Credit Card Segmentation

Table of Contents

Problem Statement

This case requires to develop a customer segmentation to define marketing strategy. The sample dataset summarizes the usage behavior of about 9000 active credit card holders during the last 6 months. The file is at a customer level with 18 behavioral variables.

Data Description:

- CUST_ID : Credit card holder ID
- BALANCE : Monthly average balance (based on daily balance averages)
- BALANCE_FREQUENCY : Ratio of last 12 months with balance
- PURCHASES : Total purchase amount spent during last 12 months
- ONEOFF_PURCHASES : Total amount of one-off purchases
- INSTALLMENTS_PURCHASES :Total amount of installment purchases
- CASH_ADVANCE : Total cash advance amount
- PURCHASES_ FREQUENCY: Frequency of purchases (percentage of months with at least on purchase)
- ONEOFF_PURCHASES_FREQUENCY : Frequency of one-off-purchases
- PURCHASES_INSTALLMENTS_FREQUENCY : Frequency of installment purchases

- CASH_ADVANCE_ FREQUENCY : Cash-Advance frequency
- AVERAGE_PURCHASE_TRX :Average amount per purchase transaction
- CASH_ADVANCE_TRX : Average amount per cash-advance transaction
- PURCHASES_TRX : Average amount per purchase transaction
- CREDIT_LIMIT : Credit limit
- PAYMENTS :Total payments (due amount paid by the customer to decrease their statement balance) in the period
- MINIMUM_PAYMENTS : Total minimum payments due in the period.
- PRC_FULL_PAYMENT: Percentage of months with full payment of the due statement balance
- TENURE :Number of months as a customer

Imputing Missing Values:

There are 313 missing values in 'MINIMUN_PAYMENTS' whereas 1 in 'CREDIT_LIMIT '. Both have float64 as datatypes.

As Mean value is affected by extreme values, we will impute the missing values with Median value.

Key Performance Indicator (KPIs):

KPI represent a set of measures that focus on important aspects of business performance for the overall success of the business.

We derived 6 KPIs for the provided data. They are as below:

1. Monthly Average Purchase:
   MONTHLY_AVG_PURCHASE = PURCHASES / TENURE
2. Cash Advance Amount:
   CASH_ADV_AMOUNT = CASH_ADVANCE / TENURE
3. Purchase by Type:
   It has 4 types:
   - Both ONEOFF_PURCHASES and INSTALLMENTS_PURCHASES having value as zero.
   - Both ONEOFF_PURCHASES and INSTALLMENTS_PURCHASES having value as zero.
   - ONEOFF_PURCHASES have value as zero and INSTALLMENTS_PURCHASES greater than zero.
   - ONEOFF_PURCHASES having value greater than zero and INSTALLMENTS_PURCHASES having value as zero.

4. Cash Advance Transaction:
   Already available in the dataset.
5. Limit Usage:
   LIMIT_USAGE = BALANCE / CREDIT_LIMIT
6. Payments to Minimum Payment Ratio:
   PAYMENT_MIN_PAY = PAYMENTS / MINIMUM_PAYMENTS

Pandas Profiling:

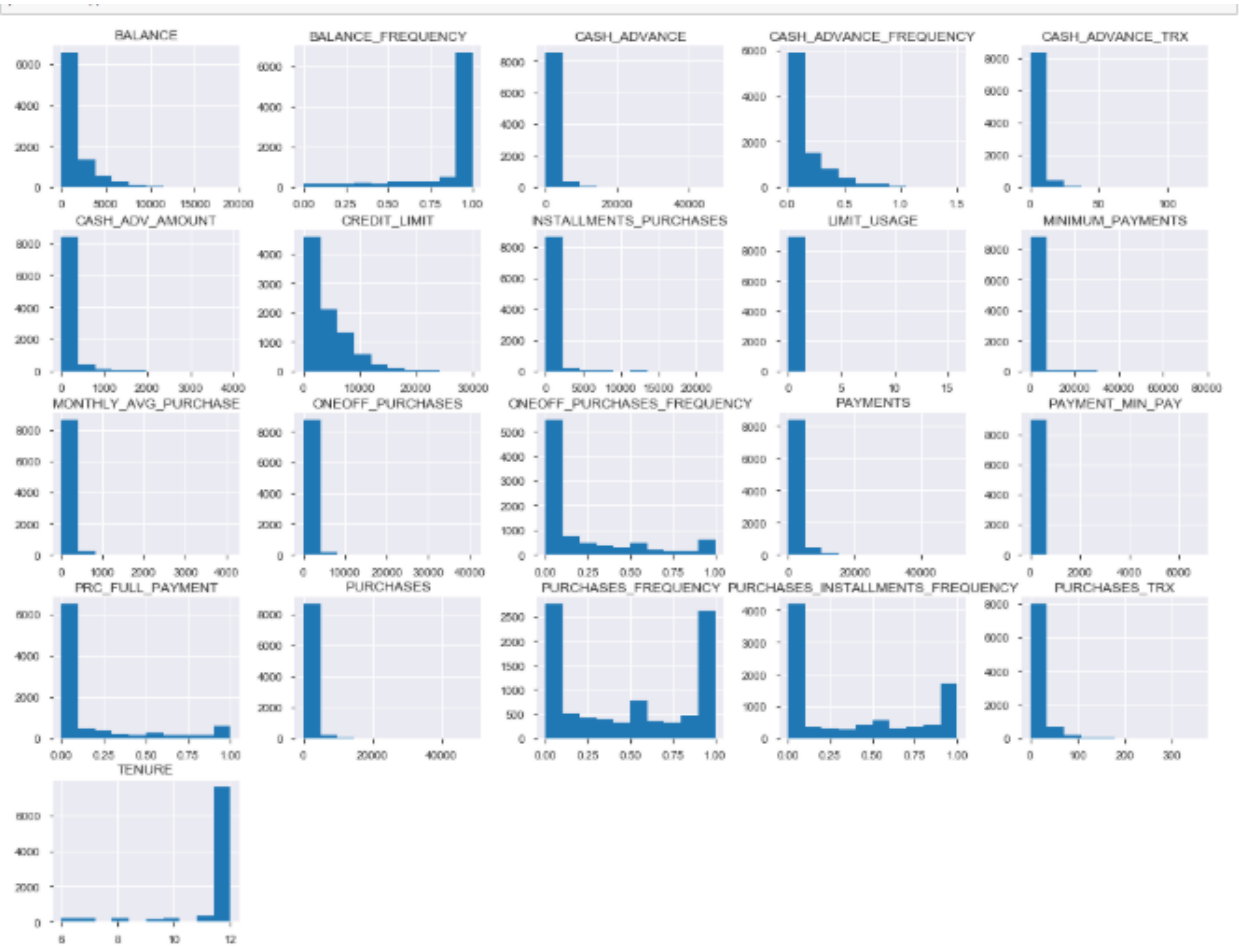
We run a Pre Profiling on the data set which gives us information regarding th e columns and detailed report regarding the correlation between the variables.

A detailed HTML report is generated which provides us with the visualization which further can be used for detailed analysis.

Observation:

- **CASH_ADV_AMOUNT** has 4628 (51.7%) zeros. Acceptable
- **CASH_ADVANCE** is highly correlated with CASH_ADV_AMOUNT (ρ = 0.9763639831) . The other is the derived KPI, Hence Acceptable.
- **CASH_ADVANCE_FREQUENCY** has 4628 (51.7%) zeros. Acceptable.
- **CASH_ADVANCE_TRX* has 4628 (51.7%) zeros. Acceptable.
- **INSTALLMENTS_PURCHASES** has 3916 (43.8%) zeros. Acceptable.
- **MONTHLY_AVG_PURCHASE** has 2044 (22.8%) zeros. Acceptable.
- **ONEOFF_PURCHASES** is highly correlated with MONTHLY_AVG_PURCHASE (ρ =0.9130598274) The other is the derived KPI, Hence Acceptable
- **ONEOFF_PURCHASES_FREQUENCY** has 4302 (48.1%) zeros. Acceptable.
- **PAYMENT_MIN_PAY** is highly skewed (γ1 = 43.00419578) We will apply log function.
- **PAYMENT_MIN_PAY** has 240 (2.7%) zeros Acceptable.
- **PAYMENTS** has 240 (2.7%) zeros. Acceptable.
- **PRC_FULL_PAYMENT** has 5903 (66.0%) zeros. Acceptable.
- **PURCHASES** is highly correlated with ONEOFF_PURCHASES (ρ = 0.9168445587) The other is the derived KPI, Hence Acceptable.
- **PURCHASES_FREQUENCY** has 2043 (22.8%) zeros. Acceptable.
- **PURCHASES_INSTALLMENTS_FREQUENCY** has 3915 (43.7%) zeros. Acceptable.
- **PURCHASES_TRX** has 2044 (22.8%) zeros. Acceptable.

Distribution of Data:



Applying Log function to data so that data is normally distributed.

Dimension Reduction:

Principal Component Analysis (PCA):

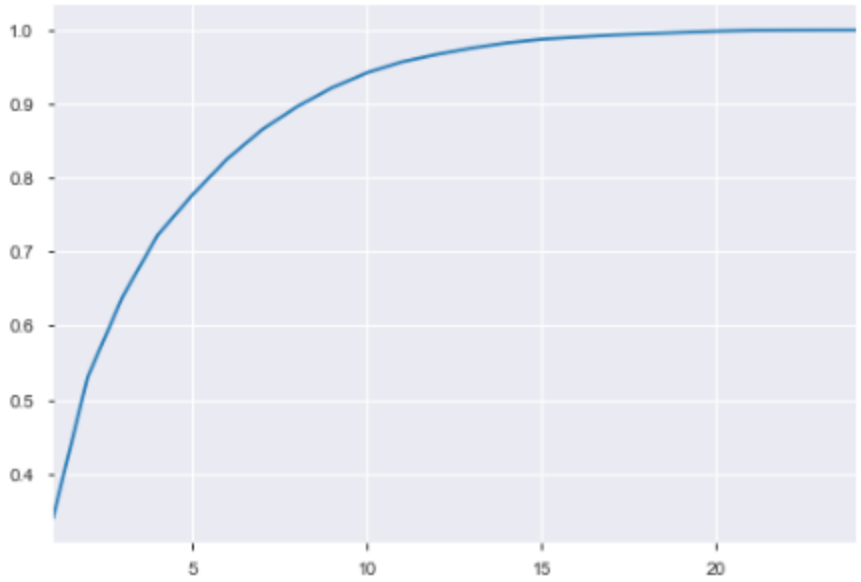
It is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables (entities each of which takes on various numerical values) into a set of values of linearly uncorrelated variables called principal components.

Each component explains its own variance. The second component explains the variance not explained by the first one and so on.

Below is the screen shot of the cumulative variance ratios of the PCs :

```
{1: 0.3393815165300376,
 2: 0.5308083460740263,
 3: 0.6393759525666121,
 4: 0.7224095329850195,
 5: 0.7771153282231585,
 6: 0.8258515432615212,
 7: 0.8654546381232499,
 8: 0.8964401559655443,
 9: 0.9219882654945665,
 10: 0.9421544636077164,
 11: 0.9569441171577663,
 12: 0.966931807032057,
 13: 0.9755495554634527,
 14: 0.9825397681633778,
 15: 0.9874370993624989,
 16: 0.9903194274405663,
 17: 0.9929812669029867,
 18: 0.9948892862754775,
 19: 0.9966267688292387,
 20: 0.9983363735473705,
 21: 0.9994599612204833,
 22: 0.9997212352014335,
 23: 0.9999666107497271,
 24: 1.0}
```

Plot for the same:



Summation of first 6 components results as ~82 i.e these 6 components contribute 82% of variance. Hence, we will finally fix n_components value as 6.

Cluster Building:

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. In other words, we try to find homogeneous subgroups within the data such that data points in each cluster are as similar as possible according to a similarity measure such as euclidean-based distance or correlation-based distance. The decision of which similarity measure to use is application-specific.

We will be clustering using k-Means Algorithm.

K-Means Clustering:

K-means algorithm is an iterative algorithm that tries to partition the dataset into *K*pre-defined distinct non-overlapping subgroups (clusters) where each data point belongs to **only one group**. It tries to make the inter-cluster data points as similar as possible while also keeping the clusters as different (far) as possible. It assigns data points to a cluster such that the sum of the squared distance between the data points and the cluster's centroid (arithmetic mean of all the data points that belong to that cluster) is at the minimum. The less variation we have within clusters, the more homogeneous (similar) the data points are within the same cluster.

The way k-means algorithm works is as follows:

1. Specify number of clusters *K*.

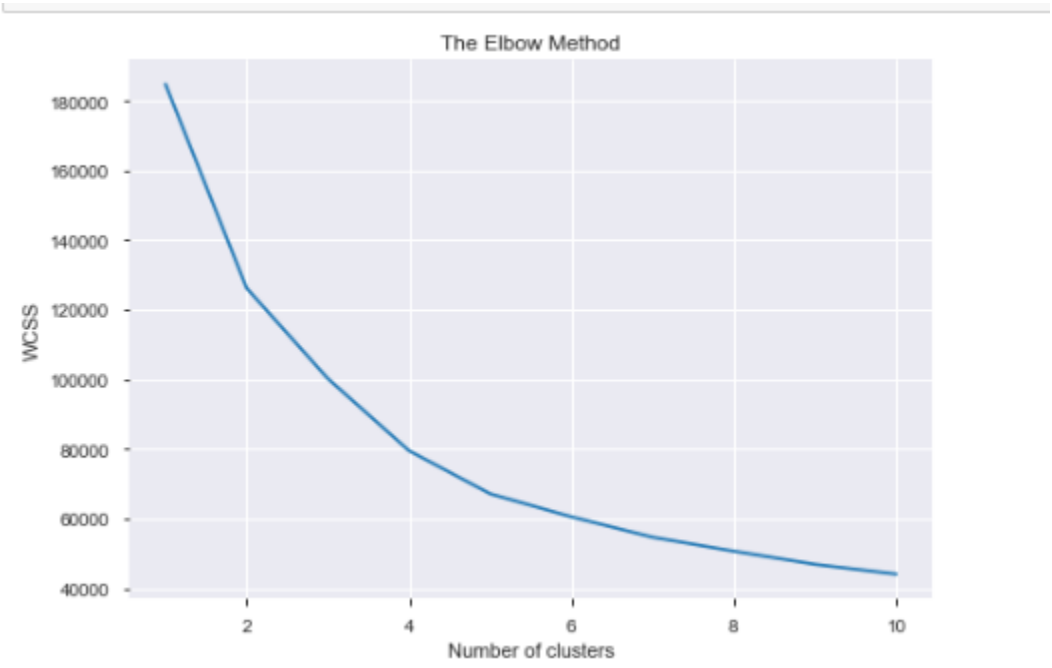
2. Initialize centroids by first shuffling the dataset and then randomly selecting *K* data points for the centroids without replacement.

3. Keep iterating until there is no change to the centroids. i.e assignment of data points to clusters isn't changing.

Finding optimum number of clusters:

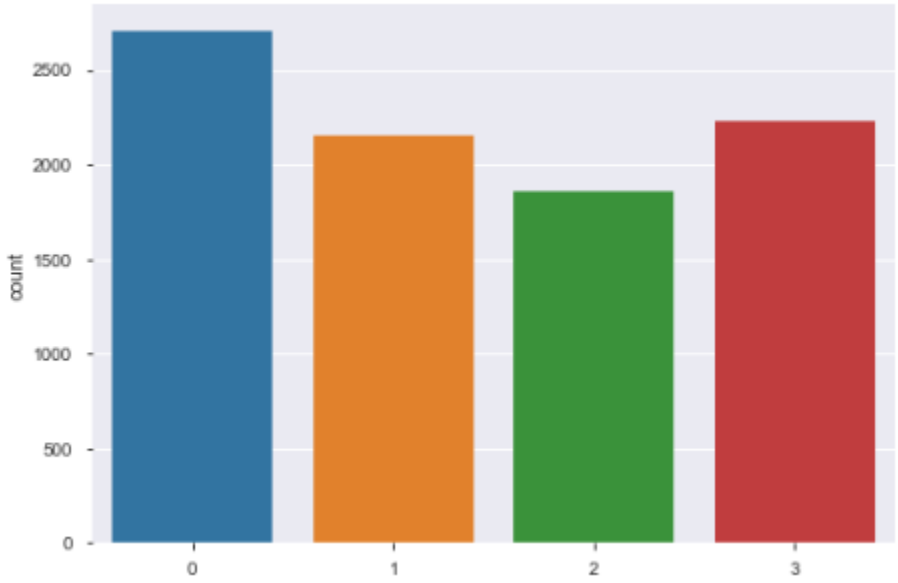
1. Elbow Method:

Elbow method gives us an idea on what a good *k* number of clusters would be based on the sum of squared distance (SSE) between data points and their assigned clusters' centroids. We pick *k* at the spot where SSE starts to flatten out and forming an elbow. We'll use the geyser dataset and evaluate SSE for different values of *k* and see where the curve might form an elbow and flatten out.



The above graph clearly suggests 4 as the optimum number of clusters. Hence fitting the K-Means with number of clusters as 4.

Distribution of data in 4 clusters:



Model Evaluation:

I. Calinski-Harabasz Index:

The Calinski-Harabasz index compares the variance between-clusters to the variance within each cluster. This measure is much simpler to calculate then the Silhouette score however it is not bounded. The higher the score the better the separation is.

Value for Model: `3948.696868270533`

II. Davies-Bouldin Index

The intuition behind Davies-Bouldin index is the ratio between the within cluster distances and the between cluster distances and computing the average overall the clusters. It is therefore relatively simple to compute, bounded – 0 to 1, lower score is better.

However, since it measures the distance between clusters' centroids it is restricted to using Euclidean distance function.

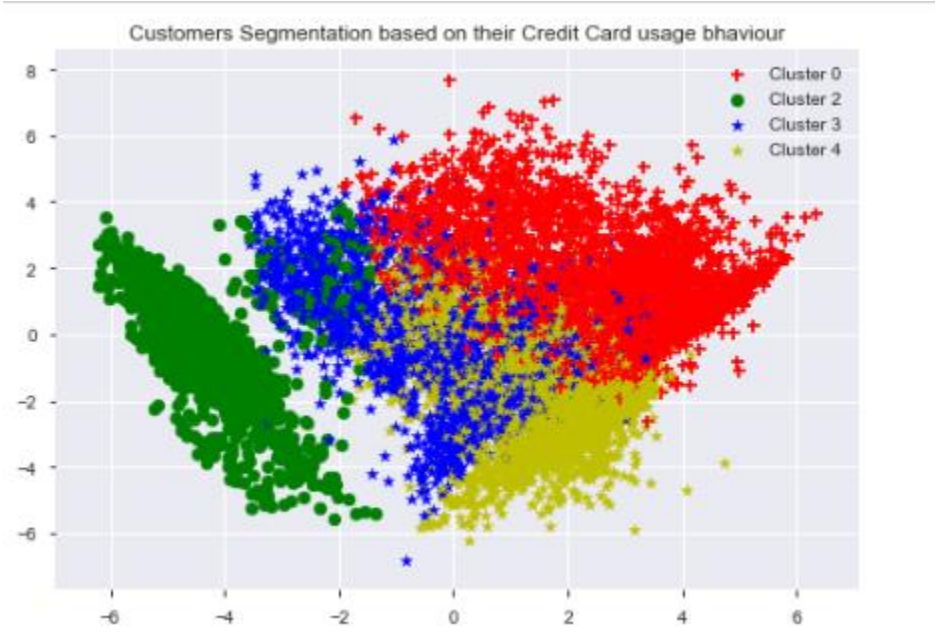
Value for Model: `1.1915432841102769`

III. The Silhouette score:

The Silhouette score reflects how similar a point is to the cluster it is associated with. i.e .for each point with compute the average distance of the point from the points in the nearest cluster minus the average distance of the point from the points in its own cluster divided by the maximum between those distances. The overall score is the average of the score per point. The Silhouette score is bounded from -1 to 1 and higher score means more distinct clusters.

Value for the Model: `0.35996655339002664`

Visualization of Clusters



Interpretations from Clusters:

- *Cluster 0* : Customer of this cluster have a **high credit limit** hence their **Purchases** and **Monthly Payments** are good. Also their credit score i.e** limit usage** is also moderate. Most of the **old customers** belong to this group. Purchases are done in both Installments and oneoff.

- *Cluster 1* : This group has the **highest** number of people opting for **Cash Advance.** Hence Monthly purchase transactions are low. **Credit limit** is moderately high. **Minimum Payment** ratio is good.
- *Cluster 2 :* * *Purchase transactions** and **Monthly purchases** are good. **Cash Advance** is comparitevly low. **Credit score** is a bit less. Purchases are done in **one time transactions.**
- *Cluster 3:* People with moderate **credit limit** with lowest number option for** Cash Advance**. Mostly purchases are done in **Installments**.

Marketing Strategy:

- ✓ *Cluster 0* : As most of the customers are associated since a long time they can be assigned some reward points upon transactions. As their purchase rates are good, can be offered shopping specific credit cards which provides some offers on different purchase categories. Credit limit can also be increased.
- ✓ Cluster 1* : A large number of these customers are only into cash advance, hence low interest rates can be provided to the same on purchases done in installments.
- ✓ *Cluster 2 : * This is a slightly risky group as most of the crowd is indulged only in one off purchases.

- ✓ *Cluster 3:* Only Installment transactions are done by this group . Credit limit can be increased with low interest fee charged on Cash Advance to lure them into the same.