 7

Cluster Number	# customers	BALANCE	PURCHASES	CASH_ADVANCE	PURCHASES_FREQUENCY	CREDIT_LIMIT	PRC_FULL_PAYMENT	TENURE	CASH_ADVANCE_TRX	PURCHASES_TRX	Cluster Name
0	3574	805	834	116	0.61	3614	6%	12	1	14	Active Customers
1	763	774	431	884	0.43	2268	13%	8	3	6	New Customers
2	3463	2873	918	2007	0.28	5592	4%	12	7	12	Revolvers
3	1150	141	1491	29	0.78	5078	80%	12	0	22	Transactors

Figure 7: Spectral Clustering Profiling results

To visualize these clusters, dataset was reduced to 2 parameters using PCA and the following was plotted as shown in Figure 22 in appendix. As mentioned in Figure 23, there is net 54% of overlap between the customers falling under same segments for both the Clustering Models. While viewing from the K Means' perspective, one can notice that Spectral Clustering has classified most of the customers for the 4 Customer Types correctly or in accordance to the K Means Clusters. However, there are 2 additional clusters in K Means, and these customers were fit in the available 4 clusters of the Spectral Clusters. As seen, Inactive customers were included mostly in either Active Customers or Revolvers. Whereas, a good chunk of VIP/Prime Customers were a part of the revolvers cluster in Spectral Clusters.

5 Conclusion

Clearly, six clusters obtained from K-Means and Agglomerative clustering are more actionable than the four clusters obtained from Spectral clustering as Revolvers from spectral clustering contain a mix of different clusters like Revolvers, VIP, and Inactive customers from K-Means. Of K-Means and Agglomerative clustering, the separation of clusters is more clear in K-Means than in Agglomerative based on observing the principle component analysis plots for both. Hence for the dataset taken after removing outliers K-Means algorithm gave the best clusters. The best clustering algorithm is also reconfirmed by the high silhouette score in K-Means than Agglomerative clustering for the chosen optimal number of clusters. The marketing strategy recommendation for different clusters of K-Means is shown in Figure 24

References

https://towardsdatascience.com/introduction-hierarchical-clustering-d3066c6b560e
http://www.umiacs.umd.edu/labs/cvl/pirl/vikas/projects/spectral_clustering.pdf
https://en.wikipedia.org/wiki/Spectral_clustering
https://www.ibm.com/blogs/research/2018/08/spectral-clustering/

6 Appendix

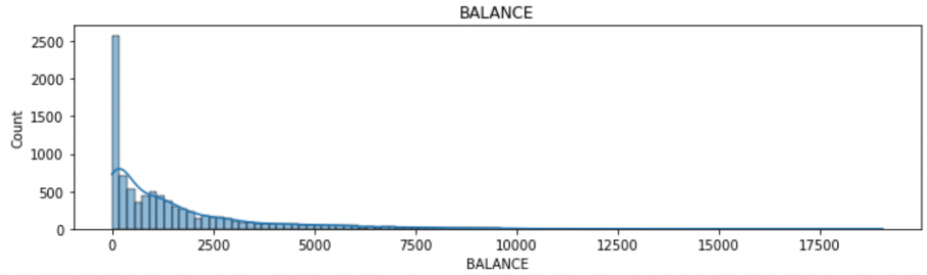

Figure 8: Histogram for Balance

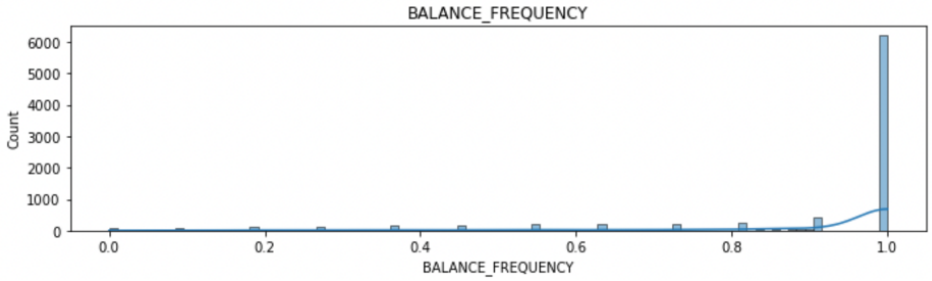

Figure 9: Histogram for Balance Frequency

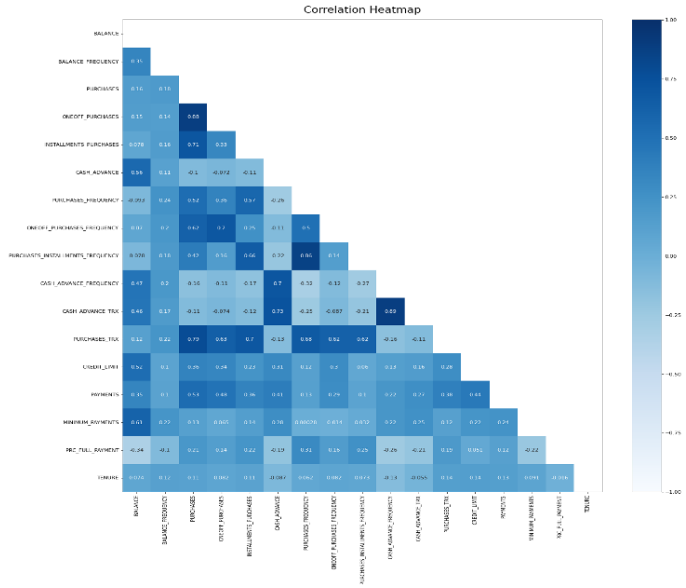

Figure 10: Correlation Matrix

```
mean     8.642065e+02
std      2.372447e+03
min      1.916300e-02
max      7.640621e+04
var      5.628503e+06
Name: MINIMUM_PAYMENTS
```

Figure 11: Minimum Payments Summary

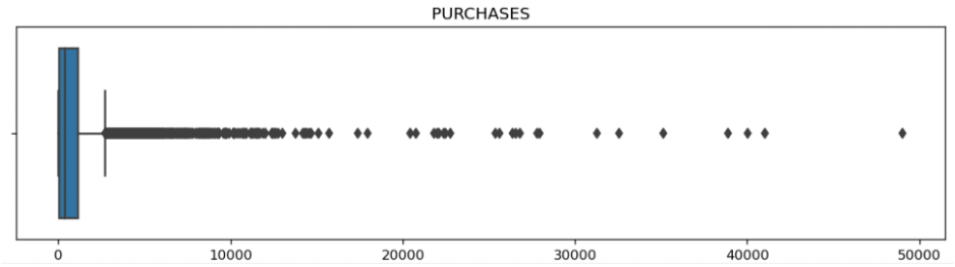

Figure 12: Boxplot for Purchases

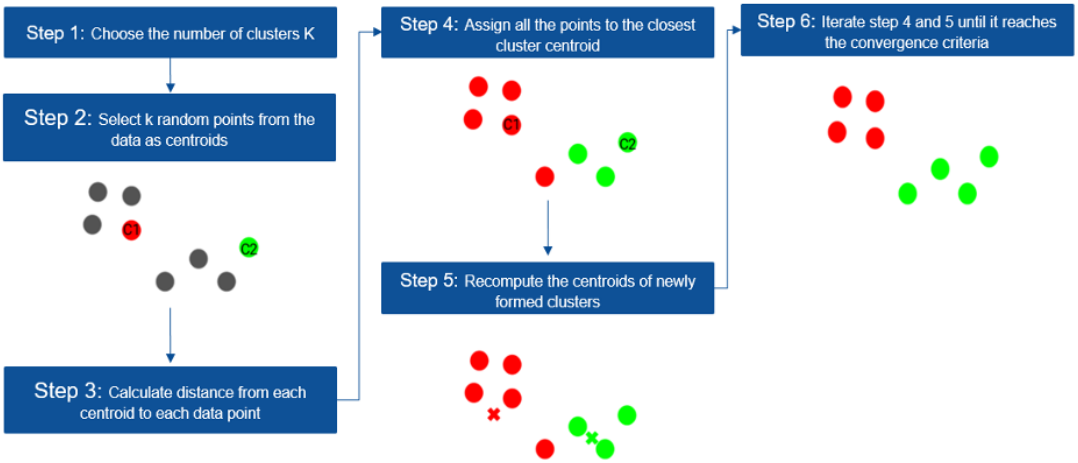

Figure 13: K-means clustering

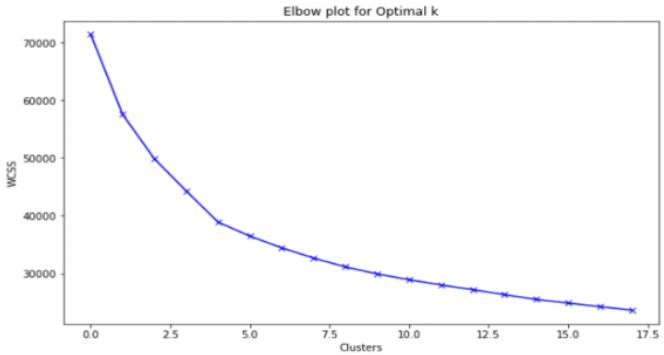

Figure 14: Elbow Plot

Figure 15: Slihoutte plot

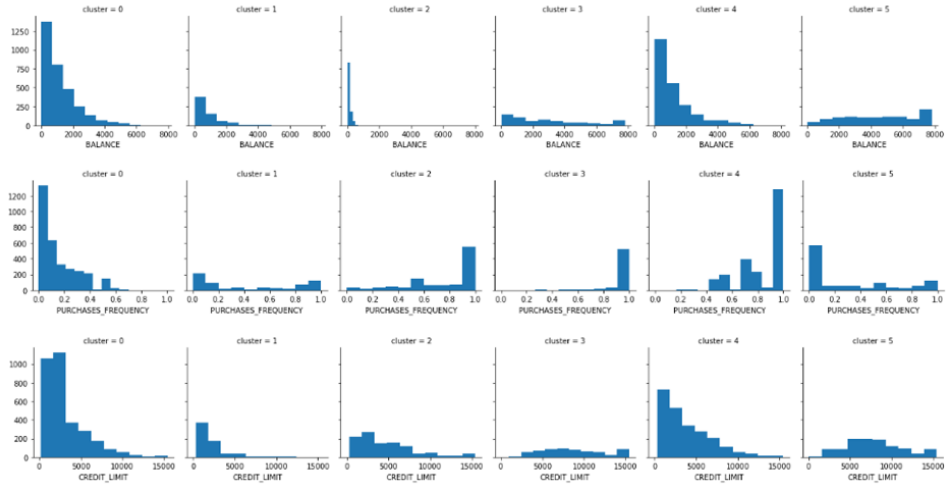

Figure 16: Distribution of few variables for each cluster

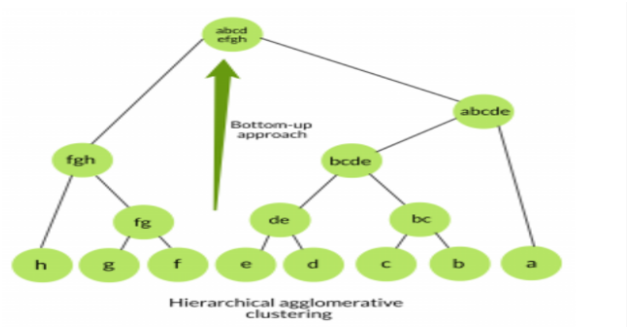

Figure 17: Agglomerative Clustering

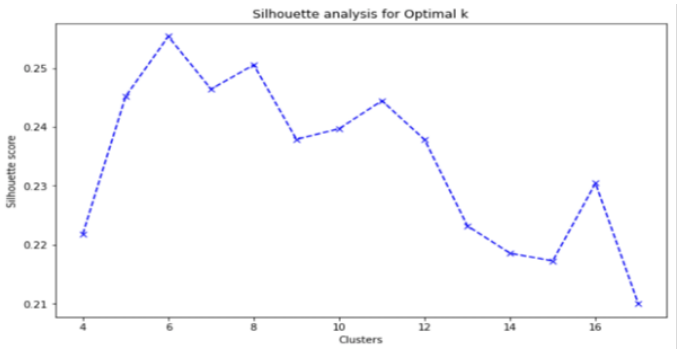

Figure 18: Silhouette score plot

Overlap of K-means and Agglomerative clusters			Agglomerative clusters					
			3125	559	961	843	1903	1559
			Inactive/Indolent customers	New customers	Transactors	VIP/Prime customers	Active customers	Revolvers
K-Means clusters	3206	Inactive/Indolent customers	2751	10	25	0	69	351
	654	New customers	46	471	0	0	21	116
	1081	Transactors	7	62	792	186	24	10
	611	VIP/Prime customers	0	2	0	555	30	24
	2302	Active customers	308	13	144	83	1685	69
	1096	Revolvers	13	1	0	19	74	989

Figure 19: Segment Overlap of K-means and Agglomerative clusters

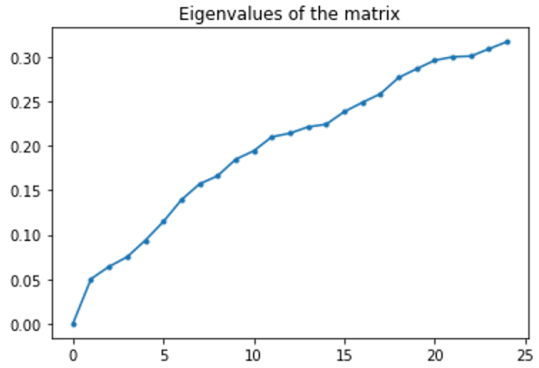

Figure 20: Eigen Values

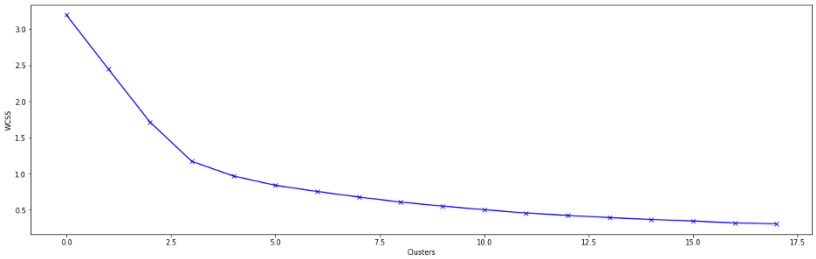

Figure 21: Elbow plot

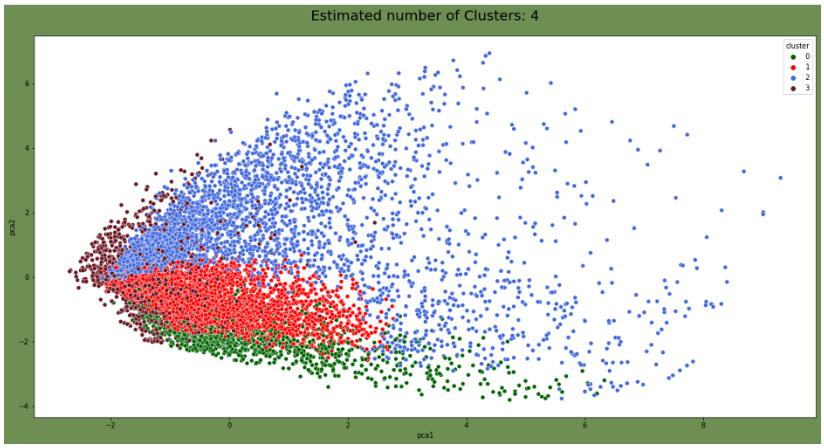

Figure 22: PCA

Overlap of K-means and Spectral clusters			Spectral clusters			
			763	1150	3574	3463
			New customers	Transactors	Active customers	Revolvers
K-Means clusters	654	New customers	641	2	0	11
	1081	Transactors	13	1014	40	14
	2302	Active customers	41	8	2074	179
	1096	Revolvers	5	0	0	1091
	3206	Inactive/Indolent customers	62	19	1400	1725
	611	VIP/Prime customers	1	107	60	443

Figure 23: Segment Overlap of K-Means and Spectral clusters

Cluster #	Cluster Name	Recommendation
0	Inactive/Indolent customers	Direct emails for cashbacks/promotions
1	New customers	Increase in credit limit based on usage, offers to encourage activity (Ex: 12 transactions in Q1, get $100)
2	Transactors	Promotions/cashbacks
3	VIP/Prime customers	New premium credit cards with higher credit limit, increased benefits
4	Active customers	Increase credit limit to drive more cativity
5	Revolvers	Lower APR %

Figure 24: Marketing Strategy Recommendation

6.1 Agglomerative Clustering profiling

Cluster 0: (1559) Revolvers - These are customers who use credit card as a loan (most lucrative sector). These customers have high balance (3767) and cash advance (3435), low purchase frequency (0.23), high cash advance transactions (10) and low percentage of full payment (4%). This can be ascribed as them revolving their credit due to high cash advances and balances, along with very low full payment percentages.
Cluster 1: (1903) Active customers – These are your active customers with high number of purchases (20), have medium balance (1306), medium credit limit (4086), and low percentage of full payment (4%). Target them for increased credit limit and increased spending habits.
Cluster 2: (843) VIP/Prime - These are premium customers who have high credit limit (7944), and high percentage of full payment (36%). High number of purchases (57), highest amount spend for purchases (4307). These are your customers who spend and use credit card a lot, which makes them very affluent and hence, potential targets of increased credit limit and purchases.
Cluster 3: (3125) Inactive/Indolent customers - These are customers who rarely use credit card for making purchases. Cluster 3 customers have low purchase frequency (0.18), low credit limit (2809), medium balance (905), medium cash advance (425) and low PRC FULL PAYMENT (3%).
Cluster 4: (559) New customers - The customers allocated to this cluster had the lowest tenure (7.4 months) and a relatively low balance amount (513), typical of people testing something new.
Cluster 5: (961) Transactors - These are customers who pay least amount of interest charges and careful with their money. Customers who belonged to Cluster 5 had the lowest amount of cash advance (13), low balance (77), and highest percentage of full payment = 71%.

6.2 Spectral Clustering profiling

Cluster 0 : (805) Active Customers- Customers with a moderate Balance (802), and a relatively high Purchase Transactions (14) are clustered in this group. They are also noticed to have a very low full payment percentage of 6%
Cluster 1: (774) New Customers- The customers under this cluster were noticed not only to have the lowest tenure of all the groups (8), but also have a typically low balance (774)
Cluster 2: (2873) Revolvers- As seen in the previous inferences, revolvers were found to have a high Balance (2873) despite having the lowest full payment percentage 4%. They also had the highest cash advance (2007) alongside a high credit limit of 5598.
Cluster 3: (141) Transactors- These customers are characterized by the lowest cash advance out of all (29) while maintaining a good percentage of full payment of 80%. They tend to have the lowest balance on average as well (141).