BA 820 Final Project Paper

Team 9: Antonio Moral, Bosoo Kim, (Sylar)Jiajian Guo, Yixuan Wang

1. Business Problem

People use credit cards everyday and all the usage information is collected by credit card companies. Americans currently have 511.4 million credit cards and 61% of American consumers have at least one credit card, while the average person has four. Credit Card Issuing in the US Market reached 98.9 billion dollars in 2021.

The purpose of this analysis is to tackle the business problem of how credit card users are segmented. With this project's analysis, advertisement companies could use these credit card users segments to improve marketing strategies such as campaign distribution optimization. Banks can also use this analysis to prevent credit card fraud, payment default and credit limit management problems with further analysis in the future.

2. Dataset

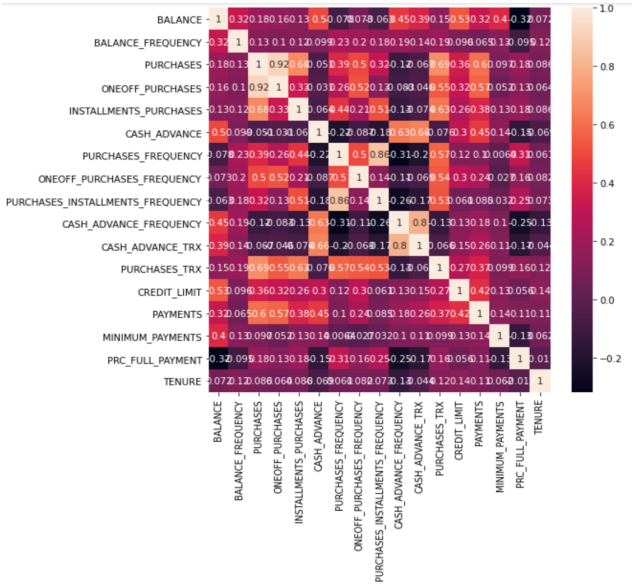
The dataset was obtained from Kaggle: https://www.kaggle.com/arjunbhasin2013/ccdata
Relevant attributes (columns) to focus on in our analysis are:

- BALANCE : Balance amount left in their account to make purchases
- 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
- ONEOFFPURCHASESFREQUENCY : How frequently Purchases are happening in one-go (1 = frequently purchased, 0 = not frequently purchased)
- CASHADVANCEFREQUENCY : How frequently the cash in advance being paid
- CREDITLIMIT : Limit of Credit Card for user
- PAYMENTS : Amount of Payment done by user
- MINIMUM_PAYMENTS : Minimum amount of payments made by user
- TENURE : Tenure of credit card service for user

We chose the Credit Card Dataset from Kaggle. This data set summarizes the usage behavior of active credit card holders for six months. It contains 8950 credit card holders and 18 different variables. There are 313 missing values in MINIMUM_PAYMENTS column and 1 missing value in CREDIT_LIMIT column. If payments equal to zero, we set minimum payments as zero, and set the remaining missing values based on the median proportion of minimum payments to payments. We dropped the only one missing value in the CREDIT_LIMIT column. Cust_ID

seems to be an unique id for each customer and hence won't play any role in our further analysis, so we dropped this column. After cleaning, the data set has 8949 rows and 17 numeric variables.

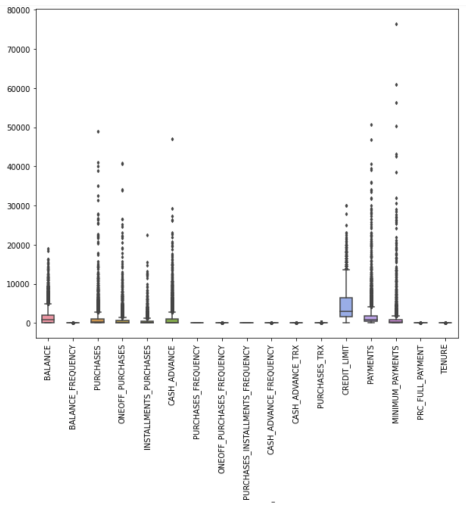
3. Exploratory Data Analysis



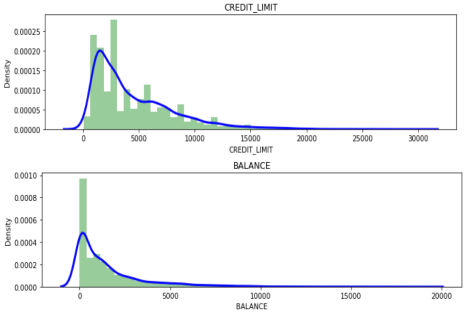
At first, we ran a heatmap style correlation matrix to identify highly correlated variables. We believe that this will allow us to streamline the feature selection process and help us to understand the dataset.

As we predicted, the "Purchases" feature has a higher level of correlation with "Oneoff_purchases" (0.92), and the "Purchases_Frequency" variable also showed a high level of correlation with "Purchase Installments Frequency" (0.86). On the other hands, the relationship between "Cash Advance TRX" and "Oneoff_Purchases" showed (-0.046), which means there is almost no relationship between these two variables.

Next, we believed that outliers may indicate bad data. Therefore, we used boxplot, one of graphical tools, in checking the normality assumption and in identifying potential outliers. 9 out of 17 variables have a value between 0 and 1. Therefore, there are no outliers for these variables. However, variables, such as "Payments" and "Minimum Payments" showed many outlier values. For example, "Minimum Payments" variable has a mean value of $864, whereas it's maximum value is $76,406. Since we are going to try different models with this data, and we believe these outliers would not hurt our hypothesis, we decided not to drop outlier values.



Last of all, we used the Kernel Density Estimation (KDE) plot to estimate the probability density function of a continuous random variable. Here are some examples. "CREDIT_LIMIT" and "BALANCE". For "CREDIT_LIMIT", we can see that most of the data are distributed between $0 and $5,000. For the "BALANCE" variable, we can also say that most of the data is



distributed between $0 and $3,000. We ran a KDE plot for all 17 variables, and we could easily identify the distribution of each variable.

4. Analytical Findings

We tried hierarchical clustering and DBSCAN, but the results did not look as good as the plot of the KMeans method. Therefore, we will only talk about the best-performing model as follows.

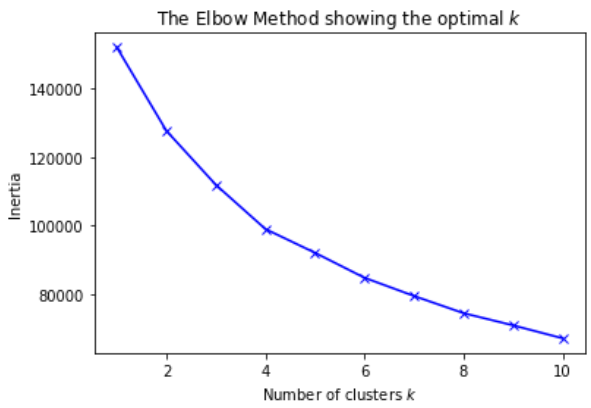
4.1 Looking for the Optimal Number of Clusters

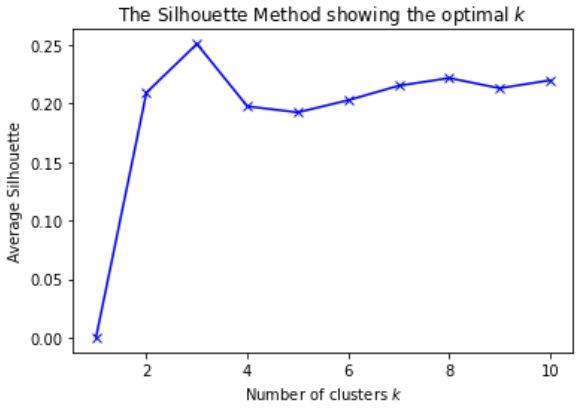
- **Elbow Method**

We calculated the Within-Cluster-Sum of Squared Errors (WCSS) for different values of k, and plotted the curve here. There is not a specific k for which WCSS apparently starts to diminish. In the plot of WCSS-versus-k, this is no visible elbow.

- **Silhouette Score**

Average silhouette method computes the average silhouette of observations for different values of k. The optimal number of clusters k is the one that maximizes the average silhouette over a range of possible values for k. So, we chose k=3 as the optimal k here with the highest score in the plot.

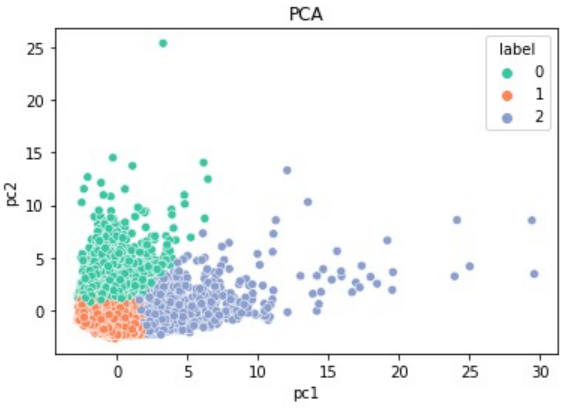


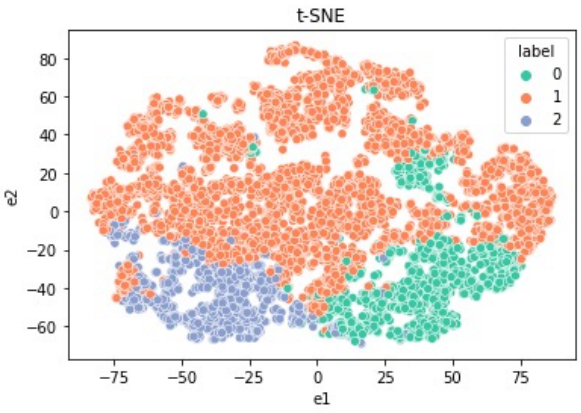


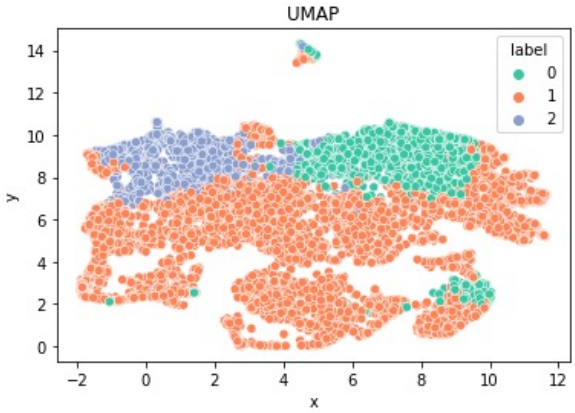
4.2 Clustering Methods

- **Clustering on original data**

Firstly, we applied the K-Means algorithm to our original dataset after standardizing features, and set the optimal number of clusters to be three, the reason explained in the previous part. Then the credit users were segmented into three like-groups. After that, we visualized our three clusters using three different dimension reduction methods: Principal Component Analysis (PCA), t-Distributed Stochastic Neighbor Embedding (t-SNE) and Uniform Manifold Approximation and Projection (UMAP) to reduce a dataset into 2 features (ie. 2D) and plot. The figures show below.



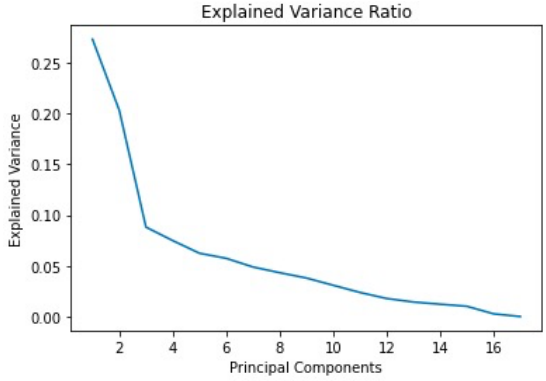


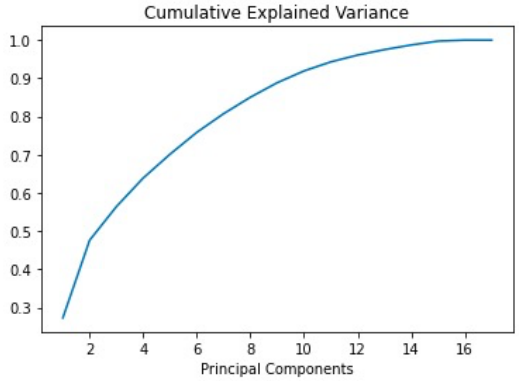


The PCA plot displays the three clusters on the two principal components created for 2-D visualization, we can see three distinct sections of the data points with strict and visible borders separating each color into groups. The t-SNE and UMAP plots also display the three clusters on the two dimensions for 2-D visualization, however, the data points between each cluster seem to overlap more than that in PCA plot.

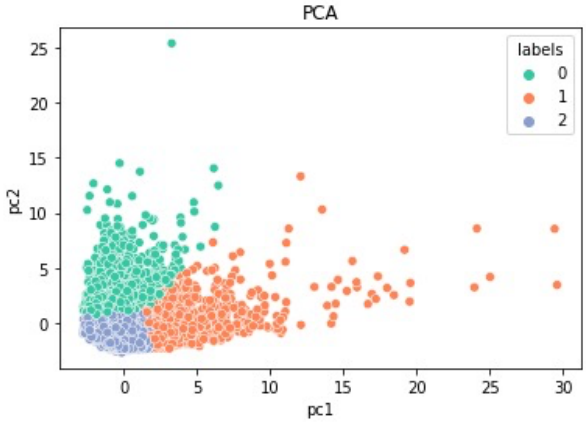
- **Clustering on reduced data**

Instead of working with the original dataset, we used PCA to create new variables. By examining the amount of variance each principal component encompasses we can see that the first 10 principal components explain about 90% of the variance.

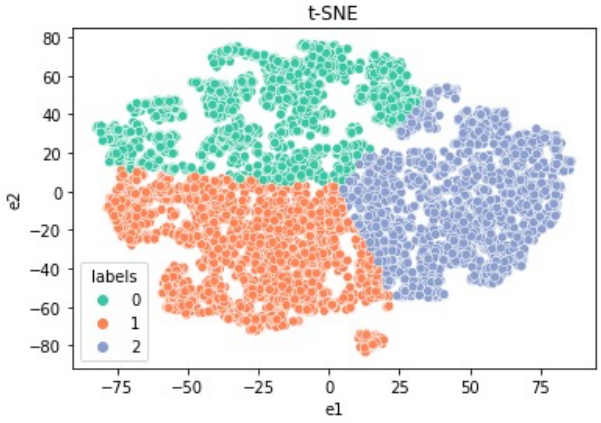


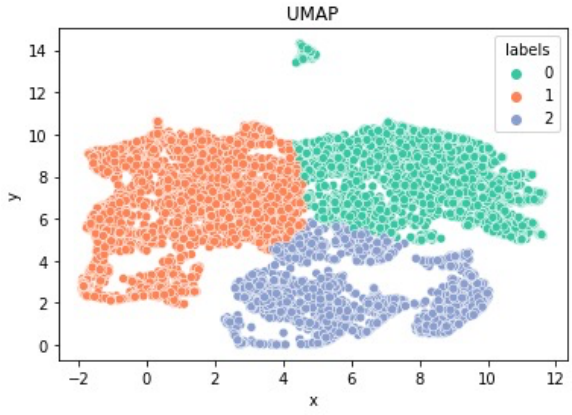


Then, we reduced our original dataset from 17 features to 10 principal components. This time, the optimal number of clusters is still equal to three from new Elbow and Silhouette plots. We applied the K-Means algorithm again on new variables and visualized the 2D plot using PCA, t-SNE and UMAP. There are still clear division lines between clusters in the PCA plot.



Finally, we reduced our dataset again using the t-SNE method. After that, we applied the K-Means algorithm on reduced data. This time, the clusters become much cleaner. We also used the UMAP method to repeat this process. We can see both figures have more significant improvements in the distinction between clusters than the figures of clustering on the original dataset.





5. Conclusions and Recommendations

Our analysis showed that 3 differentiated clusters can be found within credit card users. This number of clusters was estimated through different methodologies, such as the analysis of the elbow plot and the silhouette score. This analytical choice to group users into 3 clusters has been confirmed after our team observed the tendencies of each of the resulting clusters. We were able to find that these clusters have specific characteristics that differentiate them from each other, and after analyzing these differences, we are able to recommend specific strategies to maximize customer retention and increase profitability for credit card companies.

Starting with Cluster 0, it was evident that users in this cluster were those with the highest average balance and credit limit. This served as an indication that these customers have the highest purchasing power within our sample, so we focus on this aspect to formulate our strategy. Meanwhile, users in Cluster 2 were those with the most purchases in the time period, and thus we could establish that the most active customers would fall into this cluster. As became evident with our analysis, Cluster 0 and Cluster 2 were somewhat similar in certain aspects, however their characteristics were unique enough to warrant separate strategies. Finally, users in Cluster 1 were those with the lowest average balance, credit limit and purchases, leading us to believe that these clients were smaller and used mostly cash advances to make small purchases. After gaining insight into what characterized each cluster, we were able to elaborate 3 different marketing and sales strategies to be applied to users in their respective cluster.

The 3 proposed strategies were devised with a few key goals in mind: increase customer retention and increase revenues per user. The proposed strategies are as follows:

- For customers in Cluster 0: As these customers tend to have less purchases but of higher value, and just have higher overall purchasing power, we propose credit card companies increase the interest rate slightly. Given that these customers do not make their payments in full on average, this would increase the revenue coming into the company without affecting the users' behavior. In addition, the company could lower the interest on cash advances for this group of customers, to incentivize them to use this feature more, which could in turn lead them to requesting higher balances and credit limits, overall increasing revenues.
- For customers in Cluster 1: As these customers have the least purchasing power and lowest activity, we believe offering financial advice and capital asset management knowledge would be the most beneficial plan of action for the company. Through this, customers in this cluster will gain better understanding of financial management and, more importantly, will feel supported, which would turn them into users for life. This, paired with slightly lower interest rates on cash advances will increase customer retention, with the possibility of clients increasing their credit card expenditure.
- For customers in Cluster 2: As these customers are the most active, meaning they make the most purchases, and they make the most payments in full, we recommend the company decrease interest rates on credit card purchases. This would incentivize users to spend more per purchase as the risk of not making a full payment decreases. This would then increase the yield for the company.

Although we believe these strategies will prove beneficial for the company, it is important to note that a staggered approach should be taken when implementing them to be able to monitor their effectiveness. These strategies should also be implemented to client bases from banks similar to the one the data was collected from, as we cannot assume this distribution of clients will hold in more niche banks and markets. Finally, we believe that this project has been

successful in explaining the composition of the client base and allowing us to develop tailored strategies to customers in each cluster found.