Utilizing Clustering Algorithms to Perform Market Segmentation through Credit Card Analysis

Introduction

Market segmentation is a fundamental component of successful marketing strategies. Companies aim to divide their customer base into distinct groups with shared characteristics, allowing them to target their marketing campaigns more effectively. One way to segment customers is through credit card spending patterns. By analyzing credit card usage, companies can develop a better understanding of their customer's behaviors and preferences.

Several studies have examined credit card customer segmentation using data mining and machine learning techniques. In "The Future of Credit Cardholder Segmentation," the authors discussed the future of credit cardholder segmentation, predicting that segmentation would become increasingly sophisticated and used to target customers with greater precision. In "A novel approach to credit card customer segmentation using data mining and machine learning techniques," the authors used a unique approach that combined data mining and machine learning techniques to segment credit card customers into different groups based on risk profiles, spending habits, and needs. In "Credit card customer segmentation using a hybrid of clustering and association rule mining," the authors employed a hybrid approach that combined clustering and association rule mining to segment credit card customers into different groups based on spending habits, allowing companies to develop targeted marketing campaigns.

Problem Statement

The problem statement of the project is to perform market segmentation for a credit card company based on customer data such as demographic information, credit card usage behavior, and payment patterns. The objective of this project is to identify distinct groups of customers with similar characteristics and behaviors in order to better understand customer needs, target specific customer segments with tailored marketing strategies, and ultimately increase customer loyalty and profitability for the credit card company. Specifically, this project involves importing and cleaning the customer data, performing exploratory data analysis, and using clustering algorithms to segment the customer base.

This project employs unsupervised learning techniques to cluster customers based on their credit card spending patterns. Unlike traditional supervised learning, unsupervised learning does not rely on a set response variable to predict outcomes. Instead, it seeks to identify patterns and relationships within the data without any prior knowledge of what those patterns may be.

Method

The success of any data-driven project relies heavily on the quality of data used, and the credit card clustering project is no exception. The project involves several critical steps to prepare the data before applying unsupervised learning techniques. The initial step involves data cleaning and preparation, which includes removing duplicates, dealing with missing values, and transforming variables as necessary. Once the data is cleaned, Exploratory Data Analysis (EDA) is performed to gain insights into the data. Principal Component Analysis (PCA) is then employed to preprocess the data, visualize the data distribution, and determine the most important features. PCA helps in reducing the dimensionality of the data, which can help improve clustering accuracy. The project utilizes different types of clustering algorithms to cluster customers based on their credit card spending patterns. By applying multiple clustering algorithms, the project ensures a thorough exploration of the data to identify the most meaningful clusters.

Exploratory Data Analysis

The project begins with a dataset comprising over 8,000 records and 17 credit card attributes. The initial step involves data wrangling by removing null values and ensuring that values are identified as numeric and not strings. Exploratory data analysis (EDA) is then performed, and the results are presented in Fig 1 and Table 1.

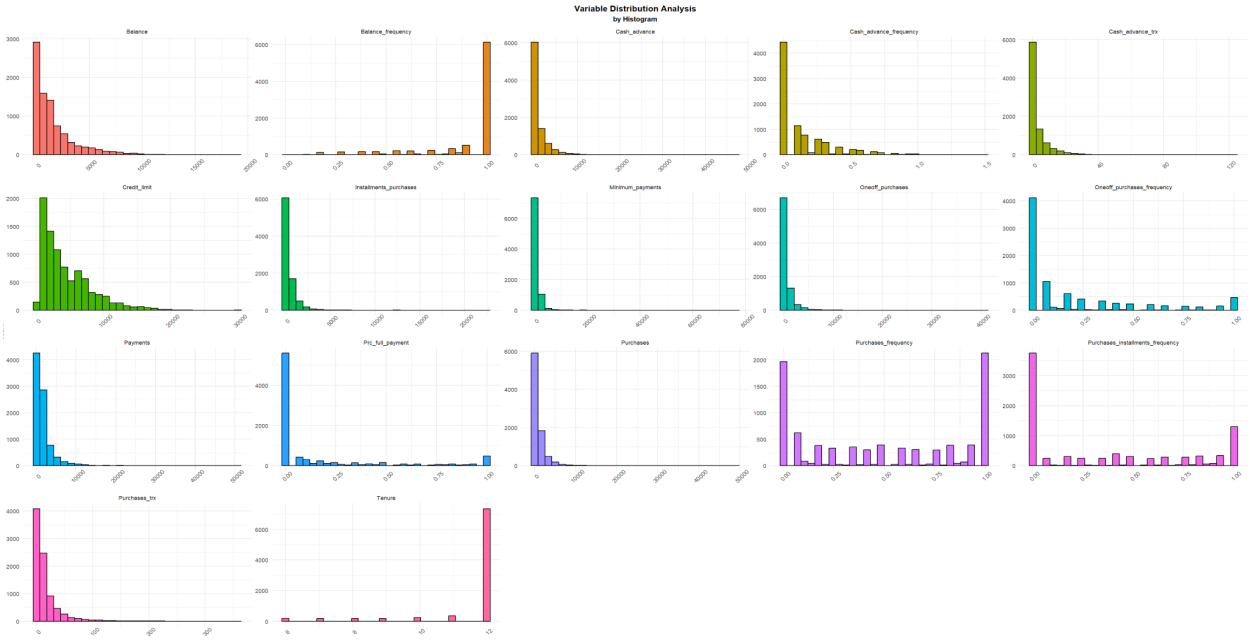

Fig 1: Explanatory Data Analysis Histogram

The variable distribution analysis shows that the data is not normally distributed, with most variables being skewed to the left or right. For instance, users tend to have a lower balance, but have a high frequency of balance changes indicating they use their cards frequently. Interestingly, customers with higher credit limits tend to have lower balances, and vice versa. Purchase frequency also has a unique distribution, with the majority of users falling into either high or low frequency categories.

Variables	Description
CUSTID	Identification of Credit Card holder (Categorical)
BALANCE	Balance amount left in their account to make purchases
BALANCEFREQUENCY	How frequently the Balance is updated, score between 0 and 1 (1 = frequently updated, 0 = not frequently updated)
PURCHASES	Amount of purchases made from account
ONEOFFPURCHASES	Maximum purchase amount done in one-go
INSTALLMENTSPURCHASES	Amount of purchase done in installment
CASHADVANCE	Cash in advance given by the user
PURCHASESFREQUENCY	How frequently the Purchases are being made, score between 0 and 1 (1 = frequently purchased, 0 = not frequently purchased)
ONEOFFPURCHASESFREQUENCY	How frequently Purchases are happening in one-go (1 = frequently purchased, 0 = not frequently purchased)
PURCHASESINSTALLMENTSFREQUENCY	How frequently purchases in installments are being done (1 = frequently done, 0 = not frequently done)
CASHADVANCEFREQUENCY	How frequently the cash in advance being paid
CASHADVANCETRX	Number of Transactions made with Cash in Advanced
PURCHASESTRX	Number of purchase transactions made
CREDITLIMIT	Limit of Credit Card for user

Table 1: Data dictionary

Correlation Analysis

The process of correlation check is essential in understanding the relationships between variables in the dataset. In this project, a correlation check was conducted on the 17 credit card attributes to identify significant relationships between them. The high dimensionality of the dataset required the removal of highly correlated features to improve model performance and reduce any bias in the clustering models. For instance, purchase_installments_frequency was found to be highly

correlated with installment_purchases, while purchase_trx was highly correlated with cash_advance, as seen in Figure 2. Based on the correlation check, variables with a correlation level of 80% or higher were removed from the dataset. By reducing the dimensionality of the dataset, we can build more accurate and robust clustering models to understand consumer spending patterns better.

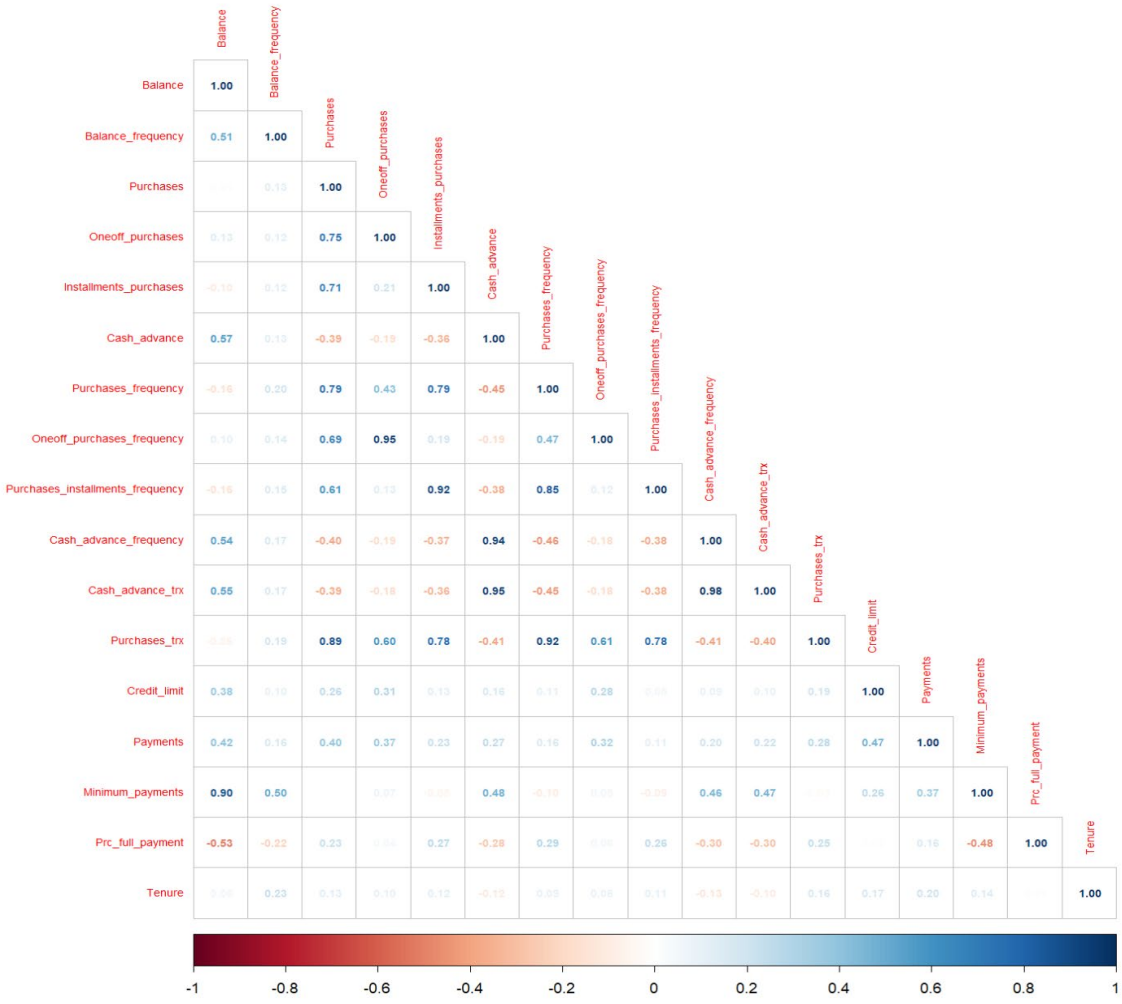

Fig 2: Correlation Analysis

The result of removing highly correlated values reduced the features by 7. So, we are left with just 10 features.

Principal Component Analysis

We utilized Principal Component Analysis (PCA) as a preprocessing technique to reduce the dimensionality of our dataset. As the original dataset had 17 credit card attributes, PCA helped us to identify the most important features that can explain the majority of the variance in the data. The importance of PCA in clustering lies in its ability to provide a lower-dimensional representation of the data while retaining the most important information. Before performing PCA, we first scaled the data to ensure that all features were on the same scale. This is important for clustering because it uses Euclidean distance to group data into cohorts. Scaling allows us to give equal importance to all the features, and it ensures that no feature introduces bias by providing incorrect significance due to a different scale.

After scaling, PCA was used to compute the principal components. These components were then ranked from high to low, providing the proportion of variance for each component and the cumulative proportion. The cumulative proportion helped us to understand how much data each component explains. Our goal was to choose the components that could explain at least 70% of the data. This value was chosen to avoid introducing bias and creating overfitting. In total, we chose 5 components for clustering. By using PCA, we were able to reduce the dimensionality of the dataset while still retaining the most important information. Furthermore, we were able to identify the most important features that could help us to cluster the data more effectively. The use of PCA as a preprocessing technique is a common practice in clustering and helps to improve the performance of clustering models.

```
Importance of components:
                          Comp.1    Comp.2    Comp.3     Comp.4     Comp.5    Comp.6     Comp.7
Standard deviation     1.5695096 1.2482362 1.0918866 0.98483795 0.96592161 0.9091019 0.82108481
Proportion of Variance 0.2463646 0.1558274 0.1192354 0.09700181 0.09331126 0.0826562 0.06742583
Cumulative Proportion  0.2463646 0.4021920 0.5214274 0.61842924 0.71174050 0.7943967 0.86182254
                           Comp.8     Comp.9   Comp.10
Standard deviation     0.78775039 0.71018072 0.5066629
Proportion of Variance 0.06206225 0.05044151 0.0256737
Cumulative Proportion  0.92388479 0.97432630 1.0000000
```

Table 3: Principal Component Analysis of Credit Card Clustering

The Scree plot below provides a visual representation of PCA. We see that the first 2 components explain about 40% of the data. This information is important for plotting the PCA factor map.

In Fig 3, the first 2 components are used to describe the 10 features in the factor map below (15.6%+24.6% = 40.2%). Each feature is represented by a vector with length (or weight) towards the positive or negative direction. In addition, the color gradient describes which features contribute significantly more than the others. Using Dim1, Purchases has the highest vector of 0.8, followed by Payments. Using Dim2, Cash_advance_trx has the highest positive vector and Prc_full_payment (Percent of full payment paid by user) has the highest negative vector. Observe also that Dim1 and Dim2 are orthogonal from each other.

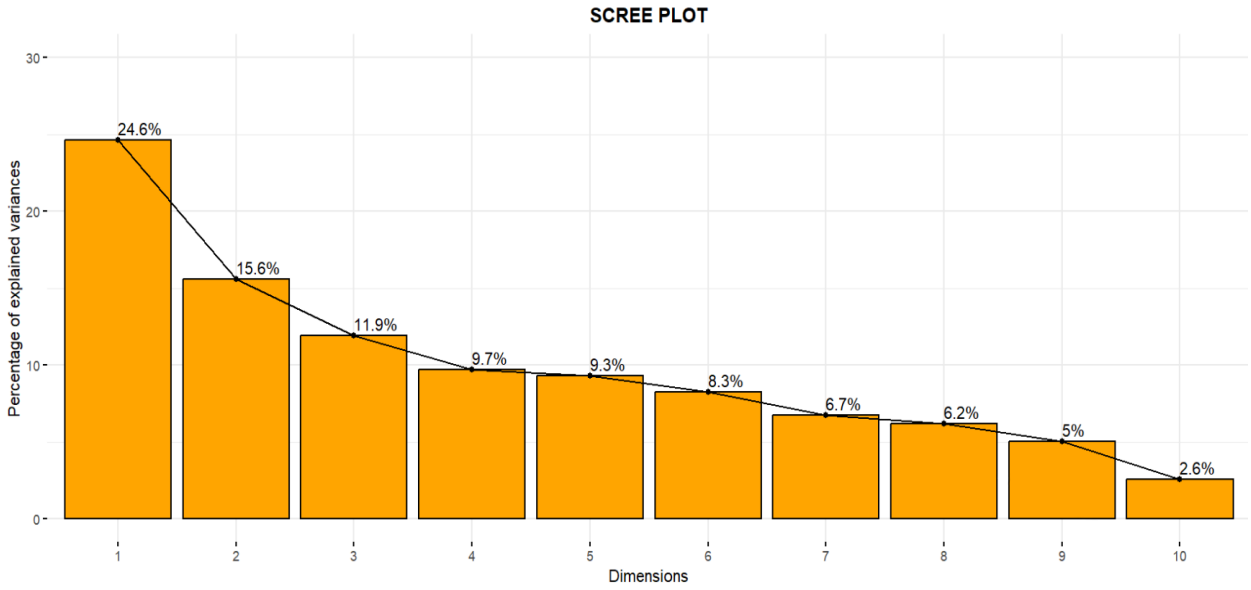

Fig 3: Scree Plot

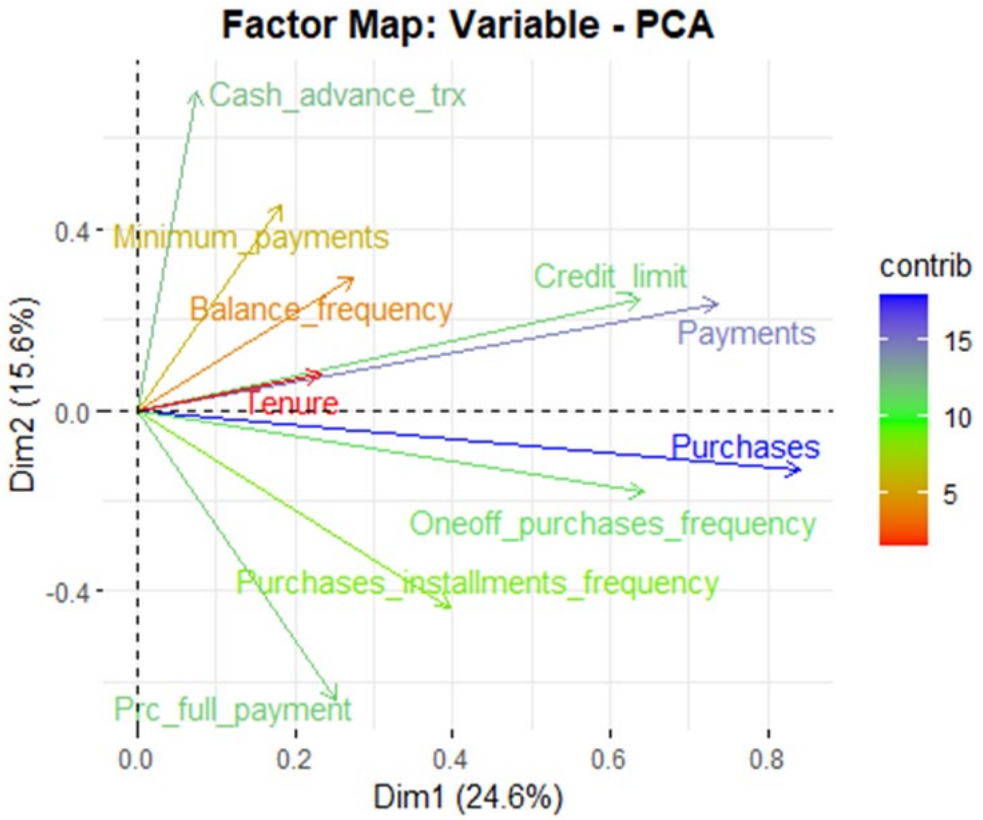

Fig 4: Factor map

Hypothesis Testing

We apply Hopkins Statistics Method to test the spatial randomness of the data. Our goal is to get uniform distribution of the data because we want to assess the clustering tendency of our dataset. This means getting to a value closet to one. Our results illustrate a Hopkins score of 0.999995. Aiming for a confidence interval of 0.95, we reject the null hypothesis hence credit card dataset is significantly clusterable data.

```
```{r}
set.seed(123)
hopkins(df_scale, m = nrow(df_scale)-1)
```
```

```
[1] 0.9999995
```

Fig 5: Hypothesis Testing using Hopkins Statistics Method

Clustering

K-Means Clustering

We started with the simplest clustering algorithm K-Means. We take the PCA output for 5 components. To determine the optimal number of clusters for KMeans we apply an elbow method. Using the fviz_nbclust library, we obtain fig 6 below.

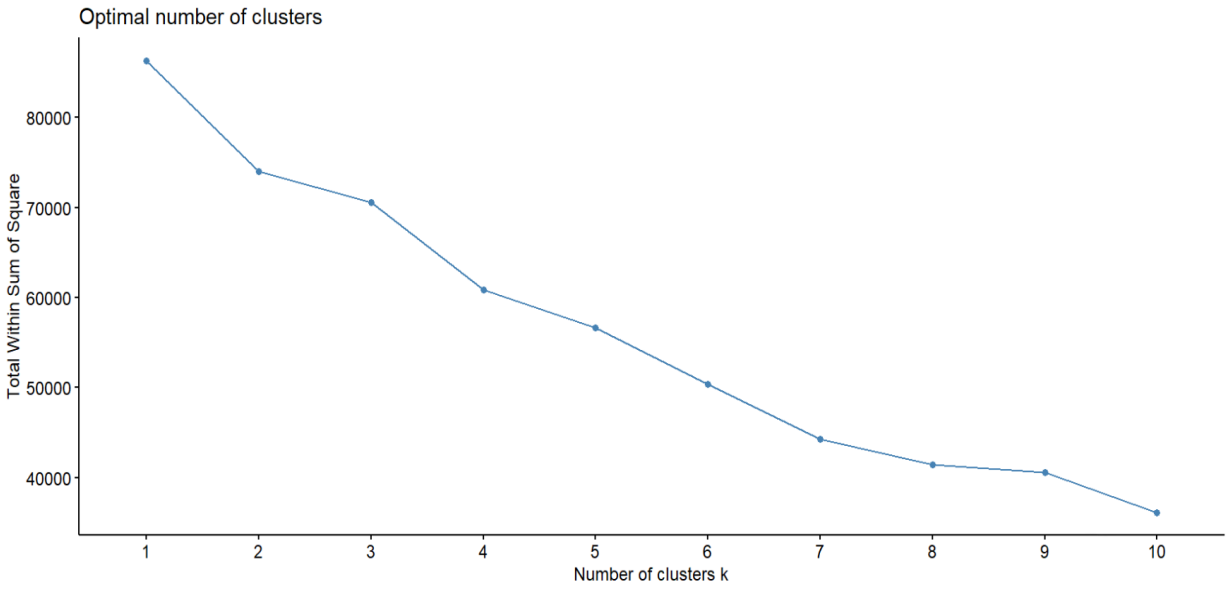

Fig 6: Elbow method for KMeans

Using fviz_nbclust package to calculate KMeans cluster, we obtain the following result

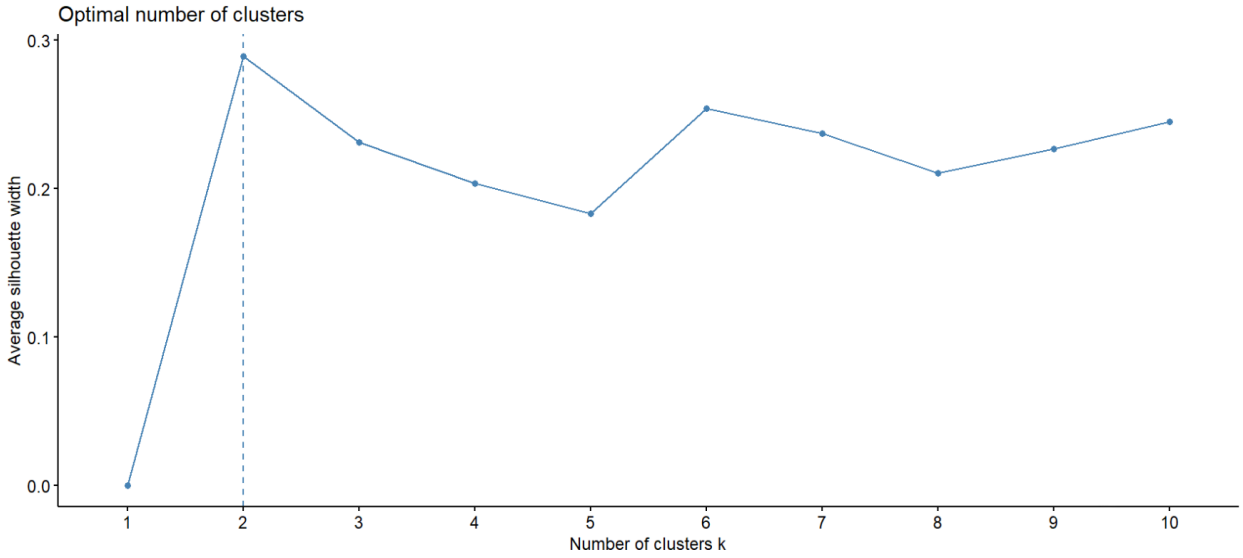

Fig 7: Silhouette method for KMeans

Using fviz_nbclust package to calculate KMeans cluster, we obtain the following result

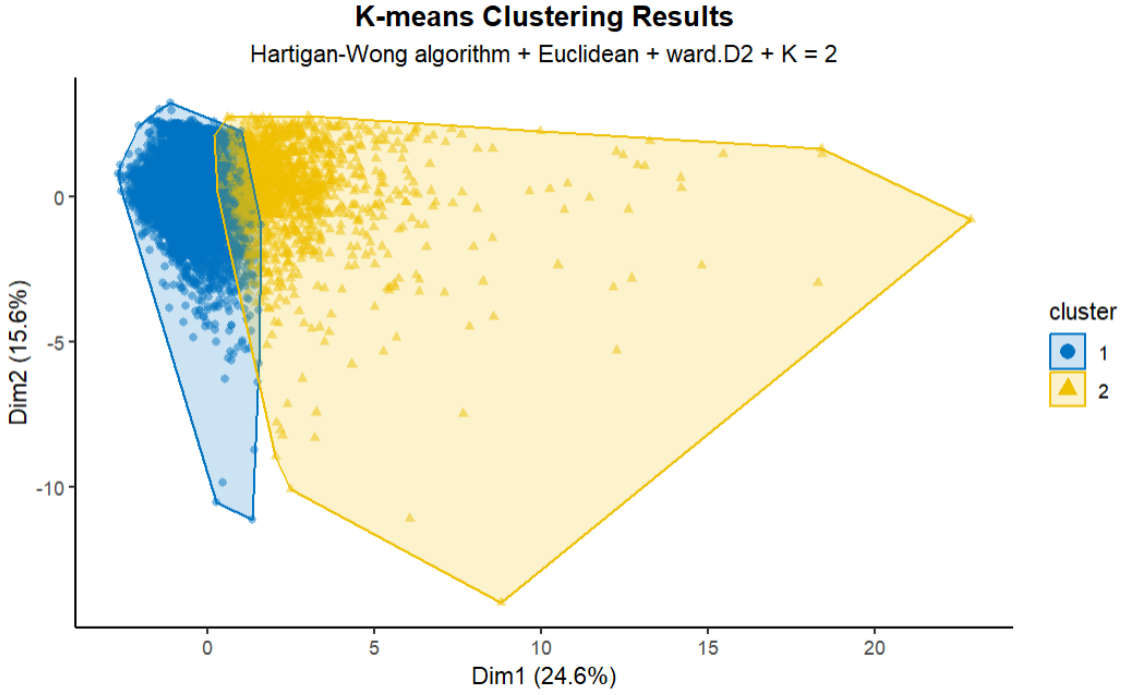

Fig 8: KMeans Cluster

The Radar plot provides more analysis of the two clusters. The radar plot or spider plot starts from zero, the center of the radar and increase in magnitude out. The first cluster (Cluster 1, red), represents less active credit card users as all the data points are close to the center of the plot. While the customers on the right represent very active users because the data points are spread out. The highest values are oneoff_purchases and purchases.

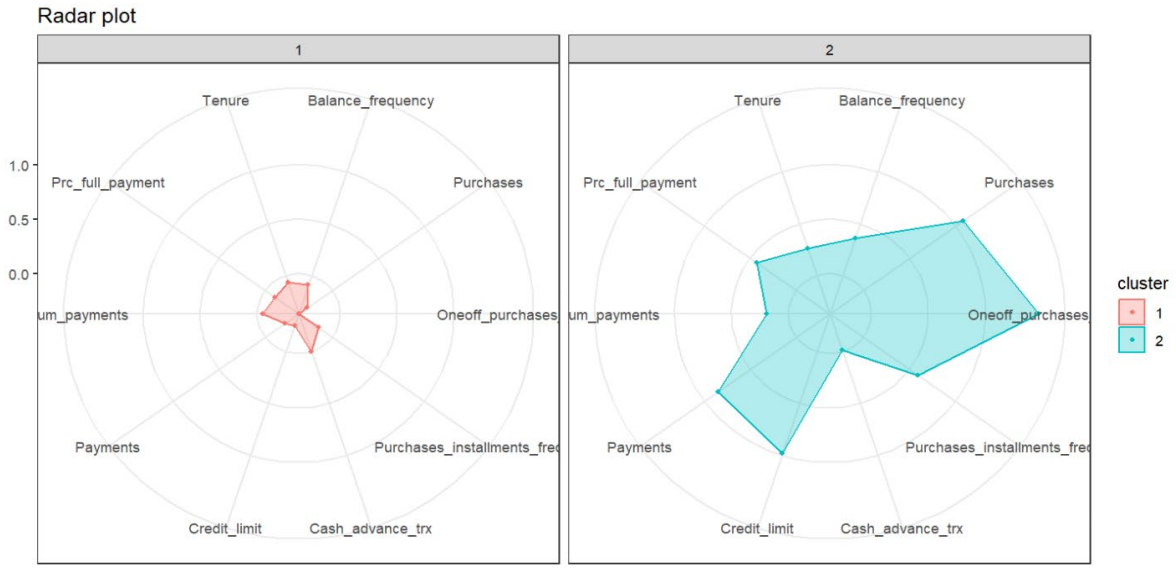

Fig 8: KMeans Radar Plot

Fuzzy Clustering

Fuzzy Clustering is a fascinating type of clustering where datapoints can belong to multiple clusters. It introduces a partial membership concept that deviates from traditional clustering methods such as KMeans. This type of clustering is ideal if there are ambiguities in the datapoints where they could belong to multiple clusters. We can explore this clustering method as part of our unsupervised learning to gain more insights into user behaviors. The results of Fuzzy Clustering, as shown in Fig 9, produce the very similar clustering as KMeans, providing two cluster groups.

However, Fuzzy Clustering offers more information about these groups that KMeans did not offer. The first cluster represents inactive cash in advance transaction users, indicating a strong preference for using cash in advance. This preference is more evident than the first cluster in KMeans in the previous section. The second cluster indicates that installment-purchases are more popular than one-off purchases. Although both fall into the same category of "Purchases," this cluster highlights the preference of users to make full repayment and less preference for minimum payment.

Fig 9: Fuzzy Clustering

Clustering Large Applications (CLARA)

We employ CLARA clustering because it is useful for large datasets. It is a much fastro algorithm and uses random sampling to create smaller subsets and does not use all the datapoints like K-Means.

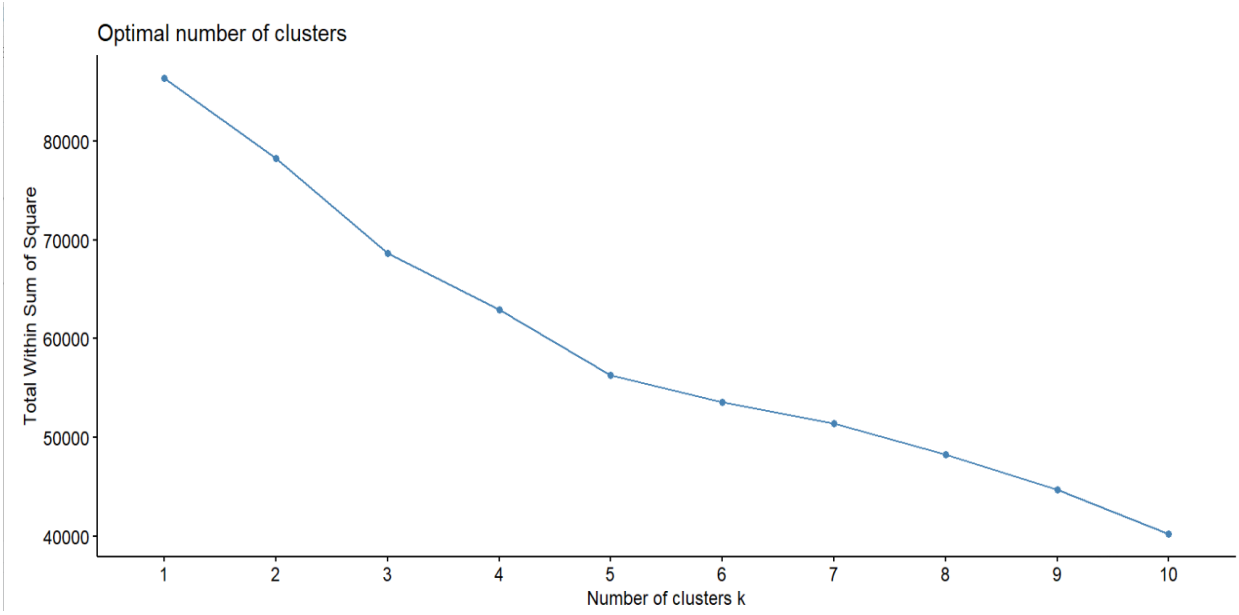

Fig 9: Elbow method for CLARA

We also want to uncover more clusters from active credit card users to understand more of their buying behaviors that we got from KMeans clustering as seen in Fig 8. We apply elbow method and silhouette method to determine the number of clusters to use on CLARA as shown in Figure 9 and 10. We obtain four clusters, much better than K-Means.

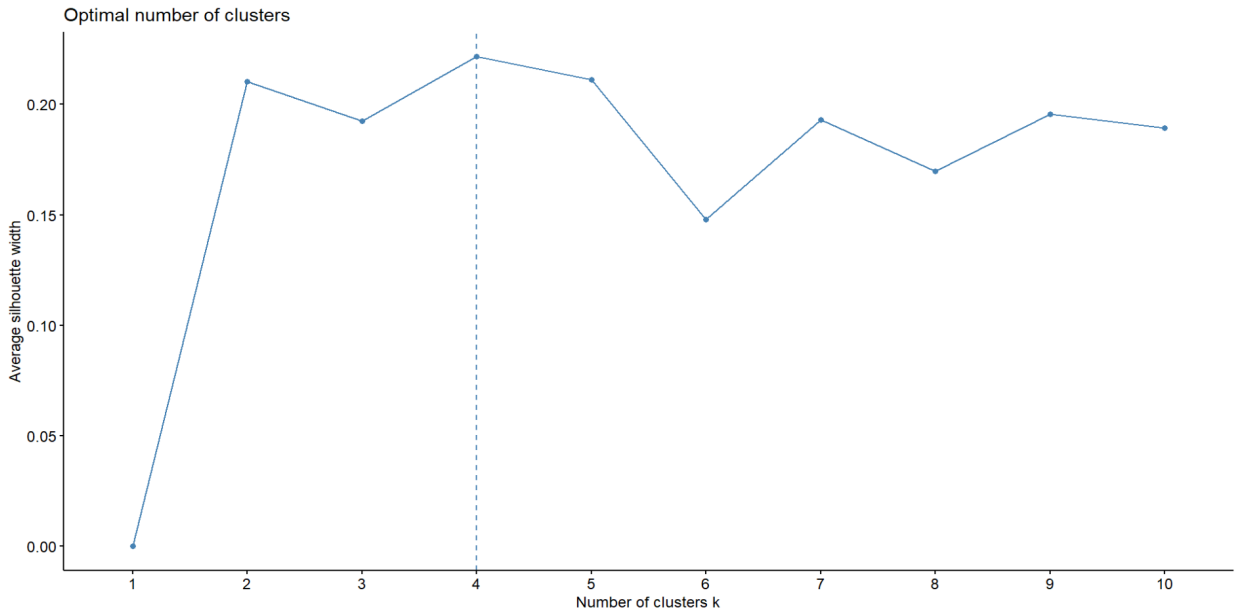

Fig 10: Silhouette method for CLARA

The Radar plot below illustrates the breakdown of the second cluster identified in KMEANS into four distinct clusters by CLARA. In total, CLARA identified four clusters, each with unique credit card usage patterns.

Cluster 1: The first cluster represents Inactive Users, who do not frequently use their credit cards and only make occasional transactions. They are less active compared to users in other clusters.

Cluster 2: The second cluster comprises One-Off Purchasers, who tend to use their cash advance options for transactions.

Cluster 3: The third cluster consists of Purchase Installment users, who prefer to pay for items over time rather than paying the entire balance at once.

Cluster 4: The fourth cluster is identified as Pay-Off Payers, who have a zero "Balance_frequency" indicating infrequent activities with the credit card company. Users in this group also have a high level of percentage of full-payment, indicating that they prefer to pay off their balance in full. This may explain their low level of "Balance frequency."

Cluster 5: The fifth and final cluster comprises Active Users, who exhibit a high level of credit card activities in terms of dollar amount. Users in this group spend larger amounts of money in

their purchases, make frequent cash advances, and make large repayments. However, they have a low level of full payment (percentage), indicating a preference for making partial payments instead of paying off their balance in full.

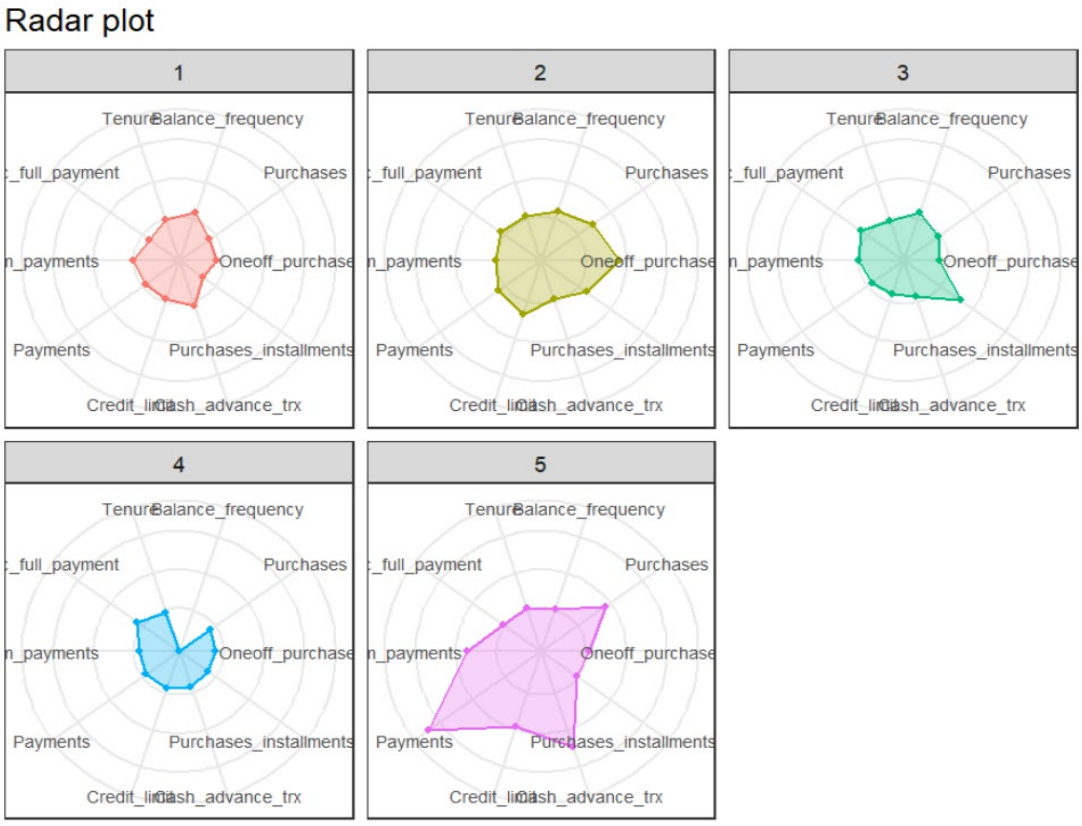

Fig 8: CLARA Radar Plot

Model Based Clustering

Model-based clustering is a powerful technique that uses probability models to assign data points to clusters. We used a sophisticated mode-based clustering R package that leverages the Gaussian Mixture Model (GMM) algorithm to identify groups of customers based on their spending behavior. The optimal number of clusters was determined using the Bayesian Information Criterion (BIC), a powerful metric that helps us to choose the best model. Our model-based clustering approach identified 8 distinct clusters, each with unique characteristics.

Cluster 1: This group represents fewer active users with a preference for cash-in-advance. Although they don't like to use credit cards much, they are more likely to use cash-in-advance for transactions.

Cluster 2: This group consists of less active users who prefer to use credit cards for purchases, especially installment purchases.

Cluster 3: The Revolvers cluster represents credit card users who prefer to make minimum payments back to the credit card company. They tend to make expensive purchases and have higher credit limits, but they don't like to make full payments back.

Cluster 4: This cluster represents fewer active users who prefer to make full payments and have very low tenure.

Cluster 5: Active card users make up this cluster, characterized by their preference for expensive products and large payments back to the credit card company. They sometimes prefer full payment, but often opt for minimum payments.

Cluster 6: This group consists of Max Payers who are less active and have zero tenure.

Cluster 7: This cluster represents less active Revolvers who prefer to make minimum payments and spend less on cheaper products.

Cluster 8: This group represents Max Payers who are active and have a high level of tenure.

Conclusion

Each clustering algorithm provides important information that may serve different needs of a credit card company. KMeans Clustering offered a simplified clustering outcome. Dividing users into active and inactive users. Fuzzy Clustering provided more insight to inactive and active users. Inactive users use their card for cash advance transactions. Active users like to purchase with a strong preference for leveraging the purchase installment options so that they can make their payments over time. CLARA broke out the second active cluster from KMeans into rather interesting results and Model-Based Clustering had the highest number of clusters 8, which is the largest number of distinct clusters compared to other clustering algorithms.

Limitations

There are several limitations to unsupervised learning that need to be considered. Firstly, the absence of demographic data such as age, gender, educational level, marital status, or profession may have limited the meaningfulness of the clusters obtained. Including this data in the clustering algorithm would have likely yielded more informative clusters. Secondly, unsupervised learning lacks performance evaluation metrics to assess the accuracy of each model. Thirdly, different algorithms produced varying cluster counts, making it difficult to determine the optimal number of clusters. Unlike supervised learning, unsupervised learning does not offer the ability to use sampling methods such as bootstrap or cross-validation to prevent overfitting. Noise in the data may have affected the outcome and the clusters may have failed to remove it. Lastly, interpreting

the clustering results may require additional research or subject matter expertise to provide more accurate insights for credit card companies seeking to take action based on the clusters.

References

Al-Khaja, M., and Al-Hmoud, A. (2003). Segmenting credit card customers: A comparative study of four methods. Journal of Business Research, 56(12), 1203-1212.

Arjun Bhasin. (2013). Credit Card Dataset for Clustering. Kaggle. Retrieved from https://www.kaggle.com/datasets/arjunbhasin2013/ccdata

Bhasin, A. (2013). Credit Card Dataset for Clustering. Kaggle. Retrieved from https://www.kaggle.com/arjunbhasin2013/ccdata

Oliver, R. L., Rust, R. T., & Varki, S. (1999). The future of credit cardholder segmentation. Journal of Direct Marketing, 13(3), 17-28.

Wang, Y., and Chen, L. (2013). Credit card customer segmentation using a hybrid of clustering and association rule mining. Expert Systems with Applications, 40(10), 4200-4208.