Vinit Shankar Nagap - 002673657

Case Study

Customer Segmentation for Personalized Banking Services

Introduction

Customer segmentation for personalized banking services refers to the process of dividing a bank's customer base into distinct groups or segments based on their characteristics, behaviors, needs, and preferences. The goal is to better understand and categorize customers in order to tailor banking services and offerings to their specific requirements.

By analyzing customer data, including demographics, transaction history, online behavior, and financial preferences, banks can identify patterns and similarities among customers. These patterns are then used to create segments or clusters of customers who share similar traits or exhibit similar banking behaviors.

The purpose of customer segmentation is to enable banks to deliver personalized and targeted banking services to each customer segment. It allows banks to develop customized marketing strategies, product recommendations, and communication approaches that are more relevant and appealing to specific customer groups.

For example, a bank might identify segments such as young professionals, retirees, small business owners, or high net worth individuals. Each segment may have different banking needs and preferences. By segmenting customers, the bank can design tailored products and services, such as specialized savings accounts, investment options, or loan packages, to address the unique requirements of each segment.

Overall, customer segmentation for personalized banking services aims to enhance customer satisfaction, deepen customer relationships, and drive customer loyalty by providing individualized experiences and offerings based on customers' specific needs and preferences.

To address the problem statement of customer segmentation for personalized banking services, several research papers have been selected for review. These papers were chosen based on their relevance to the topic, the depth of their analysis, and their focus on supervised learning algorithms. One of the selected papers is "**Data Mining for Potential Customer Segmentation in the Marketing Bank Dataset" by Maulida Ayu Fitriani , Dany Candra Febrianto [1]**. The purpose of this paper is to find the most appropriate classification method for classifying customer responses to direct telephone marketing by banks in order to increase customer response from bank marketing officers. Therefore, accuracy is an important factor for determining direct marketing results.

Another chosen paper "**A Customer Segmentation Approach in Commercial Banks" by V. Mihovaa and V. Pavlovb [2]**. This article explores the introduction of so called "loyalty program", which includes the issuance of various types of cards for such customers. Three clusters (segments) of loyal borrowers: "platinum," "gold," and "silver," are identified in the present work, using K-means clustering.

Lastly, the research paper "**Customer Behavioral Segmentation at Banking System using Principal Components Analysis and Artificial Neural Network: The Quality Management" by Md.Sarwar Kamal, Mohd. Kamal Uddin [3]** in which we analyze the data set by K-Medoids Algorithm (KMA). Besides the KMA, we also imposed Fuzzy K-Means and Fuzzy K-modes algorithms. The Principal Component Analysis

(PCA) is finally used to correlate the appropriate data set. We noticed that K-Medoids Algorithms is better only for the linear and Quantitative data set. On the contrary, Fuzzy K-Means is better when there are lots of mixed data sets, I mean both for Qualitative and Quantitative. We suggest both of the algorithms for separate environments and policies.

These selected papers were chosen to gain a comprehensive understanding of the problem at hand and to explore the various approaches, techniques, and methodologies employed in customer segmentation for personalized banking services.

In conclusion, customer segmentation for personalized banking services is a significant challenge faced by banks today. By selecting and reviewing relevant research papers, we aim to gain insights into the techniques, algorithms, and strategies employed in customer segmentation to enable banks to deliver personalized experiences and enhance customer satisfaction.

Data Consideration

The dataset in **[1]** paper is the Bank Marketing Dataset which is the marketing data for a bank in Portugal. The dataset was obtained from the University of California at Irvine (UCI) Machine Learning Repository. Bank Marketing data contains 17 attribute data, 45,211 instance data and there are 2 classes. The descriptions of the datasets are described in Table below.

TABLE I
ATTRIBUTES ON THE BANK MARKETING DATASET

No	Attribut	Type	Values
1	Age	Numeric	Real
2	Job	Categorical	Admin, Unknown Unemployed, Management, Housemaid, Enterprenuer, Student, Bluecollar, Self-empoyed, Retired, Technican, Services
3	Marital	Categorical	Maried, Diforced, Single, Widowed
4	Education	Categorical	Secondary, Unknown, Primary, Tertiary
5	Default	Binary	Yes, No
6	Balance	Numeric	Real
7	Housing	Binary	Yes, No
8	Loan	Binary	Yes, No
9	Contact	Categorical	Unknown, Telephone, Cellular
10	Day	Numeric	Real
11	Month	Categorical	Jan, Feb …. Nov, Dec
12	Duration	Numeric	Real
13	Campaign	Numeric	Real
14	Pday	Numeric	Real
15	Previous	Numeric	Real
16	Poutcome	Categorical	Unknown, Failure, Success
17	Y	Binary	Yes, No

It is explained that the first thing to do is to collect the direct marketing bank dataset and then extract the data. This study also compares the final results if the training data is pre-processed and not pre-processed. Pre-processing is done for class balancing using the SMOTE method. Furthermore, classification and testing are carried out with test data using 10-fold cross validation. After evaluation, a comparison is made between various classification methods and the effect of pre-processing on classification.

In paper **[2]** a database of 100 borrowers from a commercial bank branch that took secured consumer loans is analysed. The clients are defined as loyal based on their credit history (they have less than 3 missed payments for the last year). Three variables are used for their segmentation. Initially, the initial segmentation variables are taken as input data for the analysis, and further study on standardized segmentation variables is carried out. The potential segmentation strategies are formulated depending on the leading segmentation variable. A comparative analysis of the results of both methods examined (with initial and standardized segmentation variables) and of a two-step clustering (obtained in a previous study from one of the authors) is made within each of the strategies. It is specified which type of cluster analysis suits best to each of the strategies.

Whereas in paper **[3]** they have visited five private bank and two government banks in Patiya regions of Chittagong city. We also talked with the Branch manager as well as the senior employers of the Bank. Mr.Ifteler branch manager of Brack Bank told that customers want the benefits from them in all sides especially when there are huge congestions. Some customers are very excited when they did not get quick responses. When customers are not getting their demands, they become very much furious to the respective employers. Mr. Sumon, a senior officer of the EXIM bank told us that the female customers are more polite but they do not have any idea on banking transaction.

Bank Name	Customer Name	Loan ID	Loan Account	Location
Brack Bank	Tapos Bormon	4001	124782	Patiya
Brack Bank	Afjal Hossain	4012	783141	Do
Brack Bank	Abdul Aziz	4902	321574	Do
Brack Bank	Hira Das	4092	492345	Do
Brack Bank	Morjina Begum	4982	438921	Do
Dhaka Bank	Earsad Ullah	8723	234120	Do
Dhaka Bank	Emam Uddin	8701	234045	Do
Dhaka Bank	Abul Khair	8728	230912	Do
Dhaka Bank	Towhidul Islam	8723	234570	Do
Dhaka Bank	Karim Uddin	8703	249072	Do
UCBL Bank	Jahedul Islam	5423	764848	Do
UCBL Bank	Nishita Shaha	3452	658392	Do
UCBL Bank	Moktar Hossain	4563	652310	Do
UCBL Bank	Abutahar Mia	4567	680213	Do
UCBL Bank	Khursad Fazil	7832	602343	Do
Janata Bank	Asis Mia	0923	984536	Do
Janata Bank	Farhan Ali	0876	908765	Do
Sonali Bank	Tipu Sultan	3424	214567	Do
Sonali Bank	Korim Monshi	3214	213468	Do
Agrani Bank	Absar Mia	0123	987566	Do
Agrani Bank	Josim Uddin	0342	982342	Do

Table 2: Customer information from different banks.

Methodology Comparison

In **[1]** pre-processing and then classification method is used before evaluating the dataset. In this SMOTE (Synthetic Minority Oversampling Technique) is used to solve the class imbalance problems.

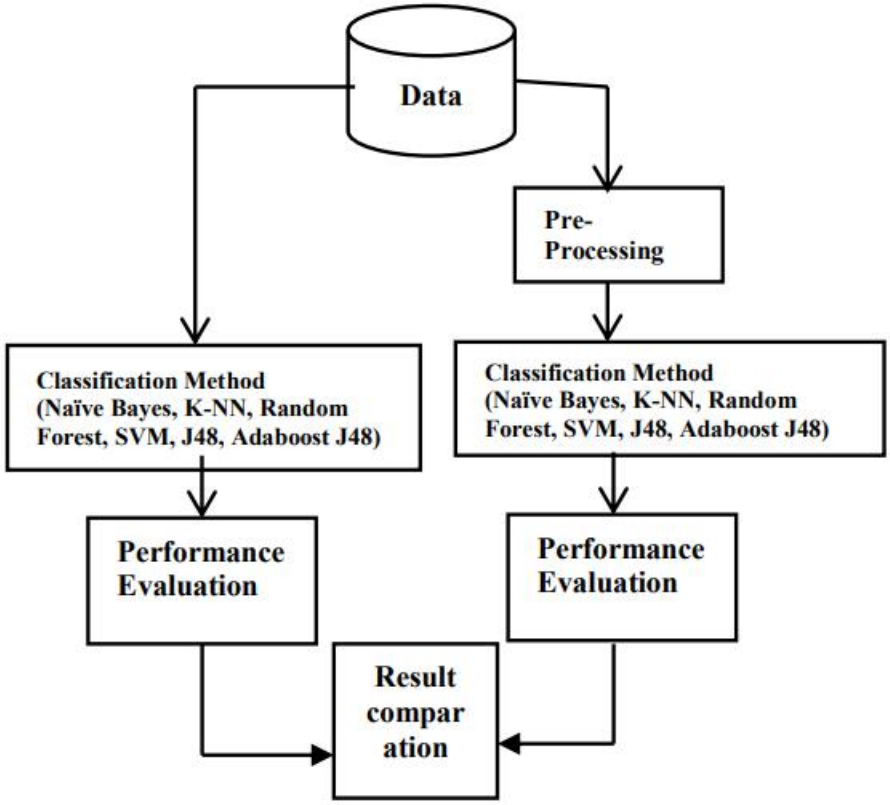


Fig. 1 Research flow

The principle of the SMOTE Method is to increase the number of data from the minority class so that it is equal to the majority class by generating artificial data. The artificial or synthetic data is made based on the k-nearest neighbor. Generating artificial data with numeric scale is different from categorical. Numerical data are measured for their proximity to Euclidean distances, while categorical data is simpler, namely the mode value. The calculation of the distance between examples of minor classes whose variables are categorical scale is done using the Value Difference Metric (VDM) formula as

$$\Delta\,(X,Y) = w_x w_y \sum_{i=1}^{N} \delta(x_i, y_i)^r$$

Δ(X, Y) is the distance between X and Y, $wxwy$ are the weight (negligible), N is: the number of explanatory variables, R is 1 (Manhattan distance) or 2 (Euclidean distance) and $\delta(xi\ ,\ yi\ )$ r is the distance between categories, with

$\delta(V1,V2\ )$ =

$\delta(V1,V2\ )$ is the distance between the values of V1 and V2, while $C1i$ is the number of V1 yang which belong to i, $C2i$ is the number of V2 which belongs to class i, I is the number of classes, i=1,2, …. M, $C1$ is the number of values 1 occurs. $C2$ is the number of values 2 occurs, N is the number of categories and R is a constant (usually 1).

- Procedure of artificial data generation for numeric data

o Calculate the difference between main vectors and their closest neighbors.

o Multiply the difference by a random number between 0 and 1.

o Add this difference to the principal value of the original main vector so that a new principal vector is obtained.

• Procedure of artificial data generation for categorical data

o Select the majority between the principal vector under consideration and its nearest k-neighbor for nominal values. If there is a similar value, choose randomly.

o Make the value data as an example of a new artificial class.

In classification method Naïve Bayes, K-NN, J48, Random Forest, SVM and AdaBoost is used.

Naïve Bayes is a simple probabilistic classification algorithm that calculates probabilities based on the frequencies and combinations of values in a dataset. It assumes independence among features within a class and uses Bayes' theorem to calculate the posterior probability of a class given a predictor. The algorithm requires a small amount of data for classification and can yield unexpected results that may not match the actual reality. The formula

$$P(C|X) = \frac{P(X|C)P(C)}{P(X)}$$

represents the posterior probability of class (C, target) given a predictor (X, attribute). $P(C)$ represents the probability of the previous class, $P(X|C)$ is the probability of the predictor given the class, and $P(X)$ is the probability of the predictor.

K-NN: In short, the KNN is a classification algorithm based on the nearest neighbor to calculate the distance, the Euclidean Distance equation can be used. Euclidean Distance is a formula for finding the distance between 2 points in two-dimensional space, equation 4 shows the calculation of Euclidean Distance.

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

J48: Decision tree with the J48 algorithm is a classification method that uses a tree structure representation where each node represents an attribute, the branches represent the value of the attribute, and the leaves represent the class.

The process of building a decision tree starts with the root node, which represents the topmost node in the tree. The following steps are involved in constructing the decision tree:

Forming a decision system: A decision system is created, consisting of condition attributes (such as sales, purchases, warehouse stock, and operating expenses) and a decision attribute (e.g., profit). The decision system includes objects (E1, E2, E3, ..., En) and their corresponding attribute values.

Calculating column data: The amount of data based on attribute members with specific conditions is determined. Initially, the conditions are empty.

Selecting attributes as Node 4: One of the attributes is selected as Node 4, and a branch is created for each member of the Node.

Checking entropy values: The entropy value of each node member is examined. If the entropy value is zero, it indicates that the leaf nodes can be formed. If all the entropy values are zero, the process continues to the next step.

Recursively repeating the process: If any node member has an entropy value greater than zero, the process is repeated from the beginning with that node as a condition. This recursion continues until all members of the node have an entropy value of zero.

Attribute selection based on gain value: The attribute with the highest gain value among the existing attributes is chosen as the next node. The gain value of an attribute is calculated using Equation (5), which subtracts the weighted sum of entropies of the resulting partitions from the entropy of the initial collection.

Calculating entropy: The entropy value of a collection is computed using accuracy, which sums the negative of the proportion of each partition to the collection multiplied by the logarithm (base 2) of the proportion.

In summary, the decision tree construction involves selecting attributes, partitioning data, evaluating entropy values, and recursively repeating the process until all nodes have zero entropy values. The attribute with the highest gain value is chosen as the next node in the tree.

$$\text{Gain}(S, A) = \text{Entropy}(S) - \sum_{i=1}^{n} \left|\frac{si}{s}\right| \times \text{Entropy}(Si) \quad (4)$$

S : Case Collections
A : Attribute
N : The number of partitions attribute A
|Si| : The proportion of Si to S
|S| : Number of cases in S.

Meanwhile, to calculate the Entropy value with (6).

$$\text{Entropy}(S) = \sum_{i=1}^{n} -pi \times \log_2 pi \quad (5)$$

S : Case Collections
N : Number of Partitions S
Pi : The proportion of Si to S

Random Forest - Random Forest is an ensemble learning method that combines multiple decision trees to perform classification, regression, or other tasks. It is derived from the concept of decision trees. In Random Forest, each tree is built using a bootstrap sample from the training data. During the tree construction, a random subset of attributes is selected for determining the best splits. The final model of the Random Forest is a combination of the results from all the individual trees. Random Forest is advantageous as it reduces overfitting and improves the model's accuracy by leveraging the collective decision-making of multiple trees.

SVM (Support Vector Machine) is a machine learning technique that separates data points using a hyperplane in attribute space, aiming to maximize the margin between instances of different classes. It

can effectively handle high-dimensional data, but training times can be slow. SVM is accurate for complex nonlinear models but may be prone to overfitting.

The key advantage of SVM is its ability to handle nonlinear data by using kernel functions. These functions map the input patterns to higher-dimensional spaces where the data points become linearly separable. The mapping is defined as the inner product between data points in the high-dimensional space.

AdaBoost (Adaptive Boosting) is an ensemble learning algorithm and a variant of the boosting algorithm. It is widely used due to its solid theoretical foundation, accurate predictions, and simplicity. The AdaBoost algorithm involves the following steps:

Input: Collection of labeled training samples and a component learning algorithm with a specified number of iterations.

Initialization: Assign equal weights to all training samples.

Iterate for each iteration (t=1 to T):

Train a classification component (h_t) using the component learning algorithm on the weighted training samples.

Calculate the training error (ε_t) by comparing the predictions of h_t with the actual labels.

Assign a weight (a_t) to the classification component based on the training error.

Update the weights of the training samples based on the weights of the misclassified samples.

Output the final prediction function f(x) as a linear combination of the weighted classification components.

In the output, f(x) is the prediction function, and the weights (a_t) determine the importance of each classification component. The final prediction is obtained by summing the weighted predictions of all the components.

In paper [2] A database of 100 borrowers from a commercial bank branch that took secured consumer loans is used for the purposes of the study. The clients are defined as loyal based on their credit history (they have less than 3 missed payments for the last year).

The purpose of this analysis is to divide the 100 objects into 3 groups using the following 3 variables: - - Loan Amount (in euro);

- Time with Bank (in months);

- Worst Status Last 12 Months – shows the number of missed payments per customer in the last 12 months (0 - “0 missed payments,” 1 - “1 missed payment,” 2 - “2 missed payments”).

K-means clustering is a method used in the study, where the initial segmentation variables are taken as input data. It calculates the distance of each unit to individual cluster centers and assigns the unit to the nearest cluster based on the distance. The number of clusters is predetermined in this non-hierarchical clustering approach. The cluster centers can be known or evaluated from the data, and they may remain

constant or get updated during the analysis process. The analysis requires the use of quantitative variables, specifically the variable "Worst Status Last 12 Months" in this case.

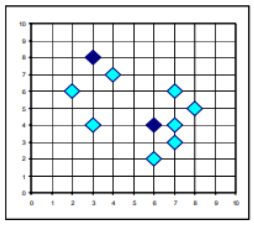
When using a statistical procedure that involves distance calculations, the measurement units of variables become important. Variables with larger values will have a greater impact on the distance compared to variables with smaller values. To address this, a further analysis is conducted using standardized segmentation variables, which have a mean of 0 and a standard deviation of 1. A comparison is made between the results obtained using the initial and standardized segmentation variables, as well as with a two-step clustering approach within each strategy.

The analysis is performed using SPSS, a statistical software package commonly used for data analysis.

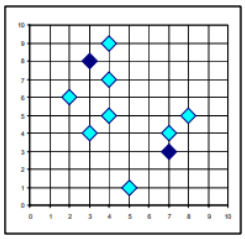
Whereas in paper **[3]** K-Medoids and K-Means clustering is used where following steps are implemented-

K-Medoids

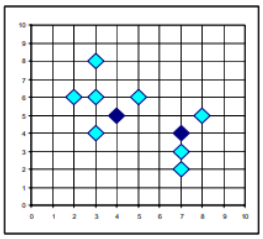
1. Choose the number of clusters, k.
2. Select k initial data points as cluster centers.
3. Calculate the swapping cost between each non-identified data point and each determined cluster center.
4. If the swapping cost is negative, replace the cluster center with the data point.
5. Assign each non-determined data point to the most similar cluster center.
6. Repeat steps 3-5 until there is no further change in the cluster assignments.
7. The algorithm has converged, and the resulting cluster centers and assignments represent the final clustering solution.



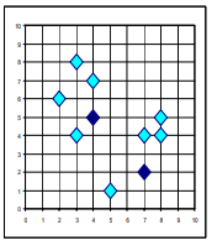
$C_{jih} = d(j, h) - d(j, i)$



$C_{iih} = 0$



$C_{iih} = d(j, t) - d(j, i)$



$C_{iih} = d(j, h) - d(j, t)$

K-Means

Given m, the m-Means algorithm is implemented in 4 steps:

i. Partition objects into m nonempty subsets

ii. Calculate pivotal points as the centroids of the clusters of the new orientation.

iii. Partitioned each object to the cluster with the nearest pivotal point.

iv. Go back to Step 2, stop when no more new assignment.

Figure 2: K-Means Clustering Algorithm.

Performance and Evaluation

In paper **[1]** the dataset is initially extracted, and it is noted that there are no missing values but some unknown information present. To handle the unknown values, classification methods are employed to predict them since they mostly occur in nominal attributes.

Next, the dataset undergoes SMOTE (Synthetic Minority Over-sampling Technique) preprocessing to address the imbalance in the target attribute. This technique generates synthetic samples to increase the representation of the minority class. The resulting dataset is balanced, with the number of instances for each class adjusted accordingly.

To evaluate the performance of the classification algorithm, a 10-fold cross-validation technique is applied. The dataset is divided into 10 subsets, and the classification models are trained and tested using each fold.

The evaluation process begins without SMOTE preprocessing, where various classification methods such as Naïve Bayes, KNN, Random Forest, J48, PART, SVM, and AdaBoost are employed. The results are compared based on accuracy, true positive rate (TPR), recall, precision, and F-measure. The Random Forest method achieves the highest accuracy among the tested methods.

Subsequently, the dataset is preprocessed using SMOTE, and the classification methods are reapplied to evaluate the impact of SMOTE preprocessing. Again, Random Forest attains the highest accuracy, suggesting that the combination of SMOTE and Random Forest improves the classification performance.

The study highlights that the use of SMOTE and tree-based classification methods can enhance the accuracy of the classification algorithm applied to the Bank Marketing dataset. However, SVM and Naïve Bayes methods exhibit a decrease in scoring value. The comparison of results demonstrates the effectiveness of the SMOTE + Random Forest approach, which achieves an accuracy of 92.61%.

The performance of the SMOTE + Random Forest method is further analyzed using a confusion matrix and ROC (Receiver Operating Characteristic) curve. These evaluations provide a deeper understanding of the algorithm's performance and its ability to correctly classify instances.

Overall, the study showcases the importance of preprocessing techniques like SMOTE and the selection of appropriate classification methods in improving the accuracy of classification algorithms for real-world datasets like the Bank Marketing dataset.

The paper **[2]** describes a K-means cluster analysis using initial variables and standardized variables. The analysis aims to segment clients based on their loan amount, time with the bank, and worst status in the last 12 months.

The results of the K-means cluster analysis using initial variables are presented in Table 1. Cluster 1 represents clients with the highest average loan amount, no missed payments, and a long average time with the bank. Cluster 2 includes clients with the second-highest loan amount and number of missed payments. Cluster 3 has clients with average time with the bank similar to Cluster 1 but lower loan amounts and higher missed payments.

TABLE 1. K-means cluster analysis – final cluster centers

Cluster No / Variable	**1**	**2**	**3**
Loan Amount	92667	50150	3165
Time with Bank	92	69	90
Worst Status Last 12 Months	0.00	0.25	0.44
Number of cases	**3**	**4**	**93**

Based on these results, the text suggests implementing different strategies for each cluster. Cluster 1 clients are labeled as "platinum clients," Cluster 2 as "gold clients," and Cluster 3 as the weakest group of loyal customers, referred to as "silver." By focusing on these segments, the bank can target specific types of clients they seek, such as those with no missed payments and large loan amounts.

Additionally, an analysis of variance (ANOVA) is conducted to assess the influence of each variable on cluster formation. The loan amount is found to have the greatest influence, followed by the worst status in the last 12 months, while time with the bank has the least influence.

The text acknowledges that K-means clustering is sensitive to outliers, which can form their own small groups. However, in this customer segmentation context, outliers may be important customers and shouldn't be excluded from the analysis.

The use of standardized variables in K-means clustering is also discussed. Table 3 presents the results obtained using standardized variables, showing the final cluster centers. The text highlights that the choice of clustering method depends on the creditor's goals. If the loan amount is crucial, K-means clustering with initial variables is appropriate, as it assigns the highest weight to this variable. If the impact of the loan amount is not desired, two-step clustering or standardized variables can be used.

TABLE 3. K-Means Cluster Analysis with standardized variables – final cluster centers

Cluster No / Variable	**1**	**2**	**3**
Loan Amount (Z)	-0.215	3.356	-0.280
Time with Bank (Z)	-0.955	-0.165	0.711
Worst Status Last 12 M (Z)	0.565	-0.367	-0.360
Number of cases	**39**	**7**	**54**

The comparative analysis further explores different strategies and compares the results of K-means clustering with initial variables, K-means clustering with standardized variables, and two-step clustering.

The paper provides tables for each strategy, indicating the most appropriate clustering method for each case.

- Strategy 1: The smallest number of missed payments in the last 12 months and a large loan amount;

- Strategy 2: The smallest number of missed payments in the last 12 months and long-standing bank customers;

- Strategy 3: The largest loan amount and long-standing bank customers;

- Strategy 4: The largest loan amount and a small number of missed payments in the last 12 months;

- Strategy 5: The longest-standing bank customers with a large loan amount;

- Strategy 6: The longest-standing bank customers with a small number of missed payments in the last 12

Overall, it provides an overview of the K-means cluster analysis process, the interpretation of results, and the implications for developing strategies based on customer segmentation.

In paper **[3]** statistical technique called Principal Component Analysis (PCA) is used to transform a set of correlated variables, represented by k original variables (x1, x2, ..., xk), into a set of uncorrelated variables called principal components. The goal of PCA is to reduce the dimensionality of the dataset while retaining as much information as possible.

The transformation involves creating k new variables (y1, y2, ..., yk), which are linear combinations of the original variables. Each new variable (yi) is calculated by multiplying the original variables (xi) by corresponding coefficients (aij) and summing them up. The coefficients (aij) represent the loadings of each original variable on the respective principal component.

The formulas for calculating the new variables are as follows:

y1 = a11x1 + a12x2 + ... + a1kxk

y2 = a21x1 + a22x2 + ... + a2kxk

...

yk = ak1x1 + ak2x2 + ... + akkxk

The coefficients (aij) are determined in a way that the new variables (yi) capture the maximum amount of variation in the original dataset. The first principal component (y1) accounts for the largest possible variance, the second principal component (y2) accounts for the second-largest variance orthogonal to the first, and so on.

Here they have segmented the three categories of the customers based on their attitudes, gestures, patients, movements, demands, body languages, communication skills and literacy. According to the algorithms of K-Means and K-Medoids Clustering, they have noticed that K-Medoids perform very well and finally the Principle Component Analysis (PCA) correlates all the classifications results. We define a threshold value of standard behavior after talking with Branch managers of five banks including two governmental Banks. Based on their opinion author then check this parameter with the K-Medoids

Algorithm threshold value. Then we finalized the value is θ=0.24. The flow chart of the process is given below.

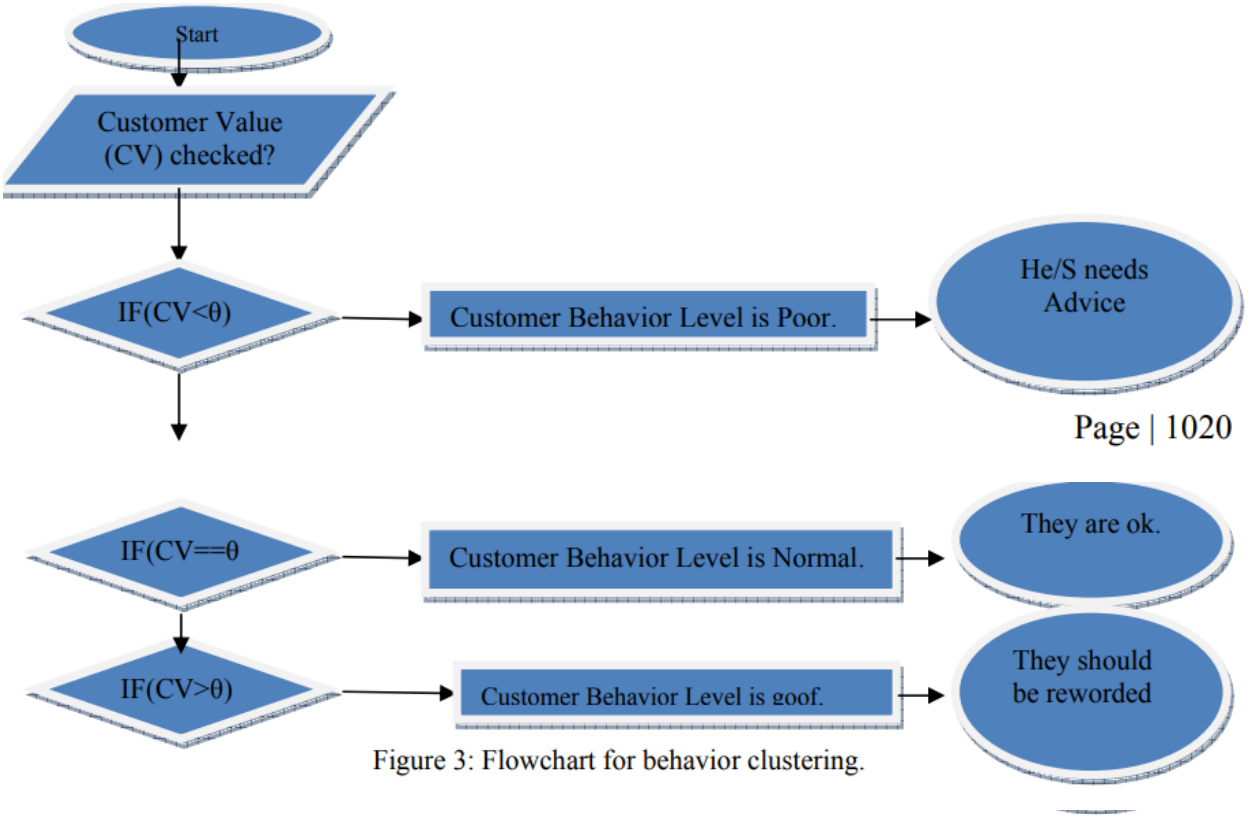


Figure 3: Flowchart for behavior clustering.

11. Result

Bank Name	Customer Name	Loan_ID	Loan Account	Location	Segmented
Brack Bank	Tapos Bormon	4001	124782	Patiya	Good
Brack Bank	Afjal Hossain	4012	783141	Do	Poor
Brack Bank	Abdul Aziz	4902	321574	Do	Ok
Brack Bank	Hira Das	4092	492345	Do	Ok
Brack Bank	Morjina Begum	4982	438921	Do	Poor
Dhaka Bank	Earsad Ullah	8723	234120	Do	Poor
Dhaka Bank	Emam Uddin	8701	234045	Do	Poor
Dhaka Bank	Abul Khair	8728	230912	Do	Poor
Dhaka Bank	Towhidul Islam	8723	234570	Do	Poor
Dhaka Bank	Karim Uddin	8703	249072	Do	Poor
UCBL Bank	Jahedul Islam	5423	764848	Do	Poor
UCBL Bank	Nishita Shaha	3452	658392	Do	Poor
UCBL Bank	Moktar Hossain	4563	652310	Do	Good
UCBL Bank	Abutahar Mia	4567	680213	Do	Ok
UCBL Bank	Khursad Fazil	7832	602343	Do	Poor
Janata Bank	Asis Mia	0923	984536	Do	Poor
Janata Bank	Farhan Ali	0876	908765	Do	Good
Sonali Bank	Tipu Sultan	3424	214567	Do	Good
Sonali Bank	Korim Monshi	3214	213468	Do	Ok
Agrani Bank	Absar Mia	0123	987566	Do	Ok
Agrani Bank	Josim Uddin	0342	982342	Do	Ok

Table 4: The resultant part of the experiment.

Conclusion

[1] The study focused on addressing the issue of imbalanced data in the Bank Marketing dataset using the SMOTE method combined with different classification algorithms. The results showed that applying

SMOTE with Random Forest led to fairly reliable outcomes and improved accuracy compared to classification without SMOTE.

The main finding was that the tree-based classification method (Random Forest) benefited the most from the SMOTE technique, resulting in a significant increase in accuracy. However, the K-NN, Naïve Bayes, and SVM methods experienced a decrease in accuracy when SMOTE was applied.

Furthermore, the study found that the computation time for SVM and Random Forest was the longest due to the increased number of instances generated by SMOTE. This suggests that the use of SMOTE can have an impact on computational efficiency.

In conclusion, the study demonstrates that the SMOTE method can effectively address the problem of imbalanced data and improve classification accuracy, particularly in tree-based algorithms like Random Forest. However, its impact on accuracy may vary depending on the choice of classification method, with SVM and Naïve Bayes showing a decrease in accuracy. Researchers should consider trade-offs between accuracy and computational time when using SMOTE and explore additional techniques to optimize both aspects.

TABLE III
TEST RESULTS WITHOUT SMOTE

	Accuracy	TPR	Recall	Precision	FMeasure
Naïve Bayes	88,00%	88,00%	88,00%	88,40%	88,20%
KNN	86,90%	87,00%	87,00%	86,00%	86,40%
R Forest	90,38%	90,40%	90,40%	89,30%	89,60%
J48	90,30%	90,30%	90,30%	89,50%	89,80%
SVM	89,20%	89,30%	89,30%	87,20%	86,60%
Adaboost	89,36%	89,40%	89,40%	87,30%	87,30%

TABLE IV
TEST RESULTS WITH SMOTE

	Accuracy	TPR	Recall	Precision	FMeasure
Naïve Bayes	82,17%	82,20%	82,20%	82,60%	82,30%
KNN	86,76%	86,80%	86,80%	86,80%	86,80%
RForest	92,61%	92,60%	92,60%	92,70%	92,60%
J48	90,52%	90,50%	90,50%	90,60%	80,30%
SVM	86,73%	86,70%	86,70%	86,70%	86,70%
Adaboost	92,35%	92,40%	92,40%	92,40%	92,40%

TABLE V
CONFUSION MATRIX FOR SMOTE + RANDOM FOREST

Yes	No	Total
37197	2725	39.922
2177	**24268**	26.445

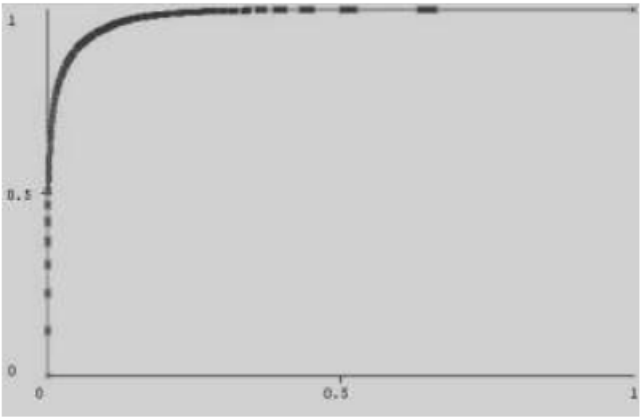


Fig. 3 ROC curve for SMOTE + random forest

[2] The analysis conducted a customer segmentation study using K-means cluster analysis, identifying three clusters of loyal customers based on different variables. By comparing results using initial and standardized segmentation variables, as well as a two-step clustering method, the study determined the most suitable cluster analysis for each segmentation strategy. The findings highlight the significance of customer segmentation in commercial banks, not only for statistical analysis but also for practical outcomes in selling financial services and improving competitiveness.

[3] The conclusions drawn from the provided analysis suggest that developing effective strategies for behavior segmentation in the banking industry requires a comprehensive understanding of customer value, behaviors, and needs. By leveraging Customer Relationship Management (CRM) systems and client databases, financial institutions can analyze data to perform profitable market segmentation. This involves assessing collected information to identify common preferences and behavioral patterns among different customer groups based on demographic, geographical, psychological, and other characteristics. The emergence of direct channels, online platforms, and social networks has facilitated easier management of multiple banking relationships and comparison of products and services for customers. The study highlights the potential of using Principal Components Analysis and Artificial Neural Network for banking customer behavior segmentation, offering a new dimension in the existing literature and aiming to accelerate the modern banking system.

In conclusion, customer segmentation for personalized banking services is a crucial task that can be effectively achieved through the application of data mining techniques. By leveraging data mining, financial institutions can analyze vast amounts of customer data to identify distinct customer segments based on various criteria such as demographics, behaviors, preferences, and needs. This enables banks to offer personalized services and tailored product recommendations to individual customers, enhancing their overall banking experience. Data mining techniques such as clustering, decision trees, and

association rule mining play a significant role in identifying patterns and relationships within the data, enabling banks to gain valuable insights into customer behavior.