Data Science and Analytics

Indian Bank Transactions: RFM Clustering and CLTV Prediction Model

Individual Report

Abbiegael Klara Go Chu
5-21-2023

Table of Contents

I. Summary

This report analyzes an Indian bank dataset with over one million bank transactions to gain insights on customer segmentation, customer value, and customer lifetime value. The project found four key customer segments, namely the Occasional Spenders, Potential Churners, Loyal Customers, and VIP Customers, characterized by their recency, frequency, and monetary scores in a clustering algorithm. The customer value analysis highlights the average value per customer for each segment. VIP Customers have the highest average value per customer with the smallest percentage of customers, while Occasional Spenders had the largest customer base and contributes to most of the bank's business. Among the three predictive models, the random forest model appeared to have the best accuracy for predicting the customer lifetime value due to having the lowest mean squared error and mean absolute error, and highest R-squared error.

II. Introduction

Banks have become a necessity for most people in their everyday lives. On banks' ends, they need to keep accurate records of all their customers' transactions, from transaction date and time to merchant to amount. While on the customers' ends, they also benefit from these transaction records being kept by the bank to assist with accounting. With this, in the vast amounts of bank transactions, data scientists can perform several machine learning projects to help learn more about the bank's customers.

III. Data Description

The data used for this project was extracted from Kaggle. Link to the site is https://www.kaggle.com/datasets/shivamb/bank-customer-segmentation. The dataset contains more than one million transactions by over eight hundred million bank customers in India. The dataset contains nine columns: Transaction ID, Customer ID, Customer Date of Birth, Customer Gender, Customer Location, Customer Account Balance, Transaction Date, Transaction Time, and Transaction Amount.

Transaction ID columns is the unique identifier for each transaction. Customer ID is the unique identifier for each customer. Customer Date of Birth is the customer's date of birth. Customer gender is the customer's gender. Customer location provides information on the customer's city/ location. Customer account balance is the balance amount in the customer's bank account. Transaction Date and time are the timestamps of when the transaction occurred. Transaction Amount is how much money was involved in the transaction.

During the data cleaning process, any rows with missing values were taken out. Given that the total transactions are more than one million and the total rows with missing values amounted to less than 1% of the data, the rows with missing values were dropped. Customer's date of birth was examined, and the researcher found that there were almost 57 thousand rows with January 1, 1800, which is impossible. These rows were dropped. With the birthdates with plausible birthdates, the researcher created a new 'Age' row that got the difference of today's date to the customer's date of birth.

Customer gender, customer account balance, customer location, transaction date, transaction time, and transaction amount seemed to have no odd values. Hence, no feature engineering or row dropping were done.

IV. Business Questions

1. Define Customer Segmentation among bank customers using RFM.
2. Create a Customer Value Analysis based on Customer Segmentation.
3. Identify and compare machine learning models to use for a Customer Lifetime Value prediction.

V. Research Findings

A. RFM Clustering

RFM means Recency, Frequency, and Monetary value, and each value is based on key customer traits. The RFM scores are important to any business due to their insights on customer behavior, which are based on frequency and monetary, and customer retention, which is based on recency. (Nurhamid, 2020) Clustering is a type of unsupervised learning, where the program divides the dataset into similar groups of data. (Priy, 2023)

For the RFM Clustering, customers were groups into four segments, A, B, C, and D. In this study, customer recency, frequency, and monetary scores were used as features for clustering.

Cluster A could be described as the “Occasional Spenders” or the low-value customers. These customers only make occasional transactions with the bank, and they have lower overall spending. This group has moderate recency, low frequency, and low to moderate monetary value.

Cluster B are the “Potential Churners” or mid-value customers. This cluster had a moderate to high recency score, moderate frequency score, and high monetary value. Compared to the other clusters, these customers have a higher likelihood to churn because of their recency scores. However, they still give a decent amount of revenue to the bank based on their monetary scores.

Cluster C are the “Loyal Customers” or high-value customers. This cluster had a moderate to high recency score, high frequency score, and high monetary score. These customers seem to be loyal to the bank, using always using the bank for their day-to-day transactions. They also contribute a significant amount of revenue to the bank.

Cluster D are the “VIP Customers” or high-spending customers. This cluster had high recency, high frequency, and very high monetary scores. These customers are highly valuable people to the bank, making recent purchases, often uses the bank for their transactions, and moves a lot of money.

B. Customer Value Analysis

Based on the RFM Clustering done, four types of clusters were identified. The “Occasional Spenders” comprise of 884,722 customers of the bank. The “Potential Churners” make up 1,227 of the total customers, while the “Loyal Customers” were a total of 15,407. And the “VIP Customers” only 27 customers in the group.

"VIP Customers" were described as having very high monetary scores, and an average value of a Cluster D customer is INR 576,354. However, since the cluster only comprises of 27 customers, the monetary contribution of Cluster D is only 1% in relation to the other segments. "Loyal Customers" have an average value of INR 23,149 and make up 23% of monetary contributions, while the "Potential Churners" are valued at INR 92,462 per person on an average and make up 7% of monetary contributions. "Occasional Spenders" or Cluster A has an average value of INR 1,179 per person, but due to the sheer volume of the cluster, it contributes 68% of the bank's business.

C. Customer Lifetime Value

Customer Lifetime Value is a metric that indicates how much customers spend over their lifetime with the business. (Hu, 2020) In this case, since the business is a bank, the customer lifetime value was calculated by multiplying the average transaction value per customer, transaction frequency, and their recency. Based on the RFM Clustering, the customer lifetime values and the statistics for each segment are:

Cluster	Count	Mean	Std	Min	25%	50%	75%	Max
A	765,131	127,491	252,565	0	13,080	44,157	128,546	4,053,165
B	937	11,133,130	9,752,138	0	5,455,213	8,925,000	14,954,128	80,710,545
C	11,457	2,756,617	24,946,49	0	1,187,792	2,150,000	3,818,880	17,808,000
D	20	71,604,230	63,362,930	21,492	26,568,780	48,210,590	117,346,787	198,720,320

Table 1 CLTV by Cluster

VI. Analytical Findings

A. RFM Clustering

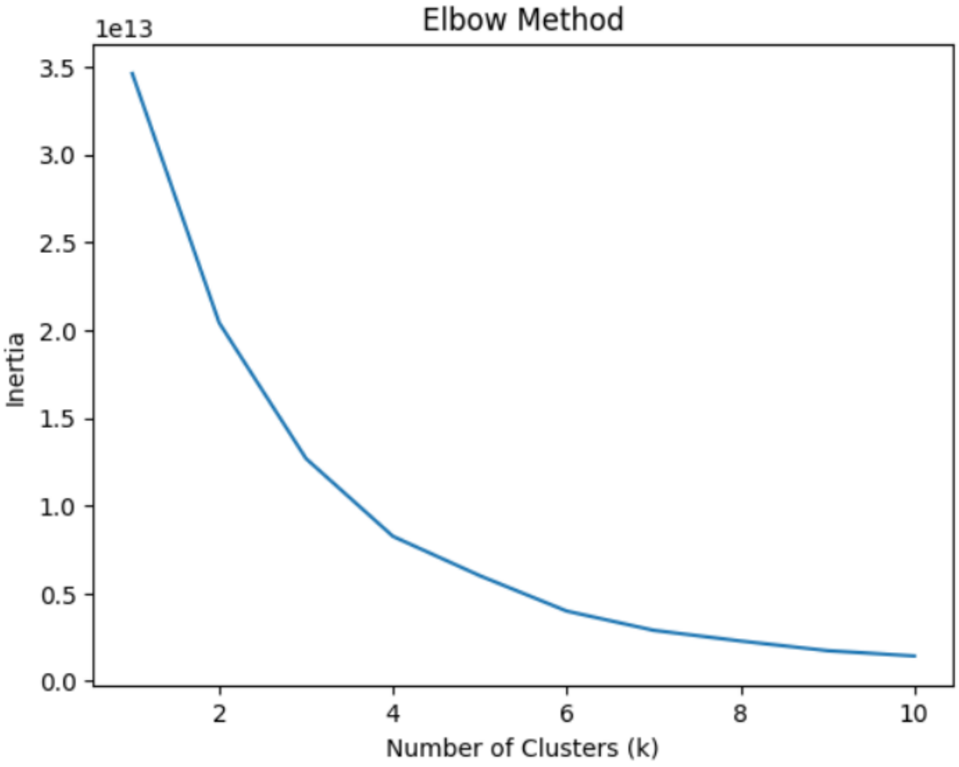

Figure 1 K Means Elbow Method

K-means Clustering was used on this dataset, with Recency, Frequency, and Monetary scores used for the parameters. K-means clustering is an unsupervised learning algorithm used for clustering problems and segments the dataset into k-number of clusters, where each datapoint is put into a cluster that is nearest to its mean. (Priy, 2023) Recency was calculated as January 1, 2017 minus the transaction date. Frequency counted the number of times Customer IDs appear. Monetary summed the transaction amounts based on Customer ID. To determine the number of K in K means, the elbow method was applied. Based on Figure 1 K Means Elbow Method, the optimal number of clusters for this data is 4.

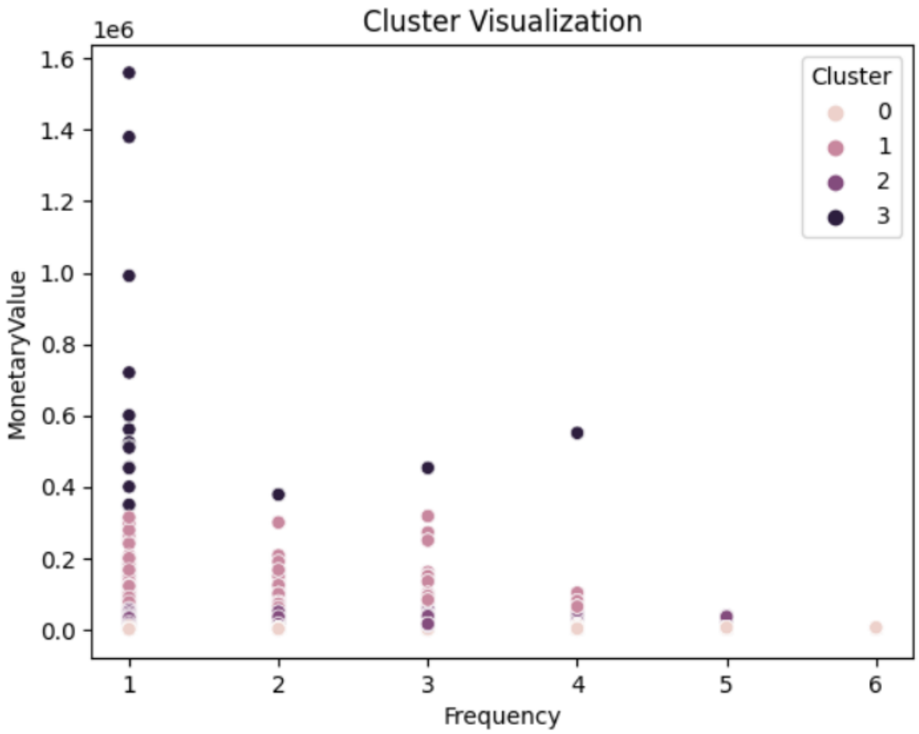

Figure 2 Cluster Visualization

After plotting the clusters, Figure 2 Cluster Visualization shows that Cluster D had large transactions but with one or two frequencies. The clusters also have these descriptive statistics:

Cluster	Recency		Frequency		Monetary	
	Mean	Median	Mean	Median	Mean	Median
A	159.54	134	1.30	1.0	1179.88	528.0
B	161.12	134	1.56	1.0	92462.31	85699.0
C	161.17	135	1.61	1.0	23149.16	20000.0
D	162.81	132	1.81	1.0	576354.79	514320.0

Table 2 Cluster RFM Statistics

B. Customer Lifetime Value

Linear regression examines the relationship between a dependent variable and one or more independent variable. It assumes a linear relationship between the identified variables. The model is used to predict the average value of dependent variables if given information about the independent variables. (Jidge, 2020) According to (What is a Decision Tree?, n.d.), a decision tree is a supervised learning algorithm that uses both classification and regression. The model starts with a root node and branches off to internal nodes. From the internal nodes, leaf nodes branch out. The cycle continues until all the features and have been branched out. A variant of the decision tree is the random forest model. The random forest model is made up off an ensemble of decision trees. The model then merges the trees together to produce more accuracy. (Donges, 2023)

In this study, linear regression, decision tree regression, and random forest regression models were used to test and train the dataset. The features used for all three models are Recency, Frequency, Monetary Value, Gender, Account Balance, Age, Transaction Amount, and pre-identified Cluster from the previous analysis, while making the target variable be CLTV. The dataset was split into 80-20. 80% was used to train the model, while 20% was used to test the accuracy of the model. Mean Squared Error, Mean Absolute Error, and R-squared Error Score were used to evaluate the error and accuracy per model.

	Linear Regression	Decision Tree	Random Forest
Mean Squared Error	174,454,667,017.04	26,053,962,838.57	13,869,947,689.29
Mean Absolute Error	108,778.90	1,946.21	1,189.57
R-Squared Score	0.7678	0.9653	0.9815

Table 3 CLTV Model Evaluation Results

In Table 3 CLTV Model Evaluation Results, based on the Mean Squared Error and Mean Absolute Error, the Random Forest model did the best, getting the lowest value and indicating the best accuracy among the three. For R-Squared Score, the Random Forest model also scored the best, getting the highest value of 0.9815, indicating that the model is a best fit for the dataset. Overall, the Random Forest model performed the best among the three and should be considered when determining the customer lifetime value for this bank.

VII. Recommendations

Based on the RFM Clustering, it seems that the bank can assign personal bankers to Cluster D since it only contains 27 high value customers. Better customer service is needed to maintain great relationship with these customers. Based on the customer segmentation, it seems that the bank has a sizable number of loyal customers in Cluster C. A strategy to maintain customer loyalty is to have a loyalty program, wherein these customers are rewarded or are given small incentives to continue using the bank for their transactions. Cluster B is seen to have high transaction amounts, even compared to Cluster C. These potential churners could be small business owners that use their personal bank accounts as a means of payment for their customers. The bank could direct marketing campaigns to them for B2B clients. However, Cluster B being small business owners is merely a speculation. More information should be included in the dataset, like transaction notes and who the money came from or was sent to. Cluster A has the largest number of customers and the largest monetary contribution among the segments. Since this cluster has low frequency, the bank could improve on their services to make sure that there is the least amount of friction for customers.

For Customer Lifetime Value prediction, it is recommended to use the Random Forest model since it had the least squared error and the highest R-squared score. Though the Decision Tree model also scored well, using the Random Forest model reduces the overfitting issue found in decision tree. Moreover, since bank have a relatively large dataset, Random Forest is more suitable for this situation. (Sharma, 2020)

VIII.Conclusions

The project used an Indian bank dataset from Kaggle to explore its customer segmentation, analyze its customer value analysis, and predict customer lifetime value. Through an RFM-based Clustering analysis, the researcher found four segments, which are Cluster A: Occasional Spenders, Cluster B: Potential Churners, Cluster C: Loyal Customers, and Cluster D: VIP Customers. The customer value analysis highlights the monetary contributions of each customer segment and the distribution of customer in each segment. For the customer lifetime value prediction, among the three models used, Random Forest performed the best due to displaying the lowest mean squared error and mean absolute error, and the highest R-squared error. Overall, based on the findings in this project, the Indian bank can turn the data into actionable insights to improve customer relation and increase customer lifetime value.

IX. References

Donges, N. (2023, March 14). *Random Forest: A Complete Guide for Machine Learning*. Retrieved from Built In: https://builtin.com/data-science/random-forest-algorithm

Hu, W. (2020, November 28). *Understanding and Forecasting Customer Lifetime Value (CLTV)*. Retrieved from Towards Data Science: https://towardsdatascience.com/understanding-and-forecasting-customer-lifetime-value-cltv-634fe34f522b

Jidge, A. (2020, May 25). *The Complete Guide to Linear Regression Analysis*. Retrieved from Towards Data Science: https://towardsdatascience.com/the-complete-guide-to-linear-regression-analysis-38a421a89dc2

Nurhamid, A. (2020, July 20). *Customer Segmentation with RFM Analysis & Kmeans Clustering*. Retrieved from Medium: https://medium.com/analytics-vidhya/customer-segmentation-with-rfm-analysis-kmeans-clustering-32c387d04dfe

Priy, S. (2023, May 9). *Clustering in Machine Learning*. Retrieved from Geeks for Geeks: https://www.geeksforgeeks.org/clustering-in-machine-learning/

Sharma, A. (2020, May 12). *Random Forest vs Decision Tree | Which Is Right for You?* Retrieved from Analytics Vidhya: https://www.analyticsvidhya.com/blog/2020/05/decision-tree-vs-random-forest-algorithm/#How_to_Choose_Between_Decision_Tree_&_Random_Forest?

What is a Decision Tree? (n.d.). Retrieved from IBM: https://www.ibm.com/topics/decision-trees