**Using KMeans, Agglomerative and Spectral Clustering to segment Credit Card Customers**

**Abstract**

Customer segmentation is an important topic for many industries and the credit card industry is no exception. All customers are never the same. Their needs can be complex to unravel and more importantly, those needs evolve. Businesses are faced with the challenge of identifying the ever-evolving complex web of needs of customers in order to tailor services to them. This has become very important for businesses to stay competitive. The existence of state-of-the-art machine learning algorithms among other powerful modern technologies have made it possible to understand customers on a large scale and more importantly develop and tailor marketing campaigns, products and services to the right market segment. In this paper, the author uses KMeans, Agglomerative and Spectral clustering to segment credit card customers into four segments. Within each profile segment, the author notes their characteristics and further recommend marketing strategies for each of the profile segment.

**Introduction and Literature Review**

People use credit card daily, weekly, monthly and even yearly to cover expenses of all nature. In the US alone, the value of all credit cards amounts to 3.06 trillion dollars. Again, the number of credit cards in the US is a little over 1.1 billion (Best, 2020). The high demand and usage of the credit cards gives an insight into the many advantages consumers enjoy from their credit cards. Credit cards enable consumers make huge purchases they otherwise could not have afforded. Additionally, it enables consumers to make emergency purchases in the event they do not have enough balance in their accounts. The sheer convenience of carrying credit cards as compared to cash and its concomitant rewards as a result of usage explains its wide-spread adoption and usage (Umuhoza et al., 2020).

The initial strategy for many credit card companies was how to increase card quantity in the system as much as possible. Thus, they cared a lot about how to get their cards into the hands of consumers without much thought about how to get majority or all customers to use them regularly. Over time, the competition for high quality customers has become the main goal (Ying & Yuanyuan, 2010). Even more insightful is the fact that increasing customer retention by 5% will increase profits by 25% to 95% (Gallo, 2014). The bottom line therefore is that retaining and upgrading customers through customer segmentation is valuable and will serve companies in the long run and make them more competitive and profitable.

Customer segmentation refers is the process of dividing customers into groups based on common attributes such as demographics, behavior among others. Segmentation helps analysts to know which group of customers may respond in a certain way to a particular message or marketing campaign (Umuhoza et al., 2020). The advantages of customer segmentation are numerous. Firstly, it allows companies to tailor marketing campaigns and optimize marketing costs to a particular group of customers who are likely to respond favorably to the campaign. Again, it helps companies to build products, services and experiences for targeted customers. Great customer and product experiences build a strong connection between companies and their customers. Additionally, it reduces customer churn rates, strengthens customer loyalty and customer relationship management.

The existence of Machine learning and artificial intelligence technologies has made it possible to build effective customer segments at scale. Daily, weekly, monthly or yearly usage of credit cards has made it possible to collect card usage data on customer. This has unlocked the possibility of extracting insights on the behavior of customers using machine learning tools. With the aid of cluster analysis, analysts are able to divide customer into groups based on homogenous features or attributes using some of the clustering algorithms available in Machine learning. There are large body of works that employ various machine learning techniques to segment users into groups to further understand their customers and provide targeted services to the different groups. A subset of those works can be found in the following paragraph.

Umuhoza et al. used anonymized credit card data of a leading Bank in Egypt, Commercial International Bank of Egypt to build a behavioral-based segmentation model that differentiates African credit card holders based on their purchase data using KMeans clustering. They sought to transform credit card business model and campaign strategy of the Bank from reliance on traditional value-based campaigns to more targeted campaigns that are informed by customers’ lifestyle, needs, and usage preferences (Umuhoza et al., 2020). Abidar et al. proposed and built a model based on the RFM model (Recency, Frequency, and Monetary value) and KMeans algorithm to segment customers using a dataset containing online transactions of a UK based retail company. This way, they were able to use clustering, scoring, and distribution to have a firm grasp of what actions were relevant to improve customer satisfaction. They believe that every firm should have a strong tool to be used to understand their customers and to provide the care that is necessary to keep them on their active customers lists (Abidar et al., 2020). Based on a Chinese commercial Bank’s credit data, Li et al. segmented credit card customers into four clusters using KMeans clustering algorithm. Following that, they built forecasting models using data mining methods such as the C5.0, neural network, chi-squared automatic interaction detector and classification and regression tree using the background information of the credit card customers. After model evaluation and the verification of the effectiveness of the regulations, they use the regulations to describe the features of the customer segment and help the bank find target customers (Li et al., 2010). Using the data of credit card customers from a commercial bank of Shanghai, Ying and Yuanyuan developed marketing strategies for three market segments as identified by their algorithms which had theoretical value and practical significance. They used two algorithms. The first being the Analytical Hierarchy Process (AHP) for indicator optimization and the KMeans algorithm for clustering.

This paper uses unsupervised machine learning algorithms (KMeans, agglomerative and spectral clustering) to cluster customers into four groups. The paper further notes the characteristics of each group and offers recommendations on potential marketing strategies to adopt to maximize customer retention, loyalty and ultimately profits.

**Problem and Data Sets Description**

In this paper, the author seeks to develop a customer segmentation to define marketing strategy. In so doing, the author will build an enriched customer profile by developing smart Key Performance Indicators (KPIs) such as monthly average purchases and cash advance amount, purchases by type, credit score, and payment to minimum payment ratio. The author will further use the derived KPIs to gain insights on customer profiles. Additionally, the author will perform variable reduction technique and apply clustering algorithm to the dataset to reveal behavioral segment of the credit card holders. Finally, the author will recommend some marketing strategies for each of the customer segment.

Customer segmentation, more so credit card segmentation is very important in business as it provides companies with the ability to market programs that are suitable for specific market segment. Again, because businesses are able to tailor services, products and experiences to relevant market segments through customer segmentation, they are able to build strong connection with their customers. This tends to reduce customer churn rate, strengthens loyalty and improves customer relationship management. Finally, market segmentation can raise new research questions as well as provide directions to its solution.

The dataset from AnalytixLabs contains the credit card usage behavior of about 9000 card holders during the last six months. It contains 18 variables. The variables are made up of the customer identification number, information on purchases, payments, cash advance, balances, credit limit and tenure.

**Methods**

To solve the problem, the author explored three unsupervised machine learning methods namely, KMeans, Agglomerative Clustering and Spectral Clustering.

**KMeans**

KMeans clustering is an iterative algorithm that aims at partitioning a dataset into a predefined number of clusters such that each row in the dataset belongs to only one group. KMeans tries to make the intra cluster data points as similar as possible and make the inter cluster data points as dissimilar as possible. It assigns a data point to a cluster such that the sum of squared distance between the data point and its centroid is at the minimum. The less variation within a cluster, the more homogeneous the data points are within the cluster (Qiu et al., 2014).

**Spectral Clustering**

Spectral clustering is another popular clustering algorithm which perform very well in various scenarios. It assumes each point is a graph node and thus treats the clustering problem as a graph-partitioning problem. The time and space complexities are very expensive thus restricting its application on large scale datasets (Huang et al., 2020).

**Agglomerative Clustering**

Agglomerative Clustering is a type of hierarchical clustering algorithm that groups similar objects into clusters. For Agglomerative clustering, it utilizes the bottom-up approach where it starts with many small clusters and merge them together to form bigger clusters (Nugraha et al., 2018). Agglomerative clustering, however, has a number of linkage criteria that explores how distance between clusters are calculated. For our dataset the ‘Ward’ linkage criteria, which calculates the distances between clusters as the sum of the squared differences within all clusters.

**Experimental Setup**

**Missing Value Treatment**

Before applying all the clustering algorithms mentioned above, we explored the dataset to observe any missing values. Two of the columns had missing values. Because the ratio of missing values was very small, hence, we replaced the missing values with the mean value of the column.

**Developing KPIs**

We generated KPIs from our dataset to be used for profiling cluster segment later on. We observed four purchasing behaviors of customers in our dataset. They included those who make both one off and installment purchases, those make none (neither one off or installment purchases), those who make only one-off purchases and finally those who make only installment purchases. These four groups of behavior helped generate new insights about our customers. Before generating visualizations on the KPIs, we performed log transformation on our dataset sets since the distribution of the all the columns were skewed. The log transformation removed the skewness from the dataset and helped to generate useful visualizations.

The KPIs included the following:

***Payment to minimum payment ratio*** - Every month customers are required to make minimum payment on their credit card balances. In the cells below, we will calculate the ratio of payments to minimum payments. In other words, we want to know now many customers pay per months as compared to their required minimum payment.

Chart, bar chart

Description automatically generated

Figure 1 - Average Payment to Minimum Payment Ratio

From the above graph, we can observe that customers who make installment purchases have the highest payment to minimum payment ratio. This means that they are far less likely to default on their credit cards. This is because they are almost always able to keep a good credit score. Customers who make one-off purchases have the lowest payment to minimum payment ratio. This means that they are more likely to default on their credit cards than other purchase type behaviors.

***Balance to credit limit ratio -*** Balance to Credit limit ratio which is also called credit utilization ratio measures the balances you owe on your credit cards relative to the credit card limit. A higher ratio means that you may get loans at a higher interest rates or may not be able to get loans at all. A ratio of 0.3 or below is ideal in many cases. In other words, a good credit score means that owing 30% or less of the credit limit. To calculate the balance to credit limit ratio, we will divide the credit balance of the customer by his or her credit limit.

Chart, bar chart

Description automatically generated

Figure 2 - Average balance to limit ratio

Balance to limit ratio of 0.3 or below denotes a good credit score. From the graph, we can see that customers who make installment purchases have the best credit scores. This is because they have a high payment to minimum payment ratio and thus are the least likely to default on their credit cards. Those who do none of oneoff or installment purchases have the worst credit scores. They usually take a lot of cash advance amount and thus are the most likely to default on their credit cards.

***Monthly average cash advance amount –*** is calculated by dividing cash advance amount by how long the customer has been with the company**.**

Chart, bar chart

Description automatically generated

Figure 3 - Monthly Average Cash Advance

From the above graph, we can see that customers who make none of oneoff or installment purchases have the highest average monthly cash advance amount. These group of customers as seen in the previous graphs, are the most likely to default on their credit card.

***Monthly average purchase*** - The monthly average purchase is calculated by dividing total purchases by how long the customer has been with the company**.**

Chart, bar chart

Description automatically generated

Figure 4 - Average Monthly Purchase

From the above graph, we can see that the group of customers who make both one-off and installment purchases are those that have the highest monthly average purchases. This makes sense because they make two types of purchases.

**Principal Component Analysis**

Principal Component Analysis (PCA) is a dimensionality reduction technique used to reduce the number of features in a dataset. Some features may be redundant leading to a situation where we are using more data than we actually need in order to reach the same goal. More features also cause high computational overhead. For this reason, we will use PCA to reduce to the number of features and also to reduce the computational overhead. Before we applied PCA, we standardized our dataet using the Standard Scaler from skicit-learn library. After applying the PCA, the study found out that 6 components explained a little over 90% of the overall variation in our dataset. Therefore, only six components will be used for our clustering.

**Modeling**

To ensure optimal model performance, we had to ensure the optimal number of clusters is selected for both the KMeans algorithm and the Spectral clustering algorithm. To do this we used the the researcher must set number of k, which can be achieved using methods like the elbow, Silhouette and the Calinski Herabaz score.

**Methods for choosing optimum k clusters for KMeans**

**Chart, line chart

Description automatically generated**

Figure 5 - The optimum number of clusters in Calinski Harabaz method is given by the peak bend in the curve before a sharp decrease. On this graph, the optimum number is at k=4.

**Chart, line chart

Description automatically generated**

Figure 6 - The silhouette coefficient increases up to the point where k=4 and then decreases. It exhibits a peak at k=4, which indicates the number for optimum clusters.

**Methods for choosing optimum k clusters for Spectral Clustering**

Chart, line chart

Description automatically generatedChart, line chart

Description automatically generated

Figure 7 - The optimum number of clusters in Calinski Harabaz method is given by the peak bend in the curve before a sharp decrease. On this graph, the optimum number is at k=4.

Figure 8 - The silhouette coefficient increases up to the point where k=4 and then decreases. It exhibits a peak at k=4, which indicates the number for optimum clusters.

The elbow method as shown below showed the same number of k clusters.

Chart, line chart

Description automatically generated

Figure 9 - The elbow method shows that k=4 is the optimal number of clusters.

**Results and Discussions**

The purpose of this study was to generate homogenous customer segments that represents behaviors of credit card holders. With this analysis, companies will be armed with strong customer knowledge that will help them develop more targeted campaigns, products and services to each segment. Thus, ultimately crafting better value proposition to their customers. After running the KMeans, Spectral and the Agglomerative clustering algorithm on our dataset, our analysis revealed four distinct segments. The profiles, and marketing strategy for each segment are reported in the next paragraph.

**Segment 1:**

***Profile:*** This segment has the highest Monthly average purchases and the higghest of both of oneoff and installment purchases. They make only both type of purchases (installment and oneoff). This segment also has the highest payment over minimum payment. This means that this segment pays more than the minimum payment required for a credit card than any other customer segment. They form 31% of the customer base.

***Marketing Strategy:*** This segment is making the purchase and making the payment as well. They also maintain a comparatively ok credit score. Therefore, they are potential customers. Their credit limit may be increased, or interest rates lowered. They can also be given incentives like loyalty or premium cards to encourage more transactions.

**Segment 2**

***Profile:*** This segment has the highest average cash advance amount. It also has the worst credit score among the rest of the segments. It makes the lowest monthly average purchases. They form 21% of the customer base.

***Marketing Strategy:*** This segment of customers is likely to default on their payment because of bad credit scores and lowest payment over minimum payment ratio. The company can target them by reducing interest rates on purchase transaction. This will increase the percentage of this segment and ultimately get them into segment 4.

**Segment 3**

***Profile:*** This segment has the second highest monthly average purchases. They also have second highest payment over minimum payment. They also have the second worst credit score. They make only oneoff purchases. They form 22% of the customer base

***Marketing Strategy:*** This segment of customers makes only oneoff purchases (eg. payment of utility bills). As their credit score is the second worst and has the second highest monthly cash advance amount, they are also likely to default on their payment. The company should reduce their credit limit and reduce the interests on purchases.

**Segment 4**

***Profile:*** This segment has the best credit score. They make the lowest monthly average cash advance amount. They make only installment purchases. They have a relatively better payment over minimum payment ratio as compared to customer segment 2. They form 19% of the customer base.

***Marketing Strategy:*** This segment of customers is the best performing among the rest. They have the best credit score and their payment over minimum payment is relatively good. This segment can be targeted by giving them more rewards to encourage purchase transactions.

**Conclusion**

The results of this study revealed that customers are grouped into four distinct clusters where most customer belonged to the first segment, followed by segment 3, segment 2 and then segment 4. This was achieved using the KMeans, Agglomerative and Spectral clustering. During the study, we also generated KPIs that helped to profile customers according to their purchasing tendencies. Using the results from this study will help AnalytixLabs to optimize marketing campaign costs, boost customer activity, reduce customer churn rate, improve customer experiences and enhance customer relationship management. It is highly recommended that AnalytixLabs incorporate the findings in this study in their marketing endeavors and also update the clustering model at least every quarter to observe any new behaviors from customers.

**References**

Abidar, L., Zaidouni, D., & Ennouaary, A. (2020). *Customer Segmentation With Machine Learning | Proceedings of the 13th International Conference on Intelligent Systems: Theories and Applications*. http://dl.acm.org.elibraryashesi.remotexs.co/doi/abs/10.1145/3419604.3419794

Best, R. de. (2020, November 12). *Topic: Credit cards in the United States*. Statista. https://www.statista.com/topics/1118/credit-cards-in-the-united-states/

Gallo, A. (2014, October 29). The Value of Keeping the Right Customers. *Harvard Business Review*. https://hbr.org/2014/10/the-value-of-keeping-the-right-customers

Huang, D., Wang, C., Wu, J., Lai, J., & Kwoh, C. (2020). Ultra-Scalable Spectral Clustering and Ensemble Clustering. *IEEE Transactions on Knowledge and Data Engineering*, *32*(6), 1212–1226. https://doi.org/10.1109/TKDE.2019.2903410

Li, W., Wu, X., Sun, Y., & Zhang, Q. (2010). Credit Card Customer Segmentation and Target Marketing Based on Data Mining. *2010 International Conference on Computational Intelligence and Security*, 73–76. https://doi.org/10.1109/CIS.2010.23

Nugraha, A., Perdana, M. A. H., Santoso, H. A., Zeniarja, J., Luthfiarta, A., & Pertiwi, A. (2018). Determining The Senior High School Major Using Agglomerative Hierarchial Clustering Algorithm. *2018 International Seminar on Application for Technology of Information and Communication*, 225–228. https://doi.org/10.1109/ISEMANTIC.2018.8549834

Qiu, Q., Zhang, Q., & Guo, K. (2014). Grey Kmeans algorithm and its application to the analysis of regional competitive ability. *2014 IEEE 7th Joint International Information Technology and Artificial Intelligence Conference*, 249–253. https://doi.org/10.1109/ITAIC.2014.7065044

Umuhoza, E., Ntirushwamaboko, D., Awuah, J., & Birir, B. (2020). Using Unsupervised Machine Learning Techniques for Behavioral-based Credit Card Users Segmentation in Africa. *SAIEE Africa Research Journal*, *111*(3), 95–101. https://doi.org/10.23919/SAIEE.2020.9142602

Ying, L., & Yuanyuan, W. (2010). Application of clustering on credit card customer segmentation based on AHP. *2010 International Conference on Logistics Systems and Intelligent Management (ICLSIM)*, *3*, 1869–1873. https://doi.org/10.1109/ICLSIM.2010.5461312