Bank Client Segmentation

Dominique R. Grimes

June 30, 2024

Business Problem

This data science project is focused on enhancing bank marketing initiatives through completing customer segmentation modeling. Marketing initiatives are expensive and time consuming, but personalizing marketing efforts to targeted subsets of clients can make marketing more intentional and successful. There is fierce competition in the banking industry, and targeted marketing gives organizations a competitive edge. This scenario focuses on more impactful marketing strategies related to age and location based on transaction amounts of banking clients.

Background

The term "client segmentation" was coined by Wendell R. Smith in 1956. What was once a market filled with one or two options for a product, developed into choosing one of many versions of the same product. The saturation of like products required a way for individual products to stand out and appeal to a certain sector of the marketplace. For example, ketchup was no longer just ketchup. Fancy ketchup, kid ketchup, organic ketchup, and more emerged on the shelves over time (Virg, 2024). Identifying similar sectors assists in bridging the gap between product marketing strategy and the clients that appeal to the product most.

Data Explanation

The dataset obtained from Kaggle focuses on bank transactions in India from 2016 (Bansal, 2021). It contains over one million transactions and nine features. The features are TransactionID, CustomerID, CustomerDOB, CustGender, CustLocation, CustAccountBalance, TransactionDate, TransactionTime, and TransactionAmount.

The first step in preparing the data was identifying null values in the dataset. Table 1 shows four features with minimal null values in comparison to the total size of the dataset. I

decided to form clusters based only on complete cases, reducing the dataset by seven-thousand rows or 0.07%.

Table 1

Features and sum of null values

TransactionID	0
CustomerID	0
CustomerDOB	3397
CustGender	1100
CustLocation	151
CustAccountBalance	2369
TransactionDate	0
TransactionTime	0
TransactionAmount (INR)	0

After reviewing all the data types of the features, I noticed CustomerDOB and TransactionDate needed to be adjusted from object to datetime. I proceeded in splitting the TransactionDate into separate TransactionMonth, TransactionDay, and TransactionDOW features. For CustomerDOB, I created a new feature that calculated the age of the customer, CustAge. The original CustomerDOB and TransactionDate features were dropped.

The new CustAge feature needed additional cleansing. Ages ranged from negative fifty-eight to two hundred seventeen. I dropped ages less than or equal to zero as well as ages over one hundred. The data frame after this step contained slightly less than nine hundred thousand transactions.

The CustGender column had three values: M, F, T. While M and F were imbalanced, there was only one occurrence of T. Since there was not enough of this category to impact

clustering, I dropped the row and used get dummies to transform CustGender into a binary feature. CustLocation was also adjusted to numerical values using Encoder.

The features were reviewed again for validity. CustomerID was unique and TransactionTime was not in a format that I was able to convert and interpret. I made the determination to drop these two columns prior to analysis and pre-processing.

Analysis

I observed feature distributions and looked for outliers through univariate analysis of value counts, summary statistics, histograms, and boxplots. I also observed relationships between features through bivariate and multivariate scatter plots as shows below:

- TransactionDOW vs CustAccountBalance
- TransactionAmount (INR) vs CustAccountBalance
- CustLocation vs TransactionAmount (INR)
- CustLocation vs TransactionAmount (INR) & CustAccountBalance
- CustAge vs TransactionAmount (INR)

Methods

To segment bank clients, I leveraged K-Means ++ clustering. This unsupervised modeling method identifies trends or natural groupings within data that may be unseen or take vast amounts of time to detect manually. K-Means ++ randomly assigns the first centroid. It then pushes out subsequent centroids as far as possible from each other based on maximum squared distance (Sharma, 2024). K-Means ++ is an enhancement of K-Means clustering (see Appendix for K-Means algorithm).

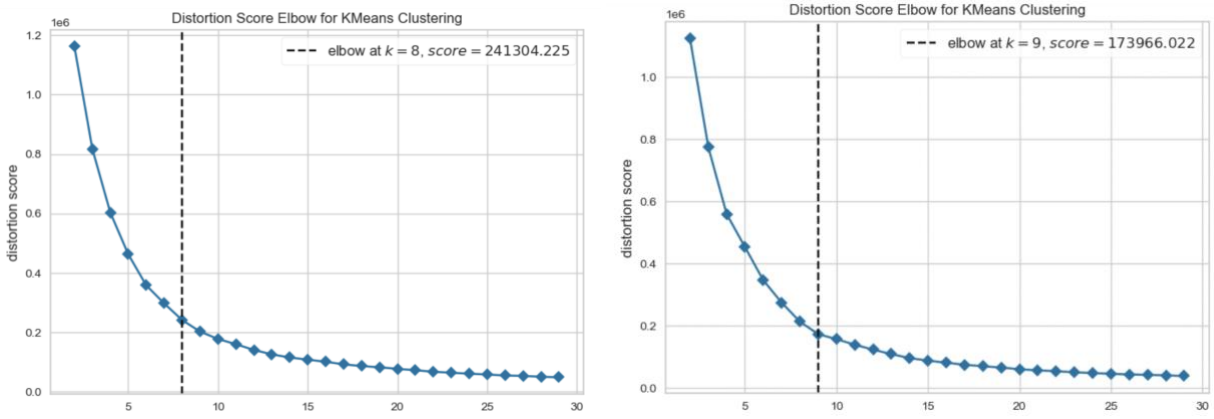
Prior to modeling, I utilized the pre-processing method of scaling the features utilizing StandardScaler. After standardizing the features, I was able to initiate the process of identifying

the optimal number of clusters to use. KElbowVisualizer provided a curve with the optimal number of clusters, also known as the elbow method, to apply in the centroid-based model.

I ran two different K-means ++ clustering scenarios focused on TransactionAmount. The first focused on clustering based on customer age with k=8 clusters. The second focused on clustering based on customer location with k=9 clusters.

Figure 1

Elbow method for Transaction Amount based on Age (L) and Location (R)



The model initializes the k centroids and assigns each data point to the closest centroid. For each cluster, the average of its assigned data points is calculated. This determines an updated location of the cluster centroid, and the data points are reassigned to the closest centroid.

Table 2 and Figure 2 show the results of the Client Age and Transaction Amount clustering, while Table 3 and Figure 3 summarize and visualize Client Location and Transaction Amounts. Clients that fall within the ranges and transaction amount summarized in the tables are generally thought to respond to similar marketing strategies.

Table 2

Customer Age and Transaction Summary by Cluster

K	Color	Age-Ave	Age-Min	Age-Max	Trans-Ave	Trans-Min	Trans-Max
0	red	27.8	26.0	30.0	834.52	0.00	12939.00
1	yellow	32.7	30.0	35.0	1152.31	0.00	12885.00
2	grey	31.0	17.0	43.0	88192.30	56340.85	247832.00
3	green	30.9	3.0	43.0	24110.12	12498.00	56000.00
4	blue	23.1	1.0	25.0	663.16	0.00	20034.00
5	orange	31.5	23.0	42.0	425609.40	265414.00	720001.16
6	pink	38.4	36.0	43.0	1378.55	0.00	15300.00
7	brown	39.3	38.0	40.0	1310390.00	991132.22	1560034.99

Figure 2

Scaled Scatter Plot of Customer Age and Transaction Amount Based on 8 Clusters

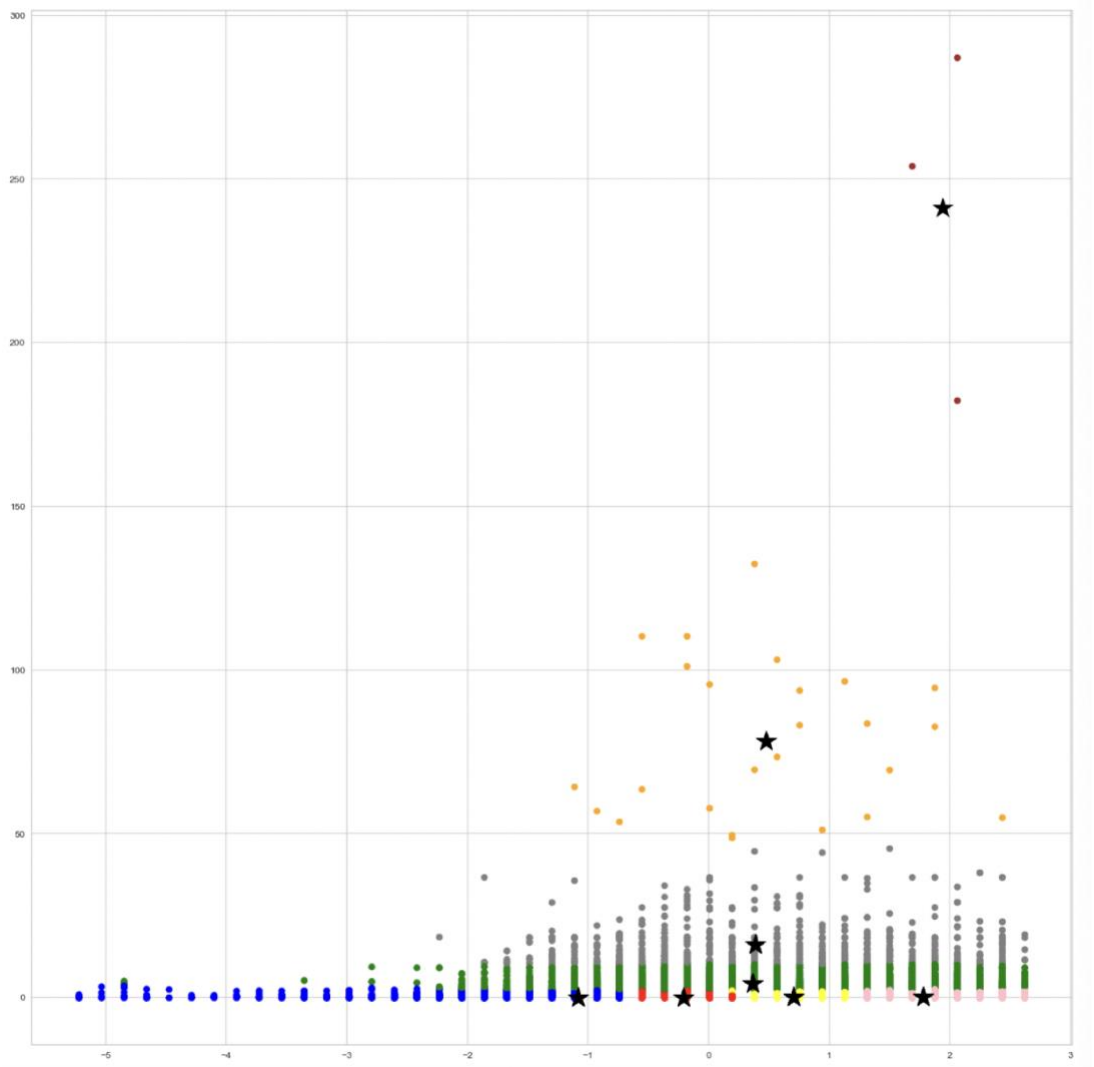


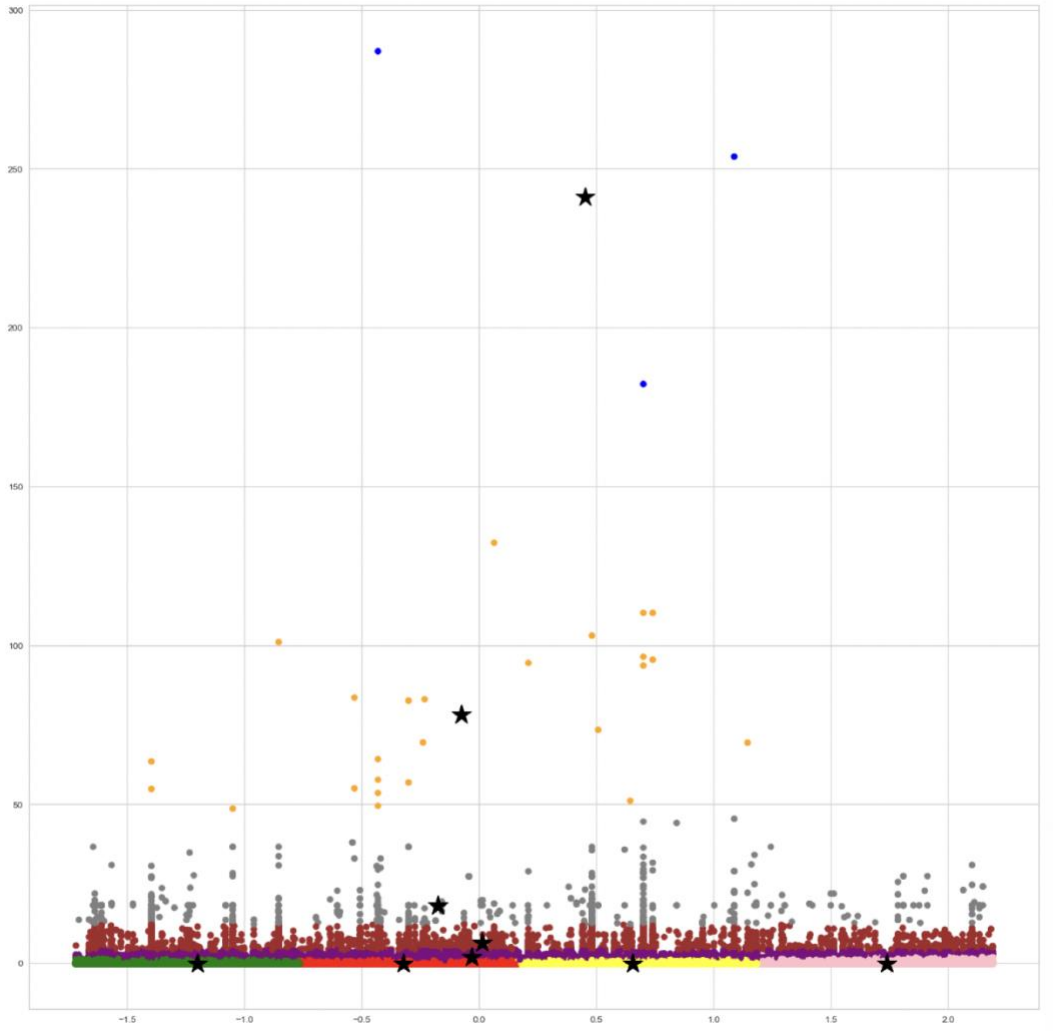
Table 3

Customer Location and Transaction Summary by Cluster

K	Color	#Locations	Trans-Ave	Trans-Min	Trans-Max
0	red	1789	761.08	0.00	6735.00
1	yellow	1937	846.78	0.00	7980.00
2	grey	121	99580.99	67470.00	247832.00
3	green	1858	788.34	0.00	9530.00
4	blue	3	1310390.00	991132.22	1560034.99
5	orange	16	425609.40	265414.00	720001.16
6	pink	1920	811.55	0.00	12375.00
7	brown	493	35422.72	23610.65	67291.00
8	purple	1326	11956.02	6132.00	23755.00

Figure 3

Scaled Scatter Plot of Customer Location and Transaction Amount Based on 9 Clusters



Conclusion

A major lifeline of business depends on client sales. Marketing is one tool that attracts clientele and enhances hit-ratio efforts of sales professionals. However, marketing can be expensive. While broad-brush, generalized marketing may help land a sale, specialized, targeted marketing is more efficient and impactful. Unsupervised learning modeling methods such as K-Means ++ can be applied to abundant, valid data to extract valuable insights for targeted marketing strategies. In this project we clustered clients from an Indian bank from 2016 based off age, location, and transaction amounts.

Assumptions

Assumptions were made when dropping values of the dataset. Even though the dataset was large, dropping values is likely to introduce bias. Bias is mitigated where the people with missing data are identical to the observations with complete data, except for the "missingness" of the data. That is, this approach involves making the strong Missing Completely at Random (MCAR) assumption (Bock, n.d.).

Limitations

The features for client demographics were limited. The transaction dates were also very sporadic and imbalanced. It is unclear whether the data is inclusive of an entire year of bank transactions or if some data was intentionally or unintentionally left out. This gap limits the ability to detect seasonality.

Challenges/Issues

Client bank data is private, secure, and heavily regulated. This makes public access to relevant data limited. Another challenge is managing the size of the data. In production, big data tools will most likely be needed to accommodate processing the size and frequency of the data.

One issue of this dataset is the questionable validity of the data. Without more information about the data source, it is a challenge to determine if outliers are valid client scenarios. Therefore, I did not remove questionable outliers that have the potential to be valid datapoints. Some of these outliers need to be reviewed since outliers can impact the location of the centroids and clustering outcomes.

Additional Applications

Three potential use cases for k-means clustering are fraud detection, document classification, and image segmentation. Document classification and image segmentation can group like text and pixel formations into clusters. In a slightly different approach, points that fall outside of a cluster as an anomaly can be leveraged to detect fraud (Virag, 2024).

Recommendations

I recommend putting data validation controls in place to ensure the validity of the source data utilized for segmentation. If possible, introduce additional demographic features that may enhance the model. While k-means is an efficient model meant to handle large sets of data, I recommend cloud computing power to process the data at a minimum.

With cleaner data, some of the clusters may be reduced due to eliminating outliers or invalid data. Reducing clusters leads to streamlined marketing strategies that will reduce costs associated with marketing efforts.

The feature with transaction data is right skewed due to a high frequency of zero transactions. It may be worth exploring log transforming this feature to see if it enhances the clustering results. In pre-processing, I suggest exploring how PCA transformation impacts clustering as well.

Implementation Plan

For client segmentation, a batch process on a periodic cadence should suffice to meet the objective of targeted marketing. I recommend periodic re-evaluation of the modeling results to update marketing strategies. Results need further analysis along with interpreted results communicated to stakeholders. This should be supplemented with data enablement through dashboard views for reference and end-user analysis.

Metrics need to be put into place to track and measure marketing initiatives generated because of the client segmentation model. Leadership will need to review the impact and effectiveness of the model to verify the value of the data collection and modeling effort.

Ethical Assessment

An ethical implication to consider is selection bias. It is essential to verify that the sampling utilized to create the model is representative of the entire client population. If not, the results may be biased. As previously mentioned, dropping incomplete client data has potential to introduce bias as well.

Be cognizant of classes or groups of clients getting more or less of a particular product offering than others due to the model. There is potential to cause disproportionate, unintentional discrimination towards certain groups of clients. This leads to unfair banking practices which are heavily regulated.

Bank client data is highly sensitive. Data protection regarding privacy must be top priority. Often data is gathered without the explicit consent of the consumer. More regulations have been implemented over the past several years due to data collection and privacy infringement. The client needs to have control over what data is being collected, how it is used, and how it is shared.

References

Arvai, K. *K-Means Clustering in Python: A Practical Guide*. Real Python. https://realpython.com/k-means-clustering-python/

Bansal, S. (2021) *Bank Customer Segmentation (1M+ Transactions).* Kaggle. https://www.kaggle.com/datasets/shivamb/bank-customer-segmentation

Bock, T. *How to Deal with Missing Values in Cluster Analysis.* DISPLAYR. https://www.displayr.com/deal-missing-values-cluster-analysis/

Hosni, Y. (2024, March 28). *Top 10 Data Science Project Ideas in 2024.* 365 Data Science. https://365datascience.com/career-advice/top-10-data-science-project-ideas/#5

Hussien, A. (2022). *Customer Segmentation Using Four Clustering Types.* https://www.kaggle.com/code/abduallahhussien/customer-segmentation-using-four-clustering-types

Sharma, N. (2024, April 15). *K-Means Clustering Explained*. neptune.ai. https://neptune.ai/blog/k-means-clustering#:~:text=K%2Dmeans%2B%2B%20is%20a,on%20the%20maximum%20squared%20distance.

Virag, J. (2024, May 20). *The No-Nonsense Guide to Market Segmentation (With Tips and Examples).* Nutshell. https://www.nutshell.com/blog/market-segmentation

Appendix

Algorithm 1 k-means algorithm

1: Specify the number k of clusters to assign.
2: Randomly initialize k centroids.
3: **repeat**
4: **expectation:** Assign each point to its closest centroid.
5: **maximization:** Compute the new centroid (mean) of each cluster.
6: **until** The centroid positions do not change.

(Arvai, n.d.)