Siddhi Padekar

25.07.2021

Bank data analysis

Table of Contents Contents Executive

|  |  |
| --- | --- |
| Summary | 2 |
| Introduction | 2 |
| Data Description . | 2 |
| Sample of the dataset: | 2 |
| Exploratory Data Analysis | 3 |
| Types of variables in the data frame | 3 |
| Missing values in the dataset | 3 |
| Correlation Plot | 4 |
| Pairplot | 4 |
| Do you think scaling is necessary for clustering in this case? Justify | 6 |
| Apply hierarchical clustering to scaled data. Identify the number of optimum clusters using Dendrogram and briefly describe them | 6 |
| Apply K-Means clustering on scaled data and determine optimum clusters. Apply elbow curve and silhouette score. Explain the results properly. Interpret and write inferences on the finalized clusters. | 8 |
| Describe cluster profiles for the clusters defined. Recommend different promotional strategies for different clusters. | 10 |

Executive Summary

A leading bank wants to promote its exclusive credit card offers to its credit card holder customers. The data set consists of different activities of users for the past few months. In this problem statement, we will explore the different activities of customers based on Income and spending pattern and will help to identify the customers who will be benefited with the new promotional offers provided by bank.

Introduction

The purpose of this whole exercise is to explore the dataset. Do the exploratory data analysis. Predict the dataset using different tools. The data consists of 211 customers. This assignment Will help us in exploring the hierarchical clustering to scaled data , K-Means clustering on scaled data and optimum clusters using dendogram.

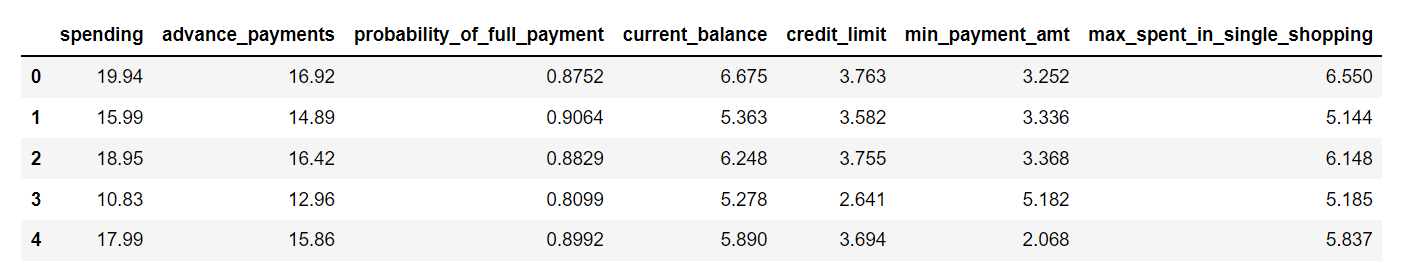
Data Description

1. spending: Amount spent by the customer per month (in 1000s)
2. advance\_payments: Amount paid by the customer in advance by cash (in 100s)
3. probability\_of\_full\_payment: Probability of payment done in full by the customer to the bank
4. current\_balance: Balance amount left in the account to make purchases (in 1000s)
5. credit\_limit: Limit of the amount in credit card (10000s)
6. min\_payment\_amt : minimum paid by the customer while making payments for purchases

made monthly (in 100s)

1. max\_spent\_in\_single\_shopping: Maximum amount spent in one purchase (in 1000s)

Sample of the dataset:



Exploratory Data Analysis

Let us check the types of variables in the data frame.

Data columns (total 7 columns):

spending float64

advance\_payments float64

probability\_of\_full\_payment float64

current\_balance float64

credit\_limit float64

min\_payment\_amt float64

max\_spent\_in\_single\_shopping float64

There are total 210 rows and 7 columns in the dataset. All the columns are Float data type.

Check for missing values in the dataset:

spending non-null

advance\_payments non-null

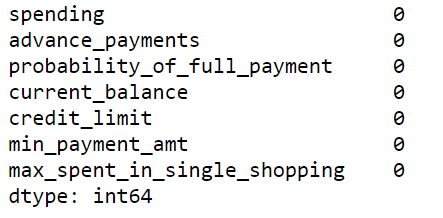
probability\_of\_full\_payment non-null

current\_balance non-null

credit\_limit non-null

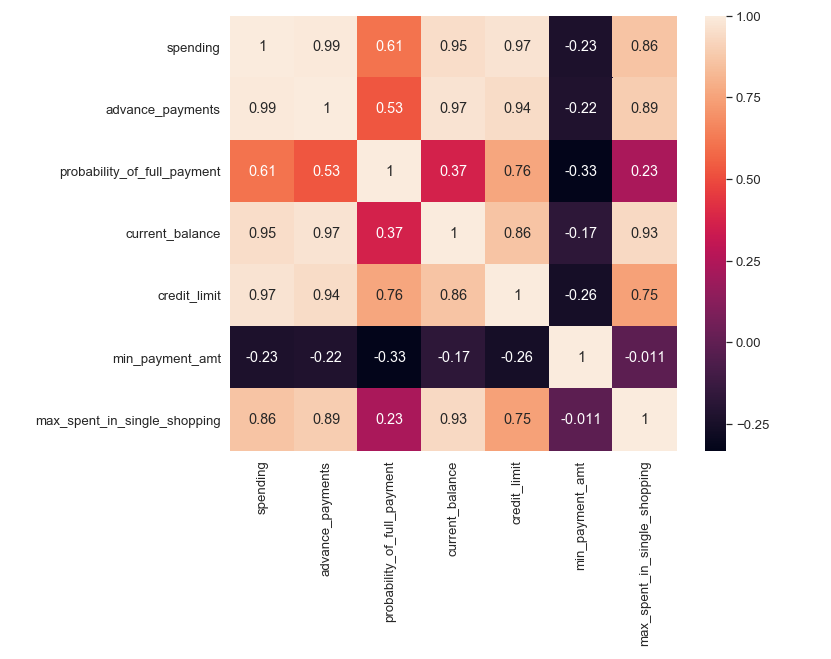
min\_payment\_amt non-null

max\_spent\_in\_single\_shopping non-null



From the above results we can see that there is no missing value present in the dataset.

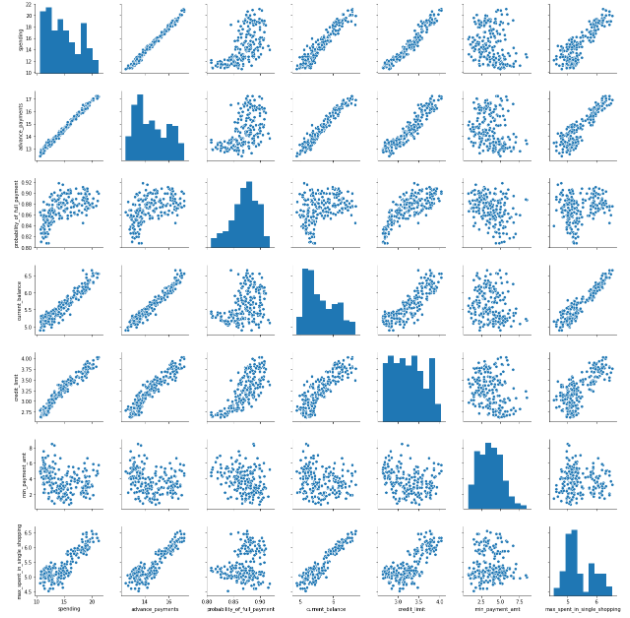
Correlation Plot



From the correlation plot, we can see that various attributes of the customers are highly correlated to each other. Correlation values are almost highly positively correlated. Correlation values near to 0 are not correlated to each other.

Pairplot

Pairplot shows the relationship between the variables in the form of scatterplot and the distribution of the variable in the form of histogram.



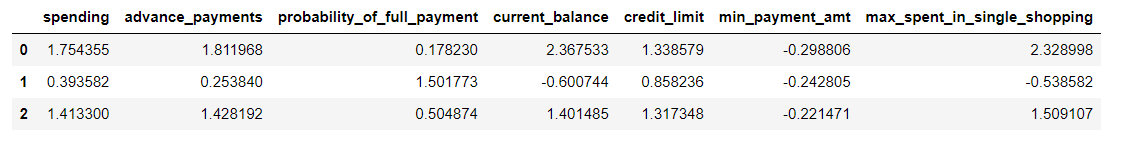
There are no duplicates in this data set.

Do you think scaling is necessary for clustering in this case? Justify

Most clustering algorithms are even highly sensitive to scaling. standardization prevents variables with larger scales from dominating how clusters are defined. It allows all variables to be considered by the algorithm with equal importance.

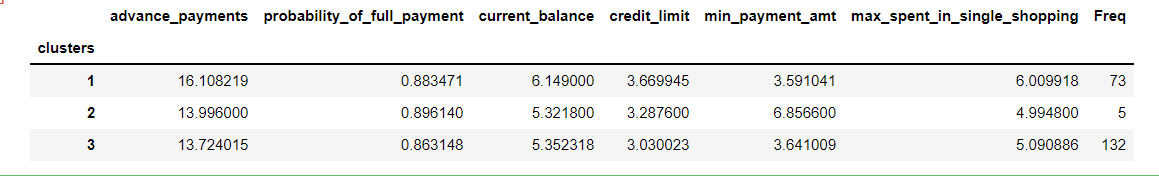
As one column of probability\_of\_full\_payment was in different unit we need to standardise it to get in cluster with other parameter’s hence, scaling is done.

Data after scaling



Hierarchical clustering to scaled data –

Hierarchical clustering, also known as hierarchical cluster analysis, is **an algorithm that groups similar objects into groups called clusters**. The endpoint is a set of clusters, where each cluster is distinct from each other cluster, and the objects within each cluster are broadly similar to each other.



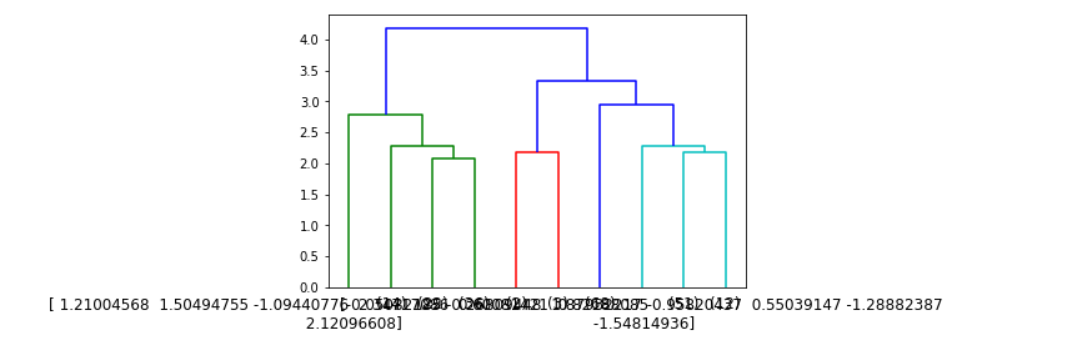
After applying Hierarchical clustering we can conclude -

Cluster 1: Medium performing client

Cluster 2: Poor performing client

Cluster 3: Top performing client

Drendogram – Used Average as link method for dendrogram. In Average linkage clustering, the distance between two clusters is defined as the average of distances between all pairs of objects, where each pair is made up of one object from each group.



Apply K-Means clustering on scaled data and determine optimum clusters. Apply elbow curve and silhouette score. Explain the results properly. Interpret and write inferences on the finalized clusters.

K-means algorithm **identifies k number of centroids, and then allocates every data point to the nearest cluster**, while keeping the centroids as small as possible.

For this data set optimum clusters is 3 as the WWS keeps dropping when we keep adding 1 cluster to the data

KM intertia(wss) – WSS means **the sum of distances between the points and the corresponding centroids for each cluster**

At 1 Cluster - 1470.0,

At 2 Cluster - 659.1717544870407,

At 3 Cluster -430.65897315130053,

AT 4 Cluster - 371.301721277542,

At 5 Cluster - 327.35442092981316,

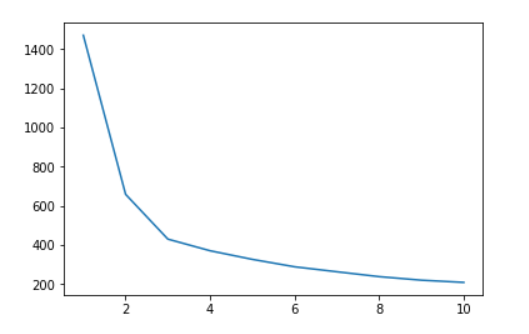
At 6 Cluster -289.4518051947612,

At 7 Cluster -264.14673804820933,

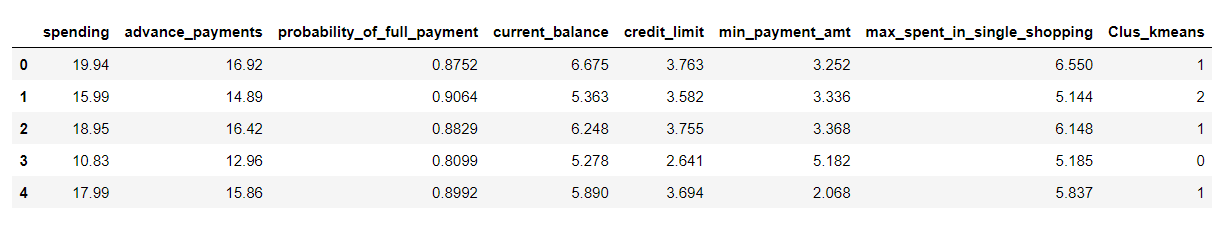
At 8 Cluster -239.28181527825586,

At 9 Cluster -221.18536872961658,

At 10 Cluster - 209.98269562947308



Data after adding cluster column.



**silhouette\_score** (**sklearn**. metrics) for the data set is used for measuring the mean of the Silhouette Coefficient for each sample belonging to different clusters. The Silhouette Index measure the distance between each data point, the centroid of the cluster it was assigned to and the closest centroid belonging to another cluster. ... For instance, the silhouette index is normalized and **a value close to 1 is always good**

silhouette\_score - 0.4007270552751299 indicates is **average score**

Cluster 0: Least spending group

Cluster 1: Premium spending group

Cluster 2: mediocure spending group

Describe cluster profiles for the clusters defined. Recommend different promotional strategies for different clusters.

**Profiling** involves generating descriptions of the **clusters** with reference to the input variables you used for the **cluster** analysis. **Profiling** acts as a class descriptor for the **clusters** and will help you to 'tell a story' so that you can understand this information and use it across your business.