MACHINE LEARNING ASSISTED PROFILING OF RETAIL BANKING CUSTOMERS

Pranjali Joshi, Anuj Mutha, Nidhi Patil, Chaitralee Datar, Sarang Agrawal
(Department of Computer Engineering, SCTR's Pune Institute of Computer Technology)

1. Abstract:
Retail banking, usually referred to as consumer banking or personal banking, is a type of banking that caters to private customers rather than enterprises. Individual customers can manage their finances through retail banking, get credit, and make secure deposits. This project aims at the profiling, i.e categorization of retail banking customers using a combination of deterministic rules and unsupervised machine learning. The transactions made by customers would be carefully studied and analyzed using clustering techniques. Patterns would be found in the data using machine learning, and segregation would be done accordingly. Real time categorical data would be used and worked upon, using the open source ETL Tool, 'Talend'. PostgreSQL, a highly scalable and open-source relational database management systsem would be used. Confidentiality would be maintained, as sensitive data of customers would be masked using a new method similar to hashing, without compromising uniqueness

Keywords - Hashing, Customer Profiling, Machine Learning, Unsupervised Learning, Data Engineering, Clustering

2. Introduction:
A subset of a company's current customers who share a number of characteristics and behaviors make up a customer profile. Businesses use this information to create in-depth, semi-fictional descriptions that aid in decision-making. These descriptions represent the client profiles. When creating customer profiles, it is vital to research consumer psychology Identify applicable funding agencies here. If none, delete this. and behavior. Customer profiling using machine learning is a popular area of study. Customers are the backbone of every business and keeping them happy is one of their most crucial responsibilities. They might ultimately fail if they don't accomplish this. The company must determine the right target market, and in order to do so, it must comprehend changes in customer behavior and preferences. Utilizing customer profiling, businesses can better understand the needs o

3. Motivation
A customer profile is a group of traits and behaviors shared by a subset of a business's current clientele. Businesses use this data to develop thorough, semi-fictional descriptions that assist in making business decisions. The customer profiles are these descriptions. Studying consumer psychology and behavior is necessary when developing customer profiles. Machine learning-based customer profiling is a popular research area. All businesses depend on their customers and keeping them is one of the most important duties they must complete. Failing to do so could ultimately result in their failure. Business must identify the correct target, and in order to do so, it must understand consumer behavior trends and preferences. Customer profiling aids businesses in comprehending the needs of their target market and the variables influencing consumer choices .There is limited research and a very few resources available with respect to customer profiling using unsupervised machine learning algorithms. Further- more, the accuracy of algorithms currently in use were not up to the mark or not very high. This gives us the motivation to bridge the gap between the machine learning based profiling currently used for customer categorization, and a more accurate and efficient model that we wish to build for the same.

ISSN: 2278-4632
Vol-13, Issue-05, No.03, May : 2023

4. Literature survey

In terms of work and resources, the study paper [1] discusses the classification of customers for the load profiles using iterative clustering. These load profiles can be used in future to predict the loads in the distribution network. The iterative self-organizing data analysis technique (ISODATA) algorithm is employed in this paper. The approach enables automatic adjustment of the number of clusters as necessary. Comparisons to other categorization techniques were done in order to confirm the precision of the ISODATA clustering algorithm.

For checking the accuracy alternatives of categorization like allocation to the closest existing customer class profile and classification in accordance with CIS customer class information were chosen. Additionally, the precision of each load profile is confirmed.

The study paper [2] delved deeper into evaluation of customers in retail banking collections. This paper outlines a study done using statistical methods for determining the kinds of likely behavior of customers having classes like whether or not customers will settle immediately; whether or not customers will pay any money at all. [2] The paper compared nine supervised classification methods like Logistic Regression, LDA, RDA, MDA, Random Forests, Support Vector Machines, Classification trees, Feed forward neural network, KNN method. It has been demonstrated that some widely accepted methods for assessing techniques for predicting the most likely future customer behavior contain significant flaws.

A two-layer clustering model has been presented in this paper [3] based on the examination of customer traits, customer contributions, and cluster segmentation. Two-layer clustering is applied to mobile telecom users and used to target the market by creating an effective marketing plan. Model divides customers into relevant clusters, allowing businesses to concentrate on their target markets before creating CRM, marketing plans, and promotional initiatives. In layer one of the model, the development of customers with high value of service usage will make high revenue to the company and in the second layer marketing strategy has been created. Hierarchical, Partitional, Density-oriented, Grid oriented, Model based clustering models are used in layer one.

The work presented in paper [4] specifies customer values which are ascertained using the LRFMP model, and then the k-means clustering technique is used to profile the consumer. The relationship between various customer profiles and different content kinds is then elicited using customer VoD rental preferences that are recovered using association rule mining. The applicability of the proposed approach is demonstrated on real-world data obtained from an Internet protocol television (IPTV) operator. In this way customers are categorized.

In the data mining literature theory of paper [5], there are two basic methods for customer segmentation: distance-based clustering techniques like k-means and parametric mixture models like Gaussian mixture models. Customer's web transactions are examined, and customers are then profiled or segmented as necessary to better understand them. As a result, many clusters that support categories like entertainment, portals, and other really unique patterns can arise. It seems sense that there are inherent categories in consumer behavior, and these categories affect the observed transactions and data. We propose that pattern-based clustering algorithms, like the one discussed in this research, may be efficient in learning such natural categories, allowing businesses to better understand their consumers and create more precise customer models.

There are several stages involved in creating customer classes. Customers are grouped using clustering techniques based on predefined features (for example, time domain data or a limited set of features obtained through data compression techniques), starting with load pattern data collection from dedicated measurements on a sample of customers (including bad data detection and elimination) [6]. The original Electrical Pattern Ant Colony Clustering (EPACC) algorithm is illustrated, highlighting its characteristics and parameters, with centroids evolution during the iterative process until stabilization.

The Length, Recency, Frequency, Monetary (LRFM) model, a variant of one of the most used segmentation strategies, is used in the study [7] to segment B2B customers. Customers are clustered into

ISSN: 2278-4632
Vol-13, Issue-05, No.03, May : 2023

categories using the k-means++ clustering technique in addition to the LRFM analysis. With the help of the clustering technique k-means++, customers are divided into 8 categories. The k-means++ algorithm is performed 100 times in a row with several cluster centers to guarantee that the best clustering solution is selected.

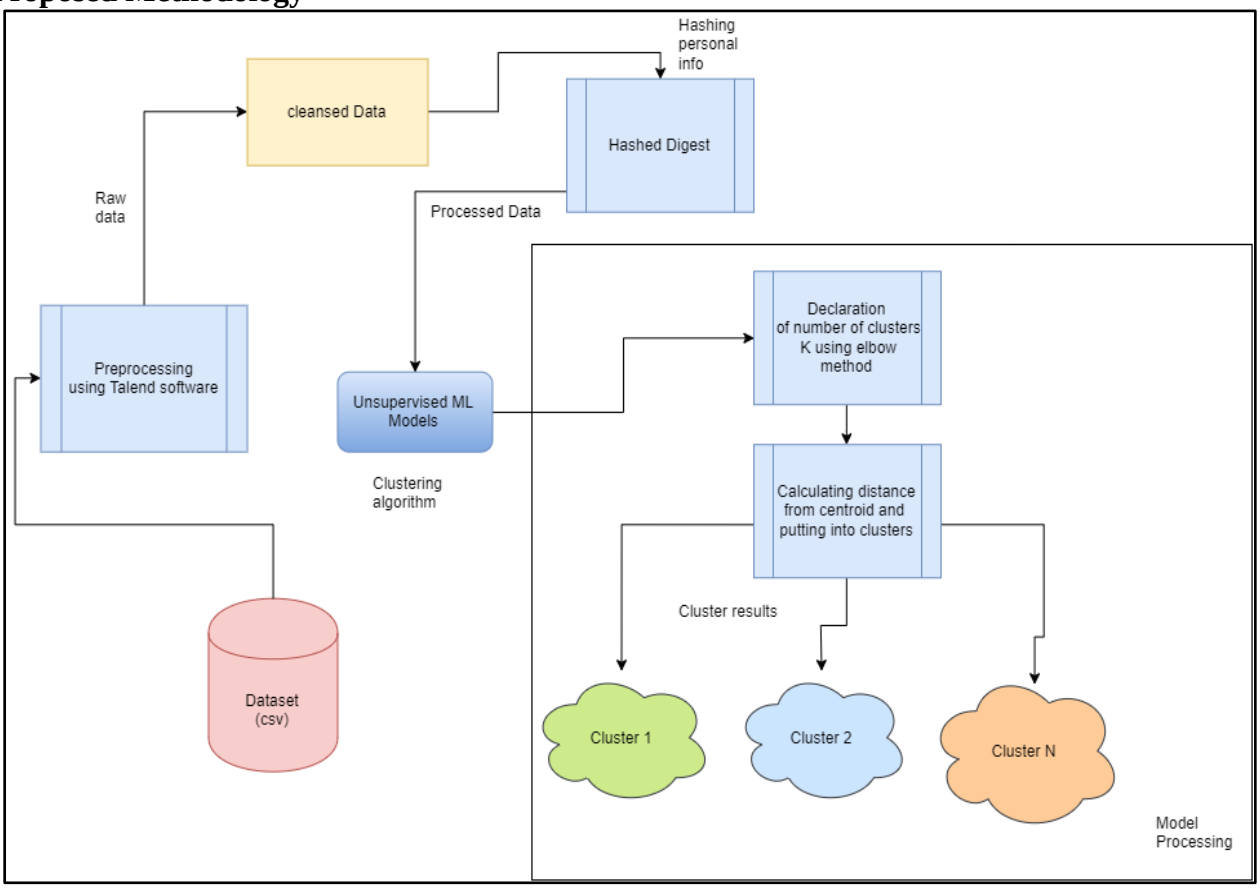
This paper [8] combines the radiation values of community relationships with the RFM model and improves the algorithm of M index to form the RFMC model, making it more suitable for e-commerce enterprises with the nature of community promotion. This was done by observing the product and sales characteristics of educational e-commerce enterprises. The segmentation accuracy of enterprise customers is directly impacted by how the segmentation model is built.

In This survey paper [9], various papers are studied for customer segmentation using Random Forest algorithm and Gradient boosting model. Clusters are visualized using various visualization tools. Hence, Using the approach, a latent model is created that can manage the numerous dependent factors and be utilized to accurately forecast the class.

For the best resource management in a smart water network, consumer habits, demands, and behaviors were considered for the research [10]. Consumers can also monitor water resource conservation and make decisions in this regard. The application of a clustering method to monitor the water supply to specific consumers and identify their water usage patterns using smart water meters is demonstrated in this study. Therefore, customer profiling will be used to ensure smooth water delivery.

The paper [11] explains the customer segmentation by analyzing the customer buying pattern at the marketplace. Every retailer wants to focus on their customers by personalizing their interests and hence customer buying patterns are analyzed and they are recommended with an appropriate product which the customer will definitely buy.

5. Proposed Methodology



ISSN: 2278-4632
Vol-13, Issue-05, No.03, May : 2023

5.1 Dataset:

The bank customer dataset is made up of a number of tables that include vital details about clients and their dealings with the bank. Data on different topics, including transactions, account specifics, and Know Your Customer (KYC) information, are contained in these connected tables. The transaction data contains information on each transaction, including the date and time of each one, the amount that was transacted, and the kind of transaction. This information can be utilized to understand consumer behavior and spot trends or patterns that will allow the bank to improve its processes and offerings. Information from consumer identity documents, contact information, and other personal details are all included in the KYC data. This information is crucial for maintaining the security and privacy of customer information as well as ensuring compliance with legal and regulatory standards.

5.2 Pre-processing with Talend Software:

Before the data is delivered to the machine learning model, it is necessary to pre-process it using the system's initial module. Pre-processing requires a number of steps, including data cleansing, missing value removal, handling of outliers, and formatting the data consistently. Data pre-processing operations including data cleansing, data enrichment, data deduplication, and data normalization can all be carried out using the software package Talend, which is used for data integration.

5.3 Cleansed Data:

A dataset free of any errors or inconsistencies is necessary to successfully train a machine learning model. This procedure, known as "data cleaning," entails finding and fixing any flaws or inconsistencies in the dataset. The Talend software, which offers a number of functions for data integration and administration, is one regularly used solution for data cleansing. The Talend software can be used to clean the data before feeding it into the machine learning model. Clean data is essential for the success of the machine learning model because errors or inconsistencies might impair the model's accuracy and performance.

5.4 Hashed Digest:

The second module in the system is in charge of utilizing a hashing method to conceal the personal information of retail banking customers. A hash function is a mathematical operation that takes an input value and produces an output value of a fixed size that is a distinct representation of the input value. The fields in this system are hashed using the SHA256 hashing method to individually identify each customer. This helps to safeguard consumer privacy while still enabling the system to use their data.

5.5 Unsupervised Machine Learning model:

The system's third module, which is the most crucial one, analyzes the pre-processed data and creates clusters based on inferences made from the patterns seen in the dataset. In this module, techniques for unsupervised machine learning are utilized to find patterns and similarities in the data without explicitly training on labeled data. Based on their transactional activity, the client segments can be identified using the clustering technique utilized in this module, which can then be used to tailor the banking experience for each segment.

5.6 Model Processing

The fourth module of the system, which is a subset of the third module, is in charge of processing the machine learning model by locating the centroid values and adding new data instances to the appropriate clusters. The accuracy of the clustering results depends on the centroid values, which are the average values of all the points in a specific cluster. In addition to making sure that new data instances are inserted into the appropriate clusters, This module is responsible for updating the centroid values and ensuring that new data instances are placed in the correct clusters.

5.7 Clusters:

In the third module of the system, the unsupervised machine learning model, groupings or segments of clients are referred to as clusters. Without the need for labelled data, this module uses clustering

algorithms to identify patterns and affinities in the pre-processed data. The system may identify the customer segments based on their transactional activity and behavior once the process of clustering is finished. The banking experience can then be customized for each of these customer categories or clusters. The bank might provide consumers in one cluster with unique investment options if they frequently conduct large transactions. In contrast, if a different group of clients predominantly use their accounts for savings, the bank might provide them with high-yield savings accounts or other savings-oriented products.

6. Limitations

1. Bias: Machine learning algorithms can be biased if they are trained on historical data that reflects the biases of the past. This can result in unfair and inaccurate profiling of customers.
2. Privacy concerns: Profiling customers using their data raises privacy concerns. It is important to ensure that the data is collected and used in a transparent and ethical manner.
3. Data quality: Machine learning algorithms require high-quality data to produce accurate results. If the data used for profiling is incomplete, inaccurate, or outdated, it can lead to flawed conclusions.
4. Over Reliance on data: Machine learning algorithms can only provide insights based on the data they are trained on. They cannot consider factors that are not captured by the data, such as the emotional state of a customer.
5. Lack of interpretability: Machine learning algorithms can be difficult to interpret, especially when they are based on complex models such as neural networks. This can make it difficult to understand how the algorithm arrived at a particular conclusion, and to identify any errors or biases in the process.
6. Cost and complexity: Implementing machine learning algorithms can be expensive and complex, requiring significant investment in technology and expertise.
7. Customer dissatisfaction: Customers may feel uncomfortable or suspicious of being profiled by algorithms, which can damage their trust in the bank and lead to customer churn.

7. Conclusion

The project is based on categorizing retail banking customers using a combination of deterministic and unsupervised machine learning. Through the project, we aim at building an unsupervised machine learning model that will be able to classify the retail customer of banks based on transaction details and usage patterns that will help retail banks to keep track of different types of customer in the form of clusters. The model would leverage the methodologies as compared to the current ones in use, thus achieving state-of-the-art accuracies in the same way.

8. References

[1] A. Mutanen, M. Ruska, S. Repo and P. Jarventausta, "Customer Classification and Load Profiling Method for Distribution Systems," in IEEE Transactions on Power Delivery, vol. 26, no. 3, pp. 1755-1763, July 2011, doi: 10.1109/TPWRD.2011.2142198.

[2] S. Guney, S. Peker and C. Turhan, "A Combined Approach for Customer Profiling in Video on Demand Services Using Clustering and Association Rule Mining," in IEEE Access, vol. 8, pp. 84326-84335, 2020, doi: 10.1109/ACCESS.2020.2992064

[3] Y. Yang and Balaji Padmanabhan, "GHIC: a hierarchical pattern-based clustering algorithm for grouping Web transactions," in IEEE Transactions on Knowledge and Data Engineering, vol. 17, no. 9, pp. 1300-1304, Sept. 2005, doi: 10.1109/TKDE.2005.145.

[04] G. Chicco, O. -M. Ionel and R. Porumb, "Electrical Load Pattern Grouping Based on Centroid Model With Ant Colony Clustering," in IEEE Transactions on Power Systems, vol. 28, no. 2, pp. 1706-1715, May 2013, doi: 10.1109/TPWRS.2012.2220159.
[05] D. A. Kandeil, A. A. Saad and S. M. Youssef, "A Two-Phase Clustering Analysis for B2B Customer Segmentation," 2014 International Conference on Intelligent Networking and Collaborative Systems, 2014, pp. 221-228, doi: 10.1109/INCoS.2014.49.
[06] Y. Huang, M. Zhang and Y. He, "Research on improved RFM customer segmentation model based on K-Means algorithm," 2020 5th International Conference on Computational Intelligence and Applications (ICCIA), 2020, pp. 24-27, doi: 10.1109/ICCIA49625.2020.00012.
[07] S. Chellaboina, M. Gembali and S. P. S, "Product Recommendation based on Customer Segmentation Engine," 2022 2nd International Conference on Intelligent Technologies (CONIT), 2022, pp. 1-7, doi: 10.1109/CONIT55038.2022.9847990.

[08] I. Figalist, M. Dieffenbacher, I. Eigner, J. Bosch, H. H. Olsson and C. Elsner, "Mining Customer Satisfaction on B2B Online Platforms using Service Quality and Web Usage Metrics," 2020 27th Asia-Pacific Software Engineering Conference (APSEC), 2020, pp. 435-444.
[09] M. F. Alam, R. Singh and S. Katiyar, "Customer Segmentation Using K-Means Clustering in Unsupervised Machine Learning," 2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N), 2021, pp. 94-98, doi: 10.1109/ICAC3N53548.2021.9725644.
[10] D. Arsene, A. Predescu, C. -O. Truică, E. -S. Apostol, M. Mocanu and C. Chiru, "Consumer profiling using clustering methods for georeferenced decision support in a water distribution system," 2022 14th International Conference on Electronics, Computers and Artificial Intelligence (ECAI), 2022, pp. 1-6, doi: 10.1109/ECAI54874.2022.9847435.
[11] S. Kaur and Sarabjeet, "Customer Se gmentation Using Clustering Algorithm," 2021 International Conference on Technological Advancements and Innovations (ICTAI), 2021, pp. 224-227, doi: 10.1109/ICTAI53825.2021.9673169.

