AREIX

# Technical Assessment Report

## Part I: Explanation of Clustering Model

| | |
|---|---|
| **Prepared by:** | **Li Lok Hang, Alan** |
| **Date:** | **27/05/2021** |

# Methodology

## Data Preprocessing

In `*preprocess.py*`, various fields from the three given CSVs are selected and joined on the same ‘psid’. There are different currencies recorded. For easier arithmetics, money-related values are all converted to equivalent values in terms of the Hong Kong dollar (HKD). The following table shows the meaning of the extracted fields.

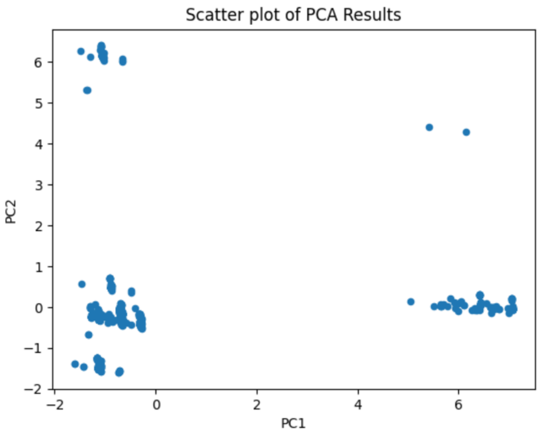
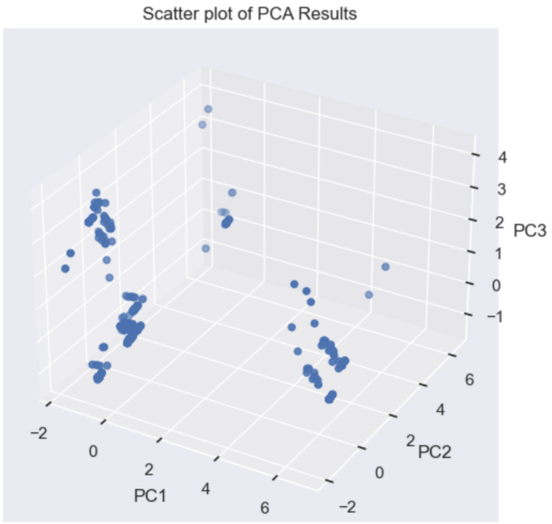
| **Features** | **Descriptions** |
|---|---|
| Dining & Beverage | Total transaction sum in category ‘Dining & Beverage’ for a given ‘psid’ |
| Income | Total transaction sum in category ‘Financials’ for a given ‘psid’ that is an “Income” |
| Investment Income | Total transaction sum in category ‘Financials’ for a given ‘psid’ that is **not** an “Income” |
| Healthcare | Total transaction sum in category ‘Healthcare’ for a given ‘psid’ |
| Home | Total transaction sum in category ‘Home’ for a given ‘psid’ |
| Leisure | Total transaction sum in category ‘Leisure’ for a given ‘psid’ |
| Others | Total transaction sum in category ‘Others’ for a given ‘psid’ |
| Shopping | Total transaction sum in category ‘Shopping’ for a given ‘psid’ |
| Transportation | Total transaction sum in category ‘Transportation’ for a given ‘psid’ |
| bal_sum | Sum of all account balances for a given ‘psid’ |
| saved_amount_sum | Sum of field ‘saved_amount’ for a given ‘psid’ |

Subsequently, the preprocessed CSV containing the above field columns and values is saved and carried to the next stage. Features are then normalized by removing their mean and scaling to unit variance. This is essential to keep all features having the same standard deviation in our subsequent steps, namely: principal component analysis (PCA) and clustering. In PCA, we do not want the relative importance of a feature distorted by its high standard deviation. In clustering, many clustering methods are based on distance metrics, and thus to be fair, having a normalized feature is necessary.

## Model Development

We observe that our data matrix has a high number of dimensions. We aim to simplify the problem to a 3D clustering problem for easier visualization and interpretation.

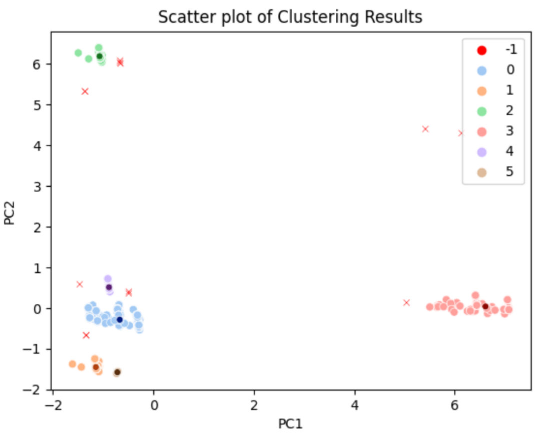
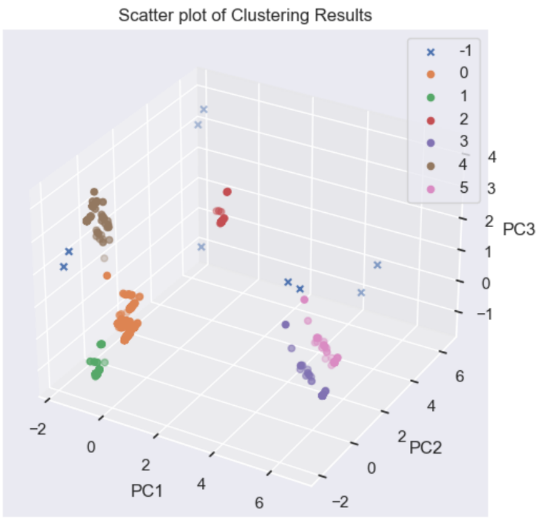
Principal component analysis (PCA), based on eigendecomposition of a high dimensional data matrix, is a commonly used dimensionality reduction algorithm. We perform PCA on the data matrix and reduce the number of features into only a few principal components, where each principal component is a linear combination of features. In general, the explained variances of the model are higher, when there are more retained principal components. However, visualization is very difficult for dimensions more than 3.

| **2D Scatter Plot** | **3D Scatter Plot** |
|---|---|
|  |  |

At this stage, we have successfully transformed the multidimensional problem into a 3D clustering problem. We noticed that there are some outliers in the above scatter plot. Thus, DBSCAN[1] is a suitable candidate algorithm for the clustering procedure. In short, DBSCAN is a robust clustering method that allows noises or outliers that do not belong to any cluster.

[1] sklearn.cluster.DBSCAN Documentations. scikit. (n.d.). https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html.

DBSCAN clusters by defining core points and reachable points. A core point has at least *min_samples* points within its neighbourhood with radius *eps*. The aforementioned *min_samples* and *eps* are parameters that can be tuned to yield different results. We perform a brute-force algorithm, namely “grid search cross-validation[2]” to search through different combinations of these parameters, in order to find the best set of parameters. For the 3D plot, it is found that the *silhouette coefficient* score (Explained in the next section) is maximized at a value of 0.703 when *min_samples* = 9 and *eps = 0.9*.

| **2D Scatter Plot** | **3D Scatter Plot** |
|---|---|
|  |  |

It is clear that our DBSCAN model is capable of characterizing the observations from the given dataset into 6 clusters. Crosses are noises that do not belong to any cluster.

The next step is to grant human-understandable meanings to the clustering results. We can represent the characteristics of a cluster by its centroid, spread and size. We have to inversely transform the coordinates of the centroid of each cluster from (PC1, PC2) coordinates back to features before PCA, and un-normalize it. The following table shows the inverse transformed features of the centroids of clusters.

[2] sklearn.model_selection.GridSearchCV Documentations. scikit. (n.d.). https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html.

These numbers are still too complicated to read. The next step is to convert these into categories of features.

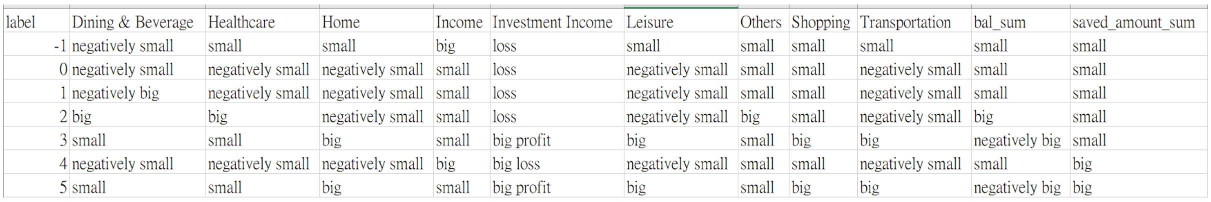
### Model Scoring

Since there is no ground truth of clustering labels, we cannot use common metrics or verification methods, such as F1 score or a contingency table. However, some statistics can still be derived to quantify the goodness of clustering. We use the mean of the silhouette coefficient[3] for all samples as the model score. In theory, the best score is 1 and the worst score is -1. The coefficient is based on intra-cluster distance and nearest-cluster distance. For example, for a very poor model, the intra-cluster distance has an order of magnitude similar to nearest-cluster distance, and hence clusters would be overlapped and ambiguous.

Our model has a score of around 0.7, which is acceptable since there exists some non-clustered outliers. Including more principal components may enhance the score in exchange of poorer model visualizability due to higher dimensions.

## Results

‘interpret.csv’ contains categorized more human-understandable results of the clustering analysis. Label “-1” refers to outliers and can be neglected. Other labels are the categorized values of the centroids of the clusters. For example, cluster group “3” typically has a big profit in investment, big expenditure in category “leisure”, “home”, “shopping” and "transportation”, big credit in account balance and big saved amount. Similar conclusions can be drawn to different clusters, given by the following image.



| label | Dining & Beverage | Healthcare | Home | Income | Investment Income | Leisure | Others | Shopping | Transportation | bal_sum | saved_amount_sum |
|---|---|---|---|---|---|---|---|---|---|---|---|
| -1 | negatively small | small | small | big | loss | small | small | small | small | small | small |
| 0 | negatively small | negatively small | negatively small | small | loss | negatively small | small | small | negatively small | small | small |
| 1 | negatively big | negatively small | negatively small | small | loss | negatively small | small | small | negatively small | small | small |
| 2 | big | big | negatively small | small | loss | negatively small | big | small | negatively small | big | small |
| 3 | small | small | big | small | big profit | big | small | big | big | negatively big | small |
| 4 | negatively small | negatively small | negatively small | big | big loss | negatively small | small | small | negatively small | small | big |
| 5 | small | small | big | small | big profit | big | small | big | big | negatively big | big |

The numerical version of the above results can be found in ‘unscale.csv’, which contains the exact value these categories correspond to.

[3] sklearn.metrics.silhouette_score Documentations. scikit. (n.d.). https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html.