

**Applied Analytics Assignment**

Diploma in Cybersecurity & Digital Forensics

Diploma in Financial Informatics

Diploma in Information Technology Year 2/3 (2024), Semester 3/5

**INDIVIDUAL ASSIGNMENT 1**

(30% of Applied Analytics Module)

**Deadline for Submission:**

**01 Jun 2024 (Sat), 23:59 HRS**

|  |  |
| --- | --- |
| Tutorial Group: | P06 |
| Student Name: | Clayton Thia |
| Student Number: |  |

**Penalty for late submission:** 10% of the marks will be deducted every day after the deadline. **NO** submission will be accepted after 08 Jun 2024, 23:59.

Contents

[Overview 3](#_Toc168177533)

[1 Building clustering models 4](#_Toc168177534)

[1.1 Data exploration and manipulation 4](#_Toc168177535)

[1.1.1 Univariate Analysis 4](#_Toc168177536)

[1.1.2 Multivariate 9](#_Toc168177537)

[1.1.3 Missing values 11](#_Toc168177538)

[1.1.4 Outlier handling 12](#_Toc168177539)

[1.1.5 Categorical encoding 13](#_Toc168177540)

[1.1.6 Scaling 14](#_Toc168177541)

[1.2 K-Means clustering 15](#_Toc168177542)

[1.2.1 Effect of outliers 15](#_Toc168177543)

[1.2.2 K-Means – Financial Stability 16](#_Toc168177544)

[1.2.3 K-Means – Financial Strain 21](#_Toc168177545)

[1.3 Hierarchical clustering 25](#_Toc168177546)

[1.3.1 Agglomerative clustering - Financial Stability 25](#_Toc168177547)

[1.3.2 Agglomerative clustering - Financial Strain 29](#_Toc168177548)

[1.4 Comparison of clustering models 33](#_Toc168177549)

[1.4.1 Silhouette score comparison 33](#_Toc168177550)

[2 Summary and interpretation 34](#_Toc168177551)

[2.1 K-Means: Financial stability 34](#_Toc168177552)

[2.1.1 Interpretation 35](#_Toc168177553)

[2.2 K-Means: Financial strain 37](#_Toc168177554)

[2.2.1 Interpretation 38](#_Toc168177555)

[2.3 Summary of key findings from data exploration 40](#_Toc168177556)

# Overview

In the banking sector, unsupervised learning, particularly clustering algorithms like K-Means and Hierarchical Clustering, plays a crucial role in customer segmentation, fraud detection, risk management, and personalized marketing.

Customer Segmentation is a significant application where clustering can provide deep insights. By grouping customers based on their transaction behaviors, account activities, credit histories, and demographic information, banks can tailor their products and services to meet the specific needs of different segments. For instance, young professionals might benefit from investment advice and loan offers, while retirees might be more interested in savings plans and wealth management services. This targeted approach enhances customer satisfaction and loyalty, driving business growth.

In this assignment, I will perform data exploration, cleaning, and manipulation to build clustering models to help a bank segment different types of customers. I will then provide actionable recommendations for the bank to support these customers.

# Building clustering models

## Data exploration and manipulation

A screenshot of a computer

Description automatically generated

After looking at some sample rows in the dataset, I determined that each row of the dataset does not represent bank accounts, but rather, loans and the attributes of its applicants.

### Univariate Analysis

To view the distribution of numerical columns, I plotted boxplots. To view the distribution of categorical columns, I plotted their frequency count. This allowed me to understand and profile my data by knowing what kind of values can be found in each column.

#### Numerical

|  |  |  |
| --- | --- | --- |
| **Column name** | **Type** | **Description** |
| Duration in month | Continuous | Right skewed with outliers. Most likely represents the duration of a loan. |
| Credit amount | Continuous | Right skewed with outliers. Most likely represents the amount of credit or loan the customer has taken. |
| Instalment rate in percentage of disposable income | Discrete | How much of the customer’s disposable income is used to pay instalments for the loan. |
| Present residence since | Discrete | How long the customer has been residing at their current address. |
| Age in years | Continuous | Right skewed with outliers. Represents age of the customer. |
| Number of existing credits at this bank | Discrete | Represents how many credits or loans the customer has at the bank. |
| Number of people being liable to provide maintenance for | Discrete | Indicates number of dependents the customer has. |
| Target | Discrete | Binary values representing “Good” or “bad”, however it is unclear as to the definition of this column. |

#### Categorical

|  |  |  |
| --- | --- | --- |
| **Column name** | **Hierarchy** | **Description** |
| Account Status | Duration | Indicates how long the account has been receiving salary assignments in years. |
| Credit History | Creditworthiness | Represents customer behaviour in terms of loan repayment punctuality. |
| Purpose |  | Purpose for customer taking a loan. |
| Savings account or bonds |  | Only contains one value. |
| Present employment since | Job stability | Shows current employment status of customer, and how long they have been employed for. |
| Personal status and sex |  | Gender & marital status of customer. |
| Other debtors or guarantors |  | How many people the customer has that can cover the loan repayment for the customer in case of unforeseen circumstance. |
| Property |  | Property the customer has, like cars or real estate. |
| Other instalment plans |  | Indicates if the customer has other plans to buy for eg. For consumer products. |
| Housing | Financial stability | Indicates whether the customer owns or rents their house as well as if they are living free of charge. |
| Job | Skill | Indicates skill level of customer in their occupation. |
| Telephone |  | Indicates whether the customer has a telephone. |
| Foreign worker |  | Indicates if customer is a foreign worker. |

#### Common themes in columns

When performing data exploration, I noticed two common themes appearing in the different columns.

|  |  |  |
| --- | --- | --- |
| **Theme** | **Involved columns** | **Reasoning** |
| **Financial stability** | Credit amount | Higher credit amounts are often extended to individuals who have demonstrated financial responsibility and creditworthiness. This means that the higher the credit amount that the customer can apply for, the better the customer’s financial stability. |
| Account Status | Having a checking account suggests that the customer has a regular income flow, implying financial stability. |
| Credit History | Indicates financial stability through repayment behaviour. Punctuality in repayment will indicate financial discipline and stability. |
| Present employment since | The longer a customer has been employed for, implies better job stability that leads to stronger financial stability. |
| Housing | If customers own/rent/live rent-free in houses, it could signify financial stability/ lack of long-term financial commitment/ dependency. |
| Job | Different jobs indicate varying levels of income and job security, and thus financial stability. |
| **Financial strains** | Instalment rate in percentage of disposable income | A higher percentage indicates that less disposable income is available for other expenses, which may lead to financial strain. |
| Number of existing credits at this bank | Multiple existing credits may suggest higher debt levels and potential difficulties in managing multiple repayments for the multiple credits. |
| Number of people being liable to provide maintenance for | A higher number of dependents means that the customer has to allocate a larger portion of their income to support others, reducing their disposable income and increasing financial stress. |
| Other instalment plans | The presence of other instalment plans indicates additional financial commitments. |

I will use these themes for feature selection later, for my clustering models.

### Multivariate

I plotted a correlation heatmap to view any obvious relationships between columns in the data. This is because features with collinearity should not be used together for clustering as high correlations can distort distance-based metrics that models like K-Means use.

A screenshot of a computer

Description automatically generated

‘Credit amount’ has a strong positive correlation with ‘Duration in month’, and is somewhat negatively correlated with ‘Instalment rate in percentage of disposable income’.

‘Age in years’ and ‘Present residence since’ has a somewhat positive correlation with each other.

I keep in mind not to use these features together when clustering later.

Additional notes

* I regrouped the various values for ‘Purpose’ into more generic categories.
* Dropped ‘Savings account or bonds’ due to it having only one unique value.
* Extracted gender information from ‘Personal status and sex’.

### Missing values

A screenshot of a computer screen

Description automatically generated

The above image shows that all columns do not host any null values. Earlier data exploration also revealed that columns don’t contain a null-representing value. Hence, no missing value imputation is performed.

### Outlier handling

Univariate analysis of numerical columns revealed 3 continuous columns that contained outlier values: ‘Credit amount’, ‘Duration in month’, and ‘Age in years.’ These outliers were handled via numerical transformations which were chosen based on how many outliers they handled.

|  |  |
| --- | --- |
| **Column** | **Transformation** |
| Duration in month | Yeo Johnson |
| Credit amount | Logarithmic |
| Age in years | Reciprocal |

As I was unsure as to how outliers affected clustering, the transformed values were stored in separate columns for later use to test the effect of outliers.

### Categorical encoding

I performed ordinal encoding for features that represent financial stability or strain as they had hierarchy, while leaving columns without a hierarchy untouched. This is because ordinal encoding allows preservation of the intrinsic order in categorical columns while converting it to numerical values.

Despite low cardinalities in columns, one-hot encoding was not due to K-means clustering relying on Euclidean distance. One-hot encoded features distorts these distances as all categories become equidistant from each other, lowering K-means performance.

|  |  |
| --- | --- |
| Column | Mapping |
| Account Status |  |
| Credit History |  |
| Present employment since |  |
| Housing |  |
| Job |  |
| Other instalment plans |  |

### Scaling

I used standard scaling on my data before feeding it to the K-means model to ensure that all features have a mean of zero and a standard deviation of one. This normalization helps the algorithm treat all features equally, preventing features with larger scales from disproportionately influencing the clustering results.

A graph of a diagram

Description automatically generated with medium confidence

## K-Means clustering

The K-Means models will aim to strike a balance between inertia and silhouette scores, as well as the number of clusters for interpretability.

### Effect of outliers

As previously mentioned in section 1.1.4 outlier handling, the effect of outliers are to be tested. This was done by comparing the inertia, silhouette, and visual interpretability of a K-Means model without outlier handling, and one with.

A chart of data clusters

Description automatically generated with medium confidenceA diagram of a number of dots

Description automatically generated with medium confidence

For the data with outliers, K-Means can segment most of the data points from the purple cluster away from the other 2 clusters. Without outliers, no clusters are properly segregated and there is inter-cluster distance between any the clusters.

A comparison of blue bars

Description automatically generated with medium confidence

Additionally, K-Means performs better with outlier values than without. This is evident by the higher inertia and silhouette scores. To summarize, clustering on this dataset should be done with the presence of outliers.

### K-Means – Financial Stability

Using the columns with the underlying theme of financial stability, I made a correlation heatmap and dropped features with relatively strong correlations.

A screenshot of a graph

Description automatically generated

* Dropped “Job” column

A K-Means model with initial cluster number of 3 was then initialized and fitted onto the data.

#### Pre-improvement Visualizing & Evaluation

Visualizing clusters was mainly done with t-dstributed Stochastic Neighbor Embedding (t-SNE) as t-SNE is able to effectively reduce dimensionality of high dimensionality data while preserving similarities, and also allows to project data points into 2D or 3D making it easier to analyse clusters. While I am unable to extract what components the dimensionality was reduced to, I can still use it to evaluate clusters as I would on a 2D visualization.

A graph showing different colored dots

Description automatically generated

|  |  |
| --- | --- |
| **Inertia** | **Silhouette** |
| ~3437.2274 | ~0.2056 |

Clusters have bad cohesion, but cluster separation is somewhat decent as there are very few overlapping data points between the different clusters. Cluster centroids can also be easily estimated by looking at the visual.

No cluster has an overwhelming amount of data points.

#### Determining optimal number of clusters

A graph of a number of clusters

Description automatically generated

I determined the optimal number of clusters to be 5 due to its balance between SSE, Silhouette score, and interpretability.

I came to this conclusion by using the elbow method for SSE and looking for the bend, which is at 5 clusters. I then looked at the silhouette score and tried to balance the number of clusters with the silhouette score. This is how I determined the optimal number of clusters for all my K-Means models.

#### Post-improvement Visualizing & Evaluation

A graph showing different colored dots

Description automatically generated

*Pre-improvement*

A graph showing different colored dots

Description automatically generated

*Post-improvement*

|  |  |  |
| --- | --- | --- |
| Pre / post | Inertia | Silhouette |
| Pre | ~3437.2274 | ~0.2056 |
| Post | ~2614.3357 | ~0.2291 |

Metrics have improved.

While cohesion of clusters has improved slightly, they are still quite bad. Clusters are still well separated with the presence of overlap between 1 or 2 clusters. Cluster centroids can still be easily estimated by looking at the visual.

No cluster has an overwhelming amount of data points.

### K-Means – Financial Strain

I plotted a correlation heatmap for features with underlying theme of financial burden.

A screenshot of a computer

Description automatically generated

* No columns were dropped.

K-Means with 3 clusters was then initialized and fitted on the data containing the above columns.

#### Pre-improvement Visualization & Evaluation

A graph showing different colored dots

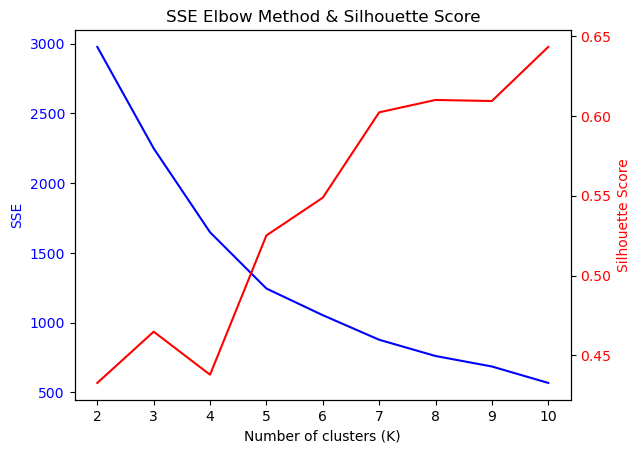
Description automatically generated

|  |  |
| --- | --- |
| **Inertia** | **Silhouette** |
| ~2249.4805 | ~0.4649 |

Cohesion of yellow and blue clusters are somewhat alright, but the cohesion of the peach cluster is bad. Clusters are well separated with little to no overlap, and centroids can be easily estimated.

Peach cluster has a lot more data points than the other 2 clusters.

#### Determining optimal number of clusters



I determined the optimal number of clusters to be 5 due to its balance between SSE, Silhouette score, and interpretability.

#### Post-improvement Visualizing & Evaluation

A graph showing different colored dots

Description automatically generated

*Pre-improvement*

A graph showing different colored dots

Description automatically generated

*Post-improvement*

|  |  |  |
| --- | --- | --- |
| Pre / post | Inertia | Silhouette |
| Pre | ~2249.4805 | ~0.4649 |
| Post | ~1244.5512 | ~0.5252 |

Metrics have improved.

Cohesion of clusters are still the same except for what was once the peach cluster (yellow now), who’s non-cohesive data points were broken down into 2 clusters. Separation of clusters remains the same and cluster centroids can still be easily estimated by looking at the visual.

The purple and blue clusters have less than average number of data points.

## Hierarchical clustering

Hierarchical models will be fed the same subset data as their K-Means counterparts to ensure fair comparison. The main and only metric used for evaluation is Silhouette score as hierarchical clustering does not support inertia scores.

### Agglomerative clustering - Financial Stability

A graph of a diagram

Description automatically generated with medium confidence

Using a dendrogram generated from ‘ward’ linkage method, I determine the number of clusters to initialize my Agglomerative clustering model with – 5.

#### Pre – improvement Visualization & Evaluation

A graph showing different colored spots

Description automatically generated

|  |  |
| --- | --- |
| Silhouette | ~0.1755 |

Cohesion and separation of clusters are horrible, with some clusters having very high intra-cluster distance and a lot of overlap between many clusters’ data points. Centroids cannot be easily estimated.

There are no clusters with an overwhelming amount of data points.

#### Determining optimal number of clusters

A graph of a number of blue bars

Description automatically generated

By plotting silhouette score against cluster number, I determine that 4 is the optimal number of clusters as it maximizes the silhouette score and interpretability of the clusters.

#### Post-improvement Visualizing & Evaluation

A graph showing different colored spots

Description automatically generated

*Pre-improvement*

A graph showing different colored spots

Description automatically generated with medium confidence

*Post-improvement*

|  |  |
| --- | --- |
| Pre silhouette | ~0.1755 |
| Post silhouette | ~0.1962 |

Cohesion of most clusters are somewhat acceptable, except for the presence of outliers and the purple cluster. Separation of clusters have improved, but there is still a presence of overlap. Centroids be estimated with a very rough estimation.

There are no clusters with an overwhelming amount of data points.

### Agglomerative clustering - Financial Strain

A graph with different colored lines

Description automatically generated

Using a dendrogram generated from ‘ward’ linkage method, I determine the number of clusters to initialize my Agglomerative clustering model with – 5.

#### Pre – improvement Visualization & Evaluation

A graph showing different colored dots

Description automatically generated

|  |  |
| --- | --- |
| Silhouette | ~0.5166 |

Cohesion of clusters are good, with presence of a few outliers in 3 clusters of 5 clusters having outliers. Clusters are well separated with little to no overlap. Centroids can be easily estimated.

There is one cluster with very little number of data points, which is very well separated from the other clusters (yellow cluster on the left-side of the chart).

#### Determining optimal number of clusters

A graph of a number of blue bars

Description automatically generated

By plotting silhouette score against cluster number, I determine that 5 is the optimal number of clusters as it maximizes by silhouette score and interpretability. However, since my model already has 5 clusters, I try to address the problem of an under-represented cluster by lowering the number of clusters to 3.

#### Post-improvement Visualizing & Evaluation

A graph showing different colored dots

Description automatically generated

*Pre-improvement*

A graph with many colored dots

Description automatically generated

*Post-improvement*

|  |  |
| --- | --- |
| Pre silhouette | ~0.5166 |
| Post silhouette | ~0.4279 |

Cohesion of orange cluster is good, followed by yellow and purple cluster. Yellow and purple clusters have quite a few outliers. Centroids can be estimated. There are no clusters with an overwhelming amount of data points.

The underrepresented clusters were merged with the other clusters. I decide to stick with the pre-improvement model, as I felt that it would be better to cluster outliers by themselves instead of allowing them to distort anything else.

## Comparison of clustering models

### Silhouette score comparison

A comparison of blue bars

Description automatically generated with medium confidence

From figure, we can tell that in both cases K-Means clustering outperforms Hierarchical clustering in terms of Silhouette score for themes of Financial Stability and Financial Strain. Hence, I will be interpreting the K-Means clusters for both themes.

# Summary and interpretation

I also added columns representing general demographic data for further interpretation.

## K-Means: Financial stability

A group of graphs on a white background

Description automatically generated

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Cluster 0** | **Cluster 1** | **Cluster 2** | **Cluster 3** | **Cluster 4** |
| **Account Status** | High | Low | Medium - High | Mostly Low - medium | Low & Medium |
| **Credit History** | Semi low - Medium | Medium | Low-Medium | Low | Semi low - Medium |
| **Present employment since** | Medium | Medium | High | High | Medium |
| **Credit amount** | Low | Low | Low-Medium | Low | High |
| **Housing** | High | High | Low | High | Medium - High |
| **Age in years** | Low | Low | Medium | Low - Medium | Low |
| **Purpose** | Mostly automobile + electronics | Automobile, Electronics, Home | Mainly automobile | Automobile, Electronics | Automobile |

### Interpretation

|  |  |  |
| --- | --- | --- |
| Cluster | Name | Interpretation |
| 0 | Stable high earners | Customers who have a stable income and have been receiving salary assignments for a long period. Their repayment behaviour is relatively good, though not perfect. They tend to take out lower amounts of loans and are likely to own their homes. These customers are generally younger and often take loans for automobiles and electronics. |
| 1 | New employees with low income | Newer entrants into the job market with lower income levels. Their credit history is average, indicating a balanced repayment behaviour. They have been employed for a moderate period and tend to take out small loans. Most of them own their homes despite their lower income. These customers are young and take loans for a variety of purposes including automobiles, electronics, and home-related expenses. |
| 2 | Established medium earners | Customers with medium to high income who have an established employment history. Their credit history is slightly below average, and they take out moderate amounts of loans. They are less likely to own their homes and fall into the middle age category. The primary purpose of their loans is to finance automobiles. |
| 3 | Low income, long-term employed | Customers with low to medium income levels who have a long-term employment history. Their credit history is poor, indicating potential repayment issues. They tend to take out small loans and are likely to own their homes. Their age ranges from young to middle-aged, and they primarily take loans for automobiles and electronics. |
| 4 | High credit, mixed employment stability | Their repayment behaviour is usually poor, and they have an average employment history. They take out larger loan amounts and are likely to own their homes or have a high housing status. These customers are younger and primarily take loans for automobiles. |

#### Recommendations

Stable high earrners

* Offer premium banking services and personalized financial planning to maximize their wealth and build rapport with such customers as they are most likely to generate the most revenue.

New employees with low income

* Provide financial literacy programs to help them manage their finances better and result in better repayment behaviour.
* Equip them with budgeting tools and savings apps to encourage prudent financial management.

Established medium earners

* Promote long-term savings plans and retirement accounts since they are middle-aged, have enough savings, and should start planning for their retirement.

Low income, long-term employed

* Implement programs focused on credit repair and financial stability.
* Create emergency fund savings accounts with favourable terms to encourage savings as they are low-income

High credit, mixed employment stability

* Create incentive programs for timely loan repayments to improve credit scores.

## K-Means: Financial strain

A screenshot of a graph

Description automatically generated

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Cluster 0 | Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 |
| Instalment rate in % of disposable income | Low | High | Semi-Low to High | Semi-Low to High | Medium to High |
| Number of existing credits at this bank | Low | Low | Low to semi-low | Low to semi-low | Semi-low |
| Other instalment plans | None | None | Banks | None | None |
| Number of people being liable to provide maintenance for | None | None | None | High | None |
| Age in years | Low - medium | Low-medium | Low-medium | Medium | Low-medium |
| Purpose | Automobiles, Electronics, Home | Electronics, Automobiles | Automobiles | Automobiles | Automobiles, Electronics |

### Interpretation

|  |  |  |
| --- | --- | --- |
| Cluster | Name | Interpretation |
| 0 | Low financial strain, diverse needs | Customers who experience low financial strain. They allocate a small portion of their disposable income to paying instalments, have few existing loans, and no additional instalment plans elsewhere. They have no dependents and are generally younger. These customers take loans for a variety of purposes, including automobiles, electronics, and home-related expenses. |
| 1 | High financial strain, focused needs | Customers experiencing high financial strain. A significant portion of their disposable income goes towards paying instalments, but they have a low number of existing loans and no additional instalment plans elsewhere. They have no dependents and are relatively young. |
| 2 | Moderate to high financial strain with other instalment plans | Customers facing moderate to high financial strain. They allocate a variable portion of their disposable income to instalments, have a low to semi-low number of existing loans, and additional instalment plans from banks. They have no dependents and are relatively young. |
| 3 | High financial strain with dependents | Customers who are under high financial strain. They spend a semi-low to high portion of their disposable income on instalments, have a low to semi-low number of existing loans, and no additional instalment plans elsewhere. They have a high number of dependents and are generally middle-aged. Their primary purpose for taking loans is for automobiles. |
| 4 | Moderate financial strain, young and diverse needs | Customers with moderate financial strain. They allocate a medium to high portion of their disposable income to instalments, have a semi-low number of existing loans, and no additional instalment plans elsewhere. They have no dependents and are relatively young. These customers take loans primarily for automobiles and electronics. |

#### Recommendations

Low financial strain, diverse needs

* Provide low-interest loans or special offers for customers with low financial strain to incentivize borrowing from the bank. Since they are not financially stressed, they would be more likely to borrow more from the bank.

High financial strain, focused needs

* Provide financial counselling, or budgeting tools and resources to help customers manage their financial stress or track their expenses and prioritize loan repayments.

Moderate to high financial strain with other instalment plans

* Since these customers have other existing debts to pay other than the one to the bank, the bank should assist customers in creating debt management plans to pay off existing loans and reduce financial strain over time.

High financial strain with dependents

* Additional dependents usually come in the form of family members. The bank should offer family financial planning services to help customers manage their finances effectively while providing for their dependents.

Moderate financial strain, young and diverse needs

* Offer savings incentives and rewards programs to encourage young customers to save and invest for their future.

## Summary of key findings from data exploration

* Rows represent loans and the attributes of their applicants, not bank accounts
* Some columns that would usually be continuous in the real-world are discrete
  + E.g. Instalment rate in % of disposable income
* Quite a few categorical columns have hierarchy
* Some categorical columns can get too specific or general
  + E.g. personal status & sex
* There are underlying themes behind the columns given like financial stability or financial strain which I used to perform feature selection for my clustering models
* Definition of target column is vague