Using Unsupervised Machine Learning Techniques for Behavioural Based Bank Customer Segmentation

Neelavathi Shanmugasundaram
MSc Big Data Analytics and Artificial Intelligence,
Letterkenny Institute of Technology,
L00163479@student.lyit.ie

***Abstract*—As we are in the competitive digital world, Understanding of customers transaction behaviour is important for segmenting which will help to predict and provide them suitable banking services. The aim of this technical project is to find common behavioural characteristics among customers of an Indian bank and segmented based on their RFM (Recency, Frequency, Monetary) value along with Customer age and and their bank account balance.K-means and Bisecting K- means machine learning algorithms were used to segment the bank customers. Elbow method is used to find out the optimal cluster value and their performance were measured using Silhouette with squared euclidean distance score. Colab interface is used for code development in Pyspark. Matplotlib and Seaborn libraries were used for visualizing and Exploratory data analysis.**

***Index Terms*—Customer Segmentation, Unsupervised machine Learning, RFM analysis, Pyspark, Big Data Analysis, Spark ML K-means, Bisecting K-Means**

I. Introduction

IN today's competitive business environment, Ultimate aim of every business is to plan clear strategies in order to provide appropriate services to each segment of customer to retain the existing customers and also able to identify future potential customers [1].Customer segmentation can be done by using RFM analysis. RFM model distinguish the customers based on their behavioural characteristics such as recent transaction made by the customer, Frequency of transactions in given time period and monetary value of each transaction [2]. The reason for using RFM model is that is easy to use and efficient implementation in different industrial sectors [3]. Customer segmentation will be profitable to the bank in the way if the bank has introduced new scheme or new product, Bank can find target customers through segmentation, bank can plan its strategies which will increase the bank profit [4].

A. Problem statement

The goal of this technical project is to answer below hypothesis by analyzing Indian bank transaction dataset and identify different segments of existing customer using clustering algorithms such as K-Means and Bisecting K-means algorithm.

1) how many segments of customer are there in the bank currently?
2) What are the characteristics of each customer segments?
3) What are the inferences made from analysis that can help bank to predict best action for each group?

II. Methodology

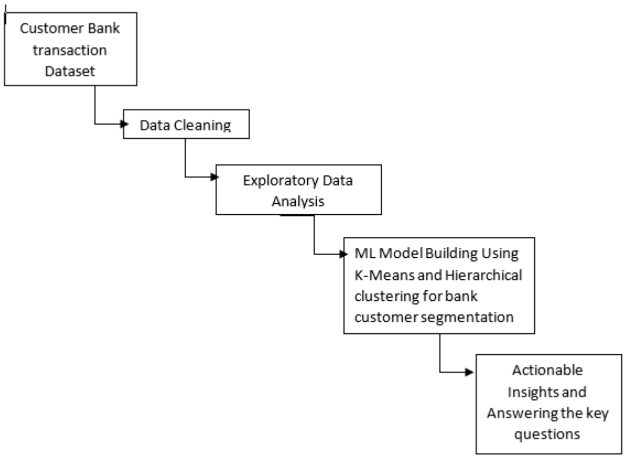

Customer Bank transaction Dataset
Data Cleaning
Exploratory Data Analysis
ML Model Building Using K-Means and Hierarchical clustering for bank customer segmentation
Actionable Insights and Answering the key questions

Fig. 1: Data Analysis Pipeline

A. Dataset description

The Data for this project is obtained from Kaggle, banking transaction's dataset of an Indian bank from August 2016 to October 2016 and has data volume of 1 million transactions and comprising of about 800k customer details such as their

Under the supervision of Dr. Shagufta Henna, Letterkenny Institute of Technology, Letterkenny, CO. Donegal age, location, customer transaction details such as transaction amount, Transaction time, Account balance at the time of every transaction.Below were the fields of the dataset.

TABLE I: Data Description

Sr.no	column	null count	Data Type
1	TransactionID	0	string
1	CustomerID	0	string
2	CustomerDOB	0	string
3	CustGender	1100	string
4	CustLocation	151	string
5	CustAccountBalance	2369	double
3	TransactionDate	0	string
3	TransactionTime	0	integer
3	TransactionAmount	0	double

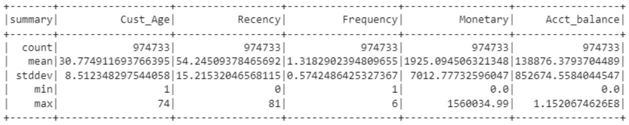
B. Data Pre-processing and Preparation

Visualizing dataset will help in finding insights from it that can be used for preprocessing.

1) From Table 1. we can see certain columns have null values. The columns with null values amount to 1 percent of total data, hence dropping the rows with null columns.
2) Transaction ID column has no effect on data analysis, hence dropping the Transaction ID column
3) There are a lot of birth date on 1/1/1800, which seems to be default birth date for customers without date of birth detail, hence removed rows with that value as column
4) In order to calculate the age of each customer, converted the format of columns CustDOB, TransactionDate into standard pyspark date format format.The recorded details were from the year 2016, in order to find the age, firstly calculated maximum transaction date and subtracted from date of birth of each customer

C. RFM (Recency Frequency Monetary)

For behavioural segmentation of customers, RFM attributes needed to be calculated.In order to provide more efficient segmentation of customers, age and account balance of customers were also taken into consideration.



```
+-------+-----------------+-----------------+------------------+-----------------+-----------------+
|summary|          Cust_Age|          Recency|         Frequency|         Monetary|      Acct_balance|
+-------+-----------------+-----------------+------------------+-----------------+-----------------+
|  count|           974733|           974733|            974733|           974733|            974733|
|   mean|30.774911693766395|54.24509378465692|1.3182902394809655|1925.094506321348|138876.3793704489|
| stddev| 8.512348297544058|15.21532046568115|0.5742486425327367| 7012.77732596047|852674.5584044547|
|    min|                1|                0|                 1|              0.0|              0.0|
|    max|               74|               81|                 6|       1560034.99|   1.1520674626E8|
+-------+-----------------+-----------------+------------------+-----------------+-----------------+
```

Fig. 2: Dataset statistics after Pre-processing

TABLE II: Data count before and after

	Data Count
Before (Pre-Processing)	1048567
After (Pre-Processing)	974733

D. System Architecture

This Project is build on Apache spark framework which efficiently process large volume of data in collaboration with other interface and distributed computing Tools. Colab interface is used to develop and test pyspark-(Apache Spark which supports Python)code.

III. Exploratory Data Analysis

For Exploratory Data Analysis python data visualization libraries were used. In order to plot graph and charts more efficiently in colab, Pyspark dataframe is converted to pandas dataframe using Topandas() method. This action will collects data from all Pyspark clusters and converts them to pandas dataframe. Matplotlib and seaborn libraries were used for data analysis to understand key insights from the dataset for modelling. Below are the inferences made from analysis that can help bank to predict best action for each group:

Age range of customers: Inference: From the Fig.3 we can clearly see that majority of customers of the bank were in the age range of 25 to 30, followed by age range of 30 to 35.

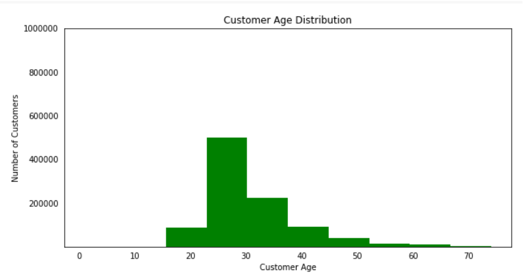


Fig. 3: Customer Age Distribution

Frequency of transactions: Inference: From the Fig.4,Fig.5 customers at all age group has made recent transactions, it is

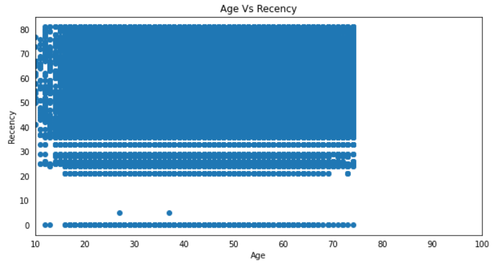


Fig. 4: Customer Age vs Recency

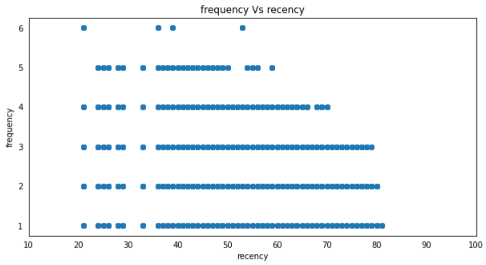


Fig. 5: Frequency vs Recency

clearly not enough to segment the customers based on their age

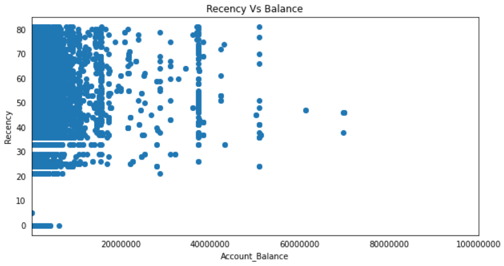


Fig. 6: Recency Vs Balance

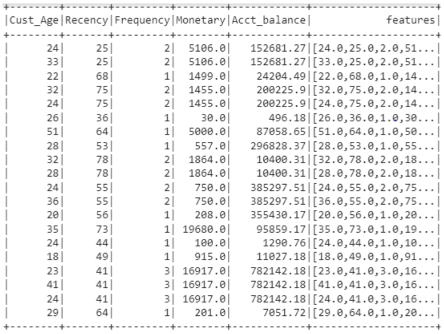
From previous inferences we can clearly see there are no unique characteristics of each customer, hence comparing the recency with bank balance to identify potential customer. from fig.6, we can infer that account balances dominate in the region of 2 crore and done more than 3 recent transactions showing that majority of the population were moderately rich. Key findings from EDA:

1) 75 percent of bank customers aged between 30 to 35, with high frequency and recency of transaction and also maintains high balance.
2) 50 percent of the bank customers aged between 25 to 30 and with medium frequency and recency of transactions in given time frame with account balance lesser than the previous segment.
3) 25 percent of the bank customers aged below 25 with minimum frequency and recency of transaction and maintains very low bank account balance.

IV. Data Modelling and Interpretation

A. Feature Vector

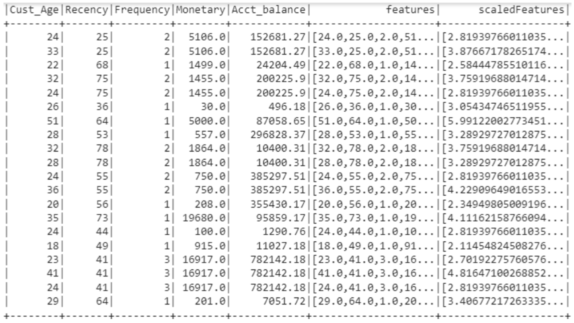
From EDA, we have made inferences to find important features from the dataset. The next step of the data pipeline is to Model the data for interpretation. Spark ML libraries were used for data modelling. In order to enable the machine learning model to be more productive, Feature vectorization is essential which will convert all features into a single multi dimensional array of features, which will be used by machine learning models to interpret data and provide better predictive analytics. Vectorassembler() Spark transformer technique has been used in this project



```
+--------+-------+---------+--------+-----------+--------------------+
|Cust_Age|Recency|Frequency|Monetary|Acct_balance|            features|
+--------+-------+---------+--------+-----------+--------------------+
|      24|     25|        2|  5106.0|  152681.27|[24.0,25.0,2.0,51...|
|      33|     25|        2|  5106.0|  152681.27|[33.0,25.0,2.0,51...|
|      22|     68|        1|  1499.0|   24204.49|[22.0,68.0,1.0,14...|
|      32|     75|        2|  1455.0|   200225.9|[32.0,75.0,2.0,14...|
|      24|     75|        2|  1455.0|   200225.9|[24.0,75.0,2.0,14...|
|      26|     36|        1|    30.0|     496.18|[26.0,36.0,1.0,30...|
|      51|     64|        1|  5000.0|   87058.65|[51.0,64.0,1.0,50...|
|      28|     53|        1|   557.0|  296828.37|[28.0,53.0,1.0,55...|
|      32|     78|        2|  1864.0|   10400.31|[32.0,78.0,2.0,18...|
|      28|     78|        2|  1864.0|   10400.31|[28.0,78.0,2.0,18...|
|      24|     55|        2|   750.0|  385297.51|[24.0,55.0,2.0,75...|
|      36|     55|        2|   750.0|  385297.51|[36.0,55.0,2.0,75...|
|      20|     56|        1|   208.0|  355430.17|[20.0,56.0,1.0,20...|
|      35|     73|        1| 19680.0|   95859.17|[35.0,73.0,1.0,19...|
|      24|     44|        1|   100.0|    1290.76|[24.0,44.0,1.0,10...|
|      18|     49|        1|   915.0|   11027.18|[18.0,49.0,1.0,91...|
|      23|     41|        3| 16917.0|  782142.18|[23.0,41.0,3.0,16...|
|      41|     41|        3| 16917.0|  782142.18|[41.0,41.0,3.0,16...|
|      24|     41|        3| 16917.0|  782142.18|[24.0,41.0,3.0,16...|
|      29|     64|        1|   201.0|    7051.72|[29.0,64.0,1.0,20...|
+--------+-------+---------+--------+-----------+--------------------+
```

Fig. 7: Feature Vectorization

Standardscaler() is used to resize the distribution of values, normalize the features value in the range of mean of observed values as 0 and the standard deviation as 1.



```
|Cust_Age|Recency|Frequency|Monetary|Acct_balance|            features|      scaledFeatures|
+--------+-------+---------+--------+-----------+--------------------+--------------------+
|      24|     25|        2|  5106.0|  152681.27|[24.0,25.0,2.0,51...|[2.81939766011035...|
|      33|     25|        2|  5106.0|  152681.27|[33.0,25.0,2.0,51...|[3.87667178265174...|
|      22|     68|        1|  1499.0|   24204.49|[22.0,68.0,1.0,14...|[2.58444785510116...|
|      32|     75|        2|  1455.0|   200225.9|[32.0,75.0,2.0,14...|[3.75919688014714...|
|      24|     75|        2|  1455.0|   200225.9|[24.0,75.0,2.0,14...|[2.81939766011035...|
|      26|     36|        1|    30.0|     496.18|[26.0,36.0,1.0,30...|[3.05434746511955...|
|      51|     64|        1|  5000.0|   87058.65|[51.0,64.0,1.0,50...|[5.99122002773451...|
|      28|     53|        1|   557.0|  296828.37|[28.0,53.0,1.0,55...|[3.28929727012875...|
|      32|     78|        2|  1864.0|   10400.31|[32.0,78.0,2.0,18...|[3.75919688014714...|
|      28|     78|        2|  1864.0|   10400.31|[28.0,78.0,2.0,18...|[3.28929727012875...|
|      24|     55|        2|   750.0|  385297.51|[24.0,55.0,2.0,75...|[2.81939766011035...|
|      36|     55|        2|   750.0|  385297.51|[36.0,55.0,2.0,75...|[4.22909649016553...|
|      20|     56|        1|   208.0|  355430.17|[20.0,56.0,1.0,20...|[2.34949805009196...|
|      35|     73|        1| 19680.0|   95859.17|[35.0,73.0,1.0,19...|[4.11162158766094...|
|      24|     44|        1|   100.0|    1290.76|[24.0,44.0,1.0,10...|[2.81939766011035...|
|      18|     49|        1|   915.0|   11027.18|[18.0,49.0,1.0,91...|[2.11454824508276...|
|      23|     41|        3| 16917.0|  782142.18|[23.0,41.0,3.0,16...|[2.70192275760576...|
|      41|     41|        3| 16917.0|  782142.18|[41.0,41.0,3.0,16...|[4.81647100268852...|
|      24|     41|        3| 16917.0|  782142.18|[24.0,41.0,3.0,16...|[2.81939766011035...|
|      29|     64|        1|   201.0|    7051.72|[29.0,64.0,1.0,20...|[3.40677217263335...|
+--------+-------+---------+--------+-----------+--------------------+--------------------+
```

Fig. 8: Feature Standardization

B. Unsupervised Machine Learning models for interpretation

To perform Customer Segmentation, K-means and Bisecting K-means unsupervised machine learning techniques were used.

C. K-means

K-means model is one of the majorly used cluster modelling technique. K-means is an unsupervised learning algorithm and has many number of application. In order to implement the model, Optimal K value has to be predetermined. For this project,Elbow technique is used to find optimal k value.

1) Elbow Method: To determine optimal number of clusters to be formed by K-means and Bisecting K-means, Elbow method is used to determine the k value. Optimal value can be found from the graph generated which looks like a elbow, the point where the elbow is formed is considered as a optimal K value. From the Fig.9, We can see that elbow is formed at two points. From EDA, we have categorized customers into 3 types, hence number of clusters to use in machine learning algorithms were k=3

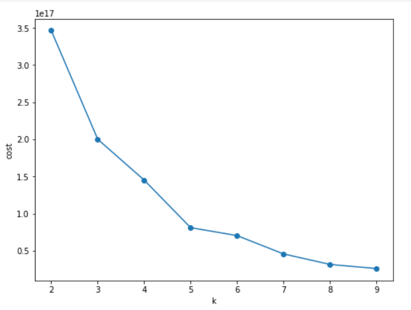
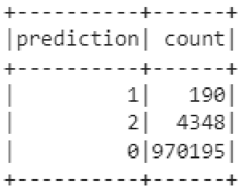


Fig. 9: Graph representing number of clusters using Elbow method

After determining the optimal k value, algorithm will initialize K points by shuffling the dataset and choosing K points for each centroid. After finding the centroids, data points will be assigned to each group. centroids will be calculated by taking average of datapoints belongs to each cluster.

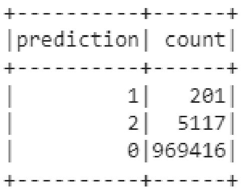


```
+----------+------+
|prediction| count|
+----------+------+
|         1|   190|
|         2|  4348|
|         0|970195|
+----------+------+
```

Fig. 10: Predictions summary of K-means

D. Bisecting K-means

Bisecting Clustering is combination of K-means and Hierachial clustering techniques. The main idea is dividing the dataset into one cluster and subdivide it to sub clusters until it finds the optimal centroids. Since we have decided the optimal K value as 3, it will be used for this technique as well, in order to compare the performance of different techniques for same K value.



```
+----------+------+
|prediction| count|
+----------+------+
|         1|   201|
|         2|  5117|
|         0|969416|
+----------+------+
```

Fig. 11: Predictions summary of Bisecting K-means

E. Validation of Models

Unsupervised machine learning models were validated based on Silhoutte Score. Silhoutte score calculates distance between centroids of each cluster, normally value will be between -1 to 1, where -1 represents poor clustering and values closer to 1 indiactes the model is properly clustered for the given data set.

Silhoutte score is calculated by the following formula. (n-i)/max(i,n) where n is the intra cluster distance and i is the mean of distance between each cluster.

TABLE III: Silhoutte score

Model	K-value	Silhoutte score
K-Means	K=3	0.9920135070352577
Bisecting K-means	K=3	0.9890936327069493

V. Results and Hypothesis Discussion

Based on the results of Exploratory data analysis and predictions made from the machine learning models which is visualized in Fig.12 we can derive answers for our proposed hypothesis.

1) Hypothesis 1: how many segments of customer are there in the bank currently?
   There are 3 different segments of customers in the Indian Bank
2) Hypothesis 2: What are the characteristics of each customer segments?

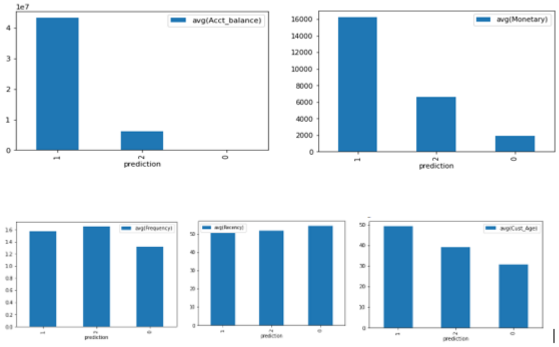


Fig. 12: Prediction Results for all dataset features

Group 1:
* Customers belong to this group have have average age of 48.
* Customers Done transaction for 0-1 times.
* Customers who have made high monetary transaction with low recency.
* Customers with average bank balance of 2 cr INR
* less than 1 percent of customers are in this group.
Group 2:
* Customers belong to this group have have average age of 39.
* Customers Done transaction for 0-1 times.
* Customers who have made less monetary transaction with low recency.
* Customers with average bank balance of 50 Lakhs INR
* Less than 1 percent of customers are in this group.
Group 0:
* Customers belong to this group have have average age of 30.
* Customers Done transaction for 0-1 times.
* Customers who have made less monetary transaction when compared to group 1 and 2 customers with low recency.
* Customers with average bank balance of 1 lakhs INR
* most of customers are in this group.

3) Hypothesis 3:
   What are the inferences made from analysis that can help bank to predict best action for each group?
   summary of Insights from 3 groups:
   *Group 1, Group 2 and Group 3 were behaving similar in terms of recency, frequency.
   *Group 1 is considered to be most valuable group as customers in that group maintaining high bank balance with high monetary transaction.
   *Group 2 is considered to be valuable customers with moderately high bank balance and with moderate monetary transaction but very less compared to that of group 1.
   *Majority of customers belong to group 0 with less bank balance and with less monetary transaction when compared to other two groups.

A. Marketing Strategies:

1) Group 0:: They are the potential target customers.There is high chances of customers in this group to leave the bank. Customers can be notified about new schemes and attractive credit offers through text messages and Email in order to attract these customers which will decrease the customer churn rate.

2) Group 1:: There are most valuable customers and bank can offer customer lifetime value (CLV) schemes to increase the brand loyalty.

3) Group 2:: This group of people considered to be valuable and bank can conduct feedback survey through call and based on their feedback schemes can be introduced to these group of people.

VI. Conclusion and Future works

This technical project analyzed existing bank customers transactional data of an Indian bank. By using unsupervised machine learning techniques, predicted existing groups of customer and provided strategies that the bank can implement to improve their business.

The performance of clustering models were evaluated using Silhoutte score. Based on the score K-means method outperformed Bisecting K-Means.

Characteristics of each groups were discussed and marketing strategies for each customers proposed in section.5.A

For Future works clustering can be performed based on Demographic information and segment customers using different machine learning clustering methods.

References

[1] M. Aryuni, E. Madyatmadja, and E. Miranda, “Customer segmentation in xyz bank using k-means and k-medoids clustering,” 09 2018, pp. 1–9.

[2] P. A. Sarvari, A. Ustundag, and H. Takci, “Performance evaluation of different customer segmentation approaches based on rfm and demographics analysis,” *Kybernetes*, 2016.

[3] A. M. Hughes, *Strategic database marketing*. McGraw-Hill Pub. Co., 2005.

[4] I. Maryani, D. Riana, R. D. Astuti, A. Ishaq, Sutrisno, and E. A. Pratama, “Customer segmentation based on rfm model and clustering techniques with k-means algorithm,” in *2018 Third International Conference on Informatics and Computing (ICIC)*, 2018, pp. 1–6.