IMPROVE BANK CUSTOMER PROFILING USING MODIFIED K-MEANS CLUSTERING

Mrs.P.Ramya[1], T.S.Nivetha[2], M.Ramya[3]

[1]Assistant Professor, [2,3]Final Year Students
Department of Information Technology,
K.L.N. College of Engineering

***Abstract*: In today's competitive markets for a business success it is essential to satisfy the desires and preferences of customers. As the data grows enormously in banking and financing sectors, data analytics plays an important role in prediction and statistics. Banking sector provides different kinds of services in which this project focuses on credit card distribution. There are various types of customers who has different types of interest such as shopping, dining, exploring new places and so on. The banks also provides a varied number of credit cards with more number of specific services. In this paper, the outlier detection will be done for the dataset by using a novel algorithm before doing clustering process. The clustering will be done using modified K-means. From the clustering result the customers will be classified into three groups and the credit card offers will be provided. The information is not only helpful for the bank to understand related characteristics of different customers, but also marketing representatives to find potential customers and to implement target marketing [3].**

I. INTRODUCTION

The face of banking industry has changed tremendously over the years, predominantly after 2008's economic crisis. Cost of compliance, regulatory reforms and risk management needs careful attention for simplifying business functions. Banking analytics is used for getting intelligent understanding on key business areas. But, no matter how much it is streamlined, the banking industry will still be bumping into an information overload. Though it continues to accumulate large chunks of data, it has been lagging to get accurate and high-value insights on the business. Banks must adopt analytics solutions that can help them in taking sound business decisions, become more compliant, drive innovation and uncover other hidden capabilities. The percentage of bank who have already invested in big data and analytics is 68%, however, 65% of them are yet to start availing their benefits. Though the concept of business intelligence (BI) in banking isn't new, the complete potential of its applications still needs to be explored. Banks in India have been using analytics as a lucrative technology to avail the advantages such as attracting new customers, retaining existing ones, maximizing gains and eliminating costly, time- consuming and redundant processes.

In this project, data analysis has been performed on the bank dataset which contains customer Id, Income, Expenditure, Loan category, Loan amount, Debt record, Returned cheque, Dishonor of bill, Overdue and so on. This dataset will be carried out by the process called Outlier Detection using DENSAT Algorithm which includes Statistical and Density based model algorithm. In the outlier detection, by using the statistical based algorithm the exceptional values will be replaced with the benchmark value and the NA values will be replaced with the values computed from the mean function. Based upon the calculation procedure of four conditions, by analyzing the history of debt record, returned cheque, dishonour of bill and overdue, the customers will be provided with the credit score. And finally the minimum density level will be removed as outliers. Then the process will be continued by moving in to the clustering, which is performed using Modified k-means calculated using Manhattan distance. Next, classification is performed using Support Vector Machine (SVM) by classifying it into three groups which includes, High Profitable Customers, Potential High Profitable Customers and Common Customers [3]. Again these three classified groups are processed in order to find the mostly used category of loan and the credit card offers will be provided according to the category.

II. LITERATURE SURVEY

The literature has reviewed several strategies and identified the pros and cons. From the Literature review it is identified that Banks are implementing "Customer Relationship Management" (CRM) and Statistical Analysis System (SAS) strategies that include segmenting and targeting customers for increased profitability. While continuing to use traditional methods to determine short-term product profitability, it is important to develop new models to determine a customer's lifetime value to the bank. The SAS System provides the technology for both types of analysis. A multidimensional database (using SAS/MDDB™ software) provides an efficient means of storing aggregate data that can then be surfaced through the multidimensional viewer of SAS/EIS® software. Enterprise Miner™ Software is an extensive environment for data mining, providing an array of modeling and statistical techniques for identifying patterns and trends in customer data. SAS/Warehouse Administrator™ software gives the Information Technology (IT) group the ability to manage the data extracted from the operational sources and organize it by subject area relating to the activities of the customers.

The existing credit card issuing systems use Statistical Analysis System (SAS) for their Business Intelligence. Over 30 years, SAS is more popular in handling large amount of data. It can provide dedicated customer support. Rather it has several limitations like, It is old-fashioned in its structure, It seem to be a bit hard, It doesn't include a 'pocket calculator', It doesn't have very fancy graphics possibilities , It is very expensive and we have to pay for any updated package to be used. In order to overcome the above limitations the proposed credit card issuing system uses R tool for implementation. As banking organizations began to recognize

the importance of credit risk control, credit risk scoring models have become more widely used in credit evaluation. The bank can use credit risk scoring models to reduce credit risks and to increase profits. Credit risk scoring is the method that banking organizations used when analyzing client data with credit scoring systems to decide whether applicants would be good or bad clients, with the former being able to pay off the debt and the latter having a higher possibility of breaching the contract[11].

III. DATA ANALYSIS USING R TOOL

There are several software packages on the market, which can do more or less the same, and by choosing SAS we don't want to indicate that the other possibilities are inferior. The proposed credit card issuing system uses R tool for implementation. This is better than SAS. R can do everything that SAS can do in terms of Statistical analysis and there are some pretty cool things R can do in which SAS can't. Anything you envisage using SAS STAT for statistical analysis and data mining, R can do it. R is free, it's an Open source project initially started in New Zealand and is now considered as one of the best Statistical analysis tools in the world. SAS Stat and other SAS packages pack a powerful punch and cover almost the whole gamut of statistical analysis and techniques. However since R is open source and people can submit their own packages/libraries, the latest cutting edge techniques are invariably released in R first. To date R has got almost 15,000 packages in the CRAN repository. Some of the latest techniques such as GLMET, RF, ADABoost are available for use in R but not in SAS. Many experimental packages are also available in R. Infact in most Kaggle competitions (which requires a blog post of it's own), the winners (who are amongst the world's best data miners) have almost invariably used R to build their models. In this aspect R is the hands down winner, however a word does need to be put in about SAS, since SAS is a paid software with support, any new innovation, or new statistical technique has to be vetted and accepted. SAS is used in many mission critical assignments where merely experimental techniques cannot be allowed to creep in. While this is necessary for the environment SAS works in, it also means that it will keep playing catch up with R in terms of latest innovations. On the other hand since anybody can upload a package in R. Therefore in terms of pure statistical capabilities, R is rated higher. R is a true programming language it gives more flexibility and power than SAS to the programmer. In addition R has advanced graphical capabilities, new statistical and machine learning techniques implementation in R will be much quicker than SAS and it has code efficiency. The contributions of the paper can be summarized as follows.

1. In this project a Novel algorithm named DENSAT for outlier detection is proposed.
2. Modified k-means for finding the initial centroids value is used. And then this initial centroids value will be given to k-means for further computations resulting with minimum number of iterations.
3. The classification will be done using "Support Vector Machine" algorithm. The maximum used category of customers will be noted and promoted as offerings for the Business Targeted Marketing.

IV. PROJECT WORK

This project proposes DENSAT algorithm for outlier detection in a first phase which is performed using statistical model and density based model. By using the algorithm several conditions had been applied to the dataset for computing the credit score. Then based upon the credit score, customers with low credit scores will be removed as outliers. The above result is given to the second phase called clustering which clusters the customers in to three groups using Modified k-means algorithm. The customers are classified in the third phase using Support Vector Machine (SVM) into three groups namely High Profitable Customers, Potential High Profitable Customers and Common Customers, again from the three classified groups customers are classified and they are provided with the credit card offers [3].

4.1 Outlier Detection:

Outlier is a data point that deviates too much from the rest of dataset. They are the abnormal data object having different behaviour than normal object in the data set. Most of real-world dataset have outlier. Outlier detection plays an important role in data mining field. Outlier detection uses many approaches like Statistical-based approach, Density-based approach, Deviation-based approach, Distance-based approach. These approaches have several demerits, in which Statistical based model has disadvantages like, it produces high positive rate and it is computationally expensive. Density based approach have been proved to be effective in detecting outliers successfully, but usually requires huge amount of computations. This paper proposes a novel algorithm named DENSAT based on Statistical model and Density based model algorithm.

4.1.1 Statistical Based Approach

By getting in to the process of outlier detection, at first the exceptional cases will be checked from the data set, and it will be replaced by some calculated benchmark value. This will be done to keep up the accuracy level, because due to those exceptional values we may result with some inaccurate results. To replace that exceptional values benchmark value will be calculated from the following formula (IRQ Rule),

$$Bench = Q3 + 1.5 * IQR\ value$$

Where,

Q 3 represents third quartile that can be calculated from the summary of the parameter.

IQR stands for Inter Quartile Range which is the difference between the upper (**Q3**) and lower (**Q1**) quartiles. It is often seen as a better measure as it is not affected by outliers.

After completing this step, next the dataset will be processed to find the NA values and it will be replaced by the values computed from the mean function. These functions are performed through the statistical model approach.

4.1.2 Density Based Approach

Based upon the calculation procedure of four conditions,

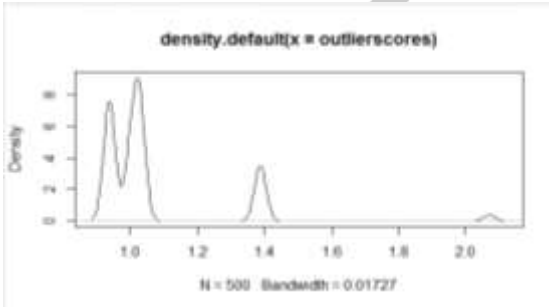
- Whether the customer has overdue or not
- Whether the customer has bad debt records or not
- Whether the customer has returned cheques or not
- Whether the customer has dishonor of bills or not

The score will provided as 1 if the above condition is True, otherwise False[3]. And finally the credit score will be calculated by adding up all these values resulted with,

Maximum credit score $\mathbf{C_{max} = 8}$
Minimum credit score $\mathbf{C_{min} = 4}$

In this approach, a local outlier factor (LOF) is computed for each point. The LOF of a point is based on the ratios of the local density of the area around the point and the local densities of its neighbors. The size of a neighborhood of a point is determined by the area containing a user-supplied minimum number of points. Based upon the LOF factor(density based outlier detection) outliers with minimum density will be removed by the following function and it is represented graphically,

```
outlierscores=lofactor(MyData1$creditscore,k=25)
density(outlierscores)
```



DENSAT Algorithm

Input: D // Dataset
Output: k identified outliers
/*Phase 1-DenSat-statistical based*/
Begin
for each record t in D
identify the exceptional cases
find the **benchmark value** and **winsorize** it to exceptional values
assign NA values using the **mean** function
for each record t in D
apply conditions using t
label result as YES or NO according to condition
/* Phase 2-DenSat – density based */
while not end of the dataset do
for each record t in D
if result is satisfied then
set value as 1
else
set value as 2
/* Phase 3-for evaluating credit score */
while not end of the dataset do
for each record t in D
evaluate credit score based on the result values
/* Phase 4-outlier identification */

while not end of the dataset do
for each record t in D
identify outlier using **LOF** factor using credit score parameter
remove the identified outliers from the dataset
End

By measuring the performance of DENSAT algorithm, it had resulted with correct classification rate as 1 and incorrect classification rate as 0 using confusion matrix. A confusion matrix contains information about actual and predicted values. Performance of such systems is commonly evaluated using the data in the matrix. The following table represents confusion matrix,

		Predicted	
Actual		Positive	Negative
	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

Where,
TP is the number of **correct** predictions that an instance is **negative**,
FN is the number of **incorrect** predictions that an instance is **positive**,
FP is the number of **incorrect** of predictions that an instance **negative**, and
TN is the number of **correct** predictions that an instance is **positive**.

The accuracy based on performance analysis is 1 where,

Accuracy = (TP+TN) / Total

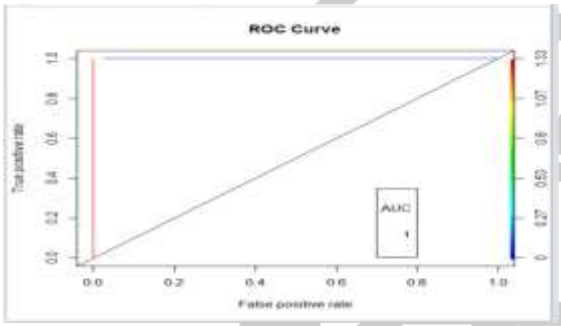
True Positive Rate has resulted 1 by using,

TPR=TP/(TP+FN)

False Positive Rate has resulted 0 by using,

FPR=FP/(FP+TN)

It is also represented in Receiver Operating Characteristics (**ROC**) graph which is a useful technique for organizing and visualizing the performance as below,



4.2 USER CLUSTERING

Clustering is a process of partitioning a set of data into a set of meaningful sub-classes, called clusters. Modified k-means is used for finding the initial centroid values [8]. And then, initial centroid values will be passed to k-means for further computations resulting with minimum number of iterations.

Limitations of k-means:

-Selection of the initial centroids is random.
-Depending upon the centroid values the number of iterations will be determined.

4.2.1 Modified k-means

Income, Expenditure and credit score will be given as parameters for finding the initial centroids. Initially random centers will be found by using the following function,

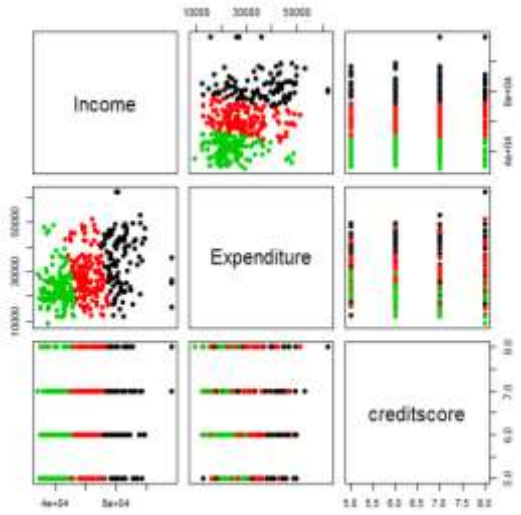
ktest[sample(nrow(ktest), 3),]

These random centers are given as input to the function which uses **Manhattan distance** to find the initial centroids. For high dimensional vectors you might find that Manhattan works better than the Euclidean distance. It is also calculated with more accuracy and speed.

Manhattan distance for Modified k-means. Formulated by,

For i=1 to n, $\sum$ ($|X_i-Y_i|$)
distance[,i]<-(abs(rowSums(t(t(points1)-points2[i,]))))

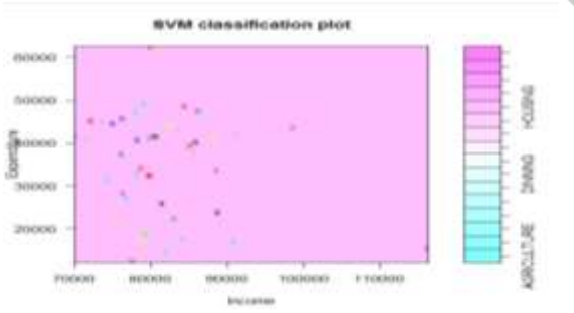
Then finally the inital centroid values returned is given as input to the k-means algorithm, the final clustering is represented below,



4.3 CLASSIFICATION

Classification is the process of organizing data into categories for its most effective and efficient use. Data classification is the process of categorizing data into various forms. In this project, the classification is done using "Support Vector Machine" algorithm [6]. Data classification includes separating customer data based on the Income, Expenditure and Credit score. It is done by using the svm function. After performing this function it results in the classification of customers with three groups namely, High Profitable Customers, Potential High Profitable Customers and Common Customers [3]. It is graphically represented below,

Further classification of each classified group is done to find the most commonly used category for loan. And the customers will be provided with the credit offers. The loan category is identified from the graph. The background colour in the graph represents the mostly used loan category and it can be identified from the list of shades in the plot. It is graphically represented as,



This is the classification plot of High Profitable Customers. From the above graph it can be identified that the most commonly used loan category is Housing.

5. CONCLUSION

In this paper, Profitable customers are examined into different categories by analyzing their credit details. Each customer is suggested with certain credit limit based upon the category they belong to. By providing such specific credit offers increases the chances of customer accepting the offer provided by the bank, which in turn helps the bank to increase its annual turnover and to retain its customers.

REFERENCES

1. Priyanka S. Patil, Nagaraj V. Dharwadkar, "Analysis of Banking Data Using Machine Learning". International conference on I-SMAC, 2020.

2. E. Kandogan. "Just-in-time annotation of clusters, outliers, and trends in point-based data visualizations". IEEE/Conference on Visual Analytics Science and Technology (VAST), 2021.

3. Wei Li, Xuemei Wu, Yayun Sun and Quanju Zhang. "Credit Card Customer Segmentation and Target Marketing Based on Data Mining". International Conference on Computational Intelligence and Security, 2018.

4. Guangli Nie, Yibing Chen, Lingling Zhang, Yuhong Guo. "Credit card customer analysis based on panel data clustering". International Conference on Computational Science, ICCS 2019.
5. Ayahiko Niimi, "Deep Learning for Credit Card Data Analysis". World Congress on Internet Security (WorldCIS-2018).

6. BenlanHea, Yong Shic, Qian Wan, Xi Zhao,
"Prediction of customer attrition of commercial banks based on SVM Model", 2020.

7. Shin-Chen Huang, Min-Yuh Day. "A Comparative Study of Data Mining Techniques for Credit Scoring in Banking". International conference on Information Reuse and Integration (IRI),2021.

8. Lavika Goel, Nilay Jai and Shivin srivastava, "A novel PSO based algorithm to find initial seeds for the k-means clustering algorithm",2022.

9. Patrick Rebentrost, Masoud Mohseni, and Seth Lloyd," Quantum Support Vector Machine for Big Data Classification", 2019.

10. Nishantha Bandara, H.M.N Dilum Bandara, "C Factors For Market Sales Promotion On Social Media In Banking Sector", 2021.

11. M.A Stelzner, "How marketers are using social media to grow their businesses", *Social media examiner*, 2020

12. J. Beese, "Complete guide to social media for banks & financial institutions", *Sprout Social*, 2022.

13. Utkarsh Srivastava, Santosh Gopalkrishnan, "Impact of Big Data Analytics on Banking Sector: Learning for Indian Banks", 2021.

14. Amir E. Khandani, Adlar J. Kim, Andrew W. Lo, "Consumer credit-risk models via machine-learning algorithms",2019.

15. Francisca Nonyelum Ogwueleka, "Neural Network and Classification Approach in Identifying Customer Behaviour in the Banking Sector: A Case Study of an International Bank", Department of Computer Science Federal University of Technology Minna, 2018.