2022 5th Asia Conference on Machine Learning and Computing (ACMLC) | 979-8-3503-3392-3/22/$31.00 ©2022 IEEE | DOI: 10.1109/ACMLC58173.2022.00017

Customer Segmentation for improving Marketing Campaigns in the Banking Industry

Celine Ganar and Patrick Hosein
Department of Computing and Information Technology
The University of the West Indies St. Augustine, Trinidad
celine.ganar@my.uwi.edu, patrick.hosein@sta.uwi.edu

***Abstract*—The internet has had a significant impact on financial institutions by allowing customers to access many bank services virtually thus creating a very competitive environment. Therefore, efficient customer segmentation is a key objective for achieving more profitable market penetration. We propose a hybrid model that predicts a financial institution client's propensity to transition to an online banking platform. In this research, we utilized a hybrid approach where the first stage is Transaction Cluster Analysis where Recency, Frequency and Monetary (RFM) segmentation and K-Means cluster analysis were performed to detect the most loyal market segments. Analytic Hierarchy Process (AHP) was used to deduce the weightings of each cluster which aided in calculating the Customer Lifetime Value (CLV) of each cluster. Then two clustering algorithms, K-Modes and K-Means, were utilized on the clients' demographic features. In the final stage, we performed experiments that compared two supervised learning algorithms, Decision Tree and Extreme Gradient Boosted (XGBoost), to predict online transition behaviour. Our results showed that K-Modes clustering algorithm and XGBoost classification model yielded the best test accuracy of 96.1%. Our results illustrate our claims by showing that the bank can attract more customers, maintain its customer base, and keep their important customers satisfied.**

***Index Terms*—Customer segmentation, financial institution, online banking, clustering, customer lifetime value, decision tree.**

I. Introduction

The coronavirus pandemic accelerated the interactions within financial institutions which placed an emphasis on digital banking in the upcoming generation. The world is rapidly changing, thus it is expected that financial institutions will also make changes to their business models and will make bigger investments in digital interactions and the cybersecurity that supports it. Financial institutions are able to deliver their services via the online platform, subsequently diminishing the importance of conventional branch networks.

With the growth and availability of customer data, Big Data and Machine Learning provide an advancement in personalized banking and gathering customer insights. By gaining a comprehensive understanding of customers and then segmenting them into categories, financial institutions are able to better optimize their marketing strategies, which results in customer retention and increased profits [1]. Therefore, for a firm in a competitive environment, achieving efficient customer segmentation is a key objective for applying high quality recommendation strategies.

Customer segmentation can effectively decrease the marketing costs of a financial institution and assist in achieving more visible and profitable market penetration [2]. This technique allows firms to design and establish different strategies and promotions to maximize the value of customers. Using a combination of both demographic and behavioral data, maximum value is obtained as financial institutions are able to better understand the demands of their customer segments, which allows them to craft and offer the right service or product. The growth in data and technologies have changed the way in which large quantities of data are accessed, such that the work of executives and other decision makers is reduced. Important decisions can be made by the executives with the assistance of big data available to them using multiple algorithms.

In this paper, the customer segmentation process is implemented to segment the customers based on their demographic features and transaction behaviour, as well as predict which customers will transition to the online banking platform. In order to reach this study's goal, customer and transaction data were acquired from an unnamed financial institution. This study proposed a hybrid approach which segments and predicts banking customers who utilize the online platform. This study's hybrid approach consists of three stages: the first stage is segmentation by transaction analysis method called RFM (Recency, Frequency, Monetary), the second stage segmentation is based on the customers' demographics and the third stage uses the best performing decision tree model and clustering combination. A key contribution of this study's model is a two dimensional (demographic and transactional behaviour) focused approach to identifying online banking customers.

II. Literature Review

A. Customer Segmentation

Segmentation is an approach aimed at partitioning customers or other entities into groups based on similar characteristics such as demographics or behavior. This technique allows analysts to identify sections of customers who may behave in a similar manner to specific marketing techniques. Reference [3] stated that the K-Means algorithm is mainly used to effectively identify valuable customers and in the development of marketing strategies. The K-Means algorithm was utilized to segment credit card users based on their purchase behaviour [4]. Their aim was to better understand customers within the

979-8-3503-3392-3/22/$31.00 ©2022 IEEE
DOI 10.1109/ACMLC58173.2022.00017

African continent in order to extract their behaviour as well as to suggest tailored marketing strategies specific to the continent. Reference [5] performed segmentation on an electricity company's clients using two techniques, firstly they used Analytical Hierarchy Process (AHP) for indicator optimization then K-Means for clustering. Reference [6] utilized the Recency, Frequency and Monetary (RFM) model and the K-Means clustering algorithm to conduct customer segmentation and value analysis on an online sales dataset. Further analysis can be done on the RFM model [5] where AHP was used to determine the weight of the RFM variables, which were used to evaluate customer lifetime value (CLV). Clustering was then performed, according to the weighted RFM value, in order to group customers into clusters with similar RFM values. Then an association rule mining approach was used to extract recommendation rules for product recommendation to each customer group.

B. Clustering and Classification Hybrid Approach

This section describes a hybrid methodology to determine customer service priorities to retain the business' customer loyalty. The author utilizes clustering in the first stage followed by a classification algorithm, Decision Tree [8]. The clustering algorithm, K-Means, was examined to segment customers into their level of loyalty based on their transaction history. The elbow graph analysis found the optimal number of clusters to be four. They used a C4.5 Decision Tree to classify loyalty where the customer base was divided into four groups. As a benchmark model, they performed classification using C4.5 without the clustering technique result. The confusion matrix was used to assess both models where the hybrid model yielded a 97.7% accuracy while the benchmark model yielded a 94.45%. The results indicated that the patterns discovered in the clustering stage improved the decision tree detection of customer loyalty.

III. Dataset Description

Two anonymized datasets were obtained from an unidentified financial institution in the English speaking Caribbean. The first dataset consisted of 26 features describing the customer profile and account information of the bank's clientele. The second dataset described 750,000 transaction history records from over a six month period. The customer profile dataset consisted of two unique identifiers, account information, demographic data and geographic data. The transaction records consisted of account information and transaction information such as date, transaction channel and transaction amount.

Both datasets were merged to form a normalized dataset which has customer data as well as an aggregation of their transaction data. As a result, this dataset consisted of the 28,915 customers who transacted in the six month time-frame. Four additional columns were added to represent the number, dollar value, mean dollar value of all transactions by the customer, as well as, a column to represent the last day a customer carried out a transaction. This dataset introduced a binary target variable called OnlineUser where indication of an online transaction was represented by a class label of 1, while a class label of 0 indicated if an online transaction was not done.

IV. Methodology

We propose a hybrid approach for online banking customer prediction which combines cluster analysis and classification. This study's approach makes the assumption that once a customer makes a transaction on the online platform, they are considered an online customer. This is represented by the binary variable, OnlineUser, where 1 represents an online transaction made and 0 represents an alternative channel was used.

In the first stage, Transaction Cluster Analysis, Recency, Frequency and Monetary (RFM) segmentation was performed and K-Means cluster analysis to detect the most loyal market segments. The weightings of these clusters were then deduced using Analytic Hierarchy Process (AHP), and these weightings were then used to calculate the Customer Lifetime Value (CLV) of each cluster. Each cluster was then ranked according to their CLV. Then two clustering algorithms, K-Modes and K-Means, were utilized on the clients' demographic features. Both clustering algorithms used a combination of attributes. In the final stage, two Decision Tree models were utilized to predict online transition behaviour.

A. Transaction Cluster Analysis

In this stage, two clustering approaches were applied to identify the homogeneous groups of clients based on their transaction behaviour from the merged Banking dataset. A marketing technique named Recency, Frequency and Monetary (RFM) segmentation, and a non-hierarchical partitioning algorithm called K-Means, were applied since most literature about marketing showed both techniques were used together. This study's model combined both techniques for weighted cluster analysis to derive Customer Lifetime Value rank.

1) Recency, Frequency and Monetary Model (RFM): The RFM model is a popular and powerful marketing technique that is widely used to rate clients based on their previous transaction history. This technique is an analytical model that differentiates important clients from large datasets based on three dimensions, Recency (R), Frequency (F) and Monetary (M). RFM analysis can be utilized in a broad spectrum of applications involving customers in banking, retail and e-Commerce [9]. It is a popular technique of customer value analysis and has been greatly used for Customer Lifetime Value (CLV) [10]. The definitions of the RFM model are described below:

Recency: When was the last time the client transacted?
Frequency: How many times did the client transact?
Monetary: How much money did the client transact?

2) K-Means Clustering: Clustering is defined as an unsupervised learning algorithm which divides the entire dataset into segments or clusters which supports the hidden patterns within the data. The K-Means algorithm uses the Euclidean

TABLE I
RFM PAIRWISE COMPARISON MATRIX

	Recency	Frequency	Monetary
Recency	1	0.543	0.170
Frequency	1.84	1	0.269
Monetary	5.89	3.720	1

distance metric to partition the clients based on their Recency Scores, Frequency Scores and Monetary Scores. The K-Means algorithm starts with a primary group of randomly selected centroids, then iterative calculations are performed to optimize the positions of the centroids. It is inherent to decide the number of clusters (value of K) to form. In this research, the elbow technique and the best silhouette score were used.

3) Analytic Hierarchy Process (AHP): AHP is considered a systematic and hierarchical technique which assists executives and other decision makers to solve complicated multi-criteria decision making problems. The AHP model has been used in CLV calculation where it is used to evaluate the relative importance of the RFM variables [11]. For the purpose of this study, the AHP assessment was utilized based on expert evaluators within the banking sector as seen in Table I [11]. Based on their evaluation, they hypothesised that Monetary is the most important dimension as the experts mainly concentrated on the revenue generated by the customer. Saati's Eigenvector method is used to calculate the weight of each dimension.

4) Customer Lifetime Value (CLV): CLV is defined as the summation of the revenues gained from the firm's clients over their lifetime of transactions, taking into account the time value of money. The CLV calculated for the institution's customers enable decision makers to improve the customer segmentation and marketing resource allocation production. Identification of the cluster with customers who have the highest loyalty level can be be derived by the calculating the CLV using Equation 1 (see [12]) for each cluster derived from the RFM segmentation and ranking the clusters according to the highest CLV. Each rank is considered the CLV rank label (CLV_L).

$$CLV_i = (W_R * a(R_i)) + (W_F * a(F_i)) + (W_M * a(M_i)) \quad (1)$$

where i represents the i^{th} cluster, $a(X)$ represents the mean value of parameter X and W_X represents the weight of parameter X derived from the AHP technique.

B. Demographic Cluster Analysis

Clustering is a popular technique used for customer segmentation. In this paper, two modelling approaches are used to cluster customers based on their demographic features. In the first approach, K-Means algorithm is used to cluster demographic features with an additional feature, OnlineUser, a binary variable which describes whether a customer transacted on the online platform (1) or not (0) (DemoOnline). For the second approach, the K-Modes algorithm replaces the K-Means algorithm with the set of attributes as described above.

While the K-Means algorithm is mainly used to effectively identify valuable customers [3], it is not suitable for data which contains non-numerical data-types. The K-Modes algorithm [13], extends the K-Means algorithm to cluster categorical data. This extension is done by using a simple matching dissimilarity measure, Hamming distance, for categorical features instead of the Euclidean distance function. Also modes were used in place of means for clusters and a frequency based technique was used to update modes in the clustering procedure to minimize the clustering cost function [14].

Clustering is performed on the most important demographic attributes: Age, Sex, Occupation, Income, Education Level and Address. Demographic segmentation assists financial institutions in understanding their customers better so that they can align their business and marketing strategies to effectively address their customers' needs. An additional factor, OnlineUser, is used to provide further insight on the behaviour of customers who transact online versus the customers who do not.

C. Classification

To predict whether a customer will transition to the online banking platform, the Decision Tree models were utilized with the dependent variable, OnlineUser. The dataset was then split, where 75% was used as training data, while the remaining 25% was used as testing data. The model with the best evaluation metrics was selected as the best performer for the identification of online customers.

In past literature about churn prediction, decision trees were proven to be accurate, fast and efficient in comparison to other techniques [15], therefore the decision tree models were adopted for in this paper. Decision trees can automatically derive the importance of attributes as well as handle the presence of outliers. Two algorithms were examined: Decision Tree and Extreme Gradient Boosting (XGBoost), for the task of identifying the most loyal customers who are most likely to transition to the online banking platform.

To enhance the performance of the decision tree, it is extended with the gradient boosting algorithm. The boosting approach is an ensemble technique which combines different classifier models by assigning weights to them. The weights are then iteratively adjusted over several trials until no further improvements can be made. Although each classifier may not have good performance, using a combination of the boosting approach models can improve the overall classification performance significantly. The boosting approach can avoid over-fitting so that the classifier can have good performance for both the training data and an unknown testing data [8].

We examined two techniques to accurately predict the class label of whether a customer would transition to the online platform. The first technique is considered the benchmark model (BM_{m_i}), where the decision tree models (m_i) did not utilize the clustering techniques. The second technique used the labels that represented the identity of clusters of each clustering algorithm as input to the decision tree models for prediction of online banking transition. In the second technique, each decision tree model was trained for each clustering technique and used labels of CLV rank label (CLV_L) and Online Demographic label (ODL_L).

TABLE II
RFM NORMALIZED MEAN VALUES BASED ON CLUSTER AND THEIR RESPECTIVE CLV.

Cluster	Recency	Frequency	Monetary	CLV	CLV Rank
1	0.27	0.71	0.81	0.728	1
2	0.47	0.10	0.52	0.431	4
3	0.29	0.42	0.55	0.492	3
4	0.41	0.33	0.79	0.658	2

Fig. 1. Non-Online User (0) and Online Banking User (1) based on CLV ranking.

V. RESULTS AND DISCUSSION

A. Transaction Cluster Analysis

After the elbow technique was performed, K was found to be four, and this was further supported by the silhouette score of 0.359. The silhouette score indicates that the clusters are well distinguishable and each cluster has its own significant customer traits. The AHP values for each RFM dimension was computed using Python's AHPy package which resulted in: $W_R = 0.112$, $W_F = 0.196$ and $W_R = 0.692$. The normalized mean values of each RFM attribute for each cluster as well as their respective CLV are provided in Table II.

Cluster 1 accounts for the majority of the dataset (33.1%) and has a low recency but high frequency and monetary value. This cluster is vital to target as they rank number one in terms of CLV, as they would have transacted frequently and in large sums of money in the past but would not have had a recent transaction. These customers can be considered a loyal segment as the monetary value is essential in the financial sector as supported by the RFM AHP weightings [11]. Cluster 1 also has the highest number of customers who are online users. Based on Table II and Figure 1, it can be seen that the CLV Rank 1 (Cluster 1) is the group to actively target as they would have the highest propensity of transitioning online. This grouping consists of working class individuals who are most likely to transact frequently and with high monetary value.

B. Demographic Cluster Analysis

Using the elbow plot method, three was chosen as the optimal value for the number of clusters, K. Based on Table III, it is observed there are no distinguishable features with the exception of Region and Occupation. The results obtained show that the K-Means algorithm is not suitable in terms of segmenting categorical variables.

Based on the six demographic attributes as well as the Online User class attribute, a value of K=7 was chosen based on the Elbow plot technique. As seen in Fig 2, cluster 2 is most likely to transition online, therefore marketing techniques should be aimed towards this specific cluster, which represents the second highest population percentage (21.6%). It is derived from Table IV that the working class, middle income male from the eastern region has the highest likelihood of transitioning to the online platform, therefore specific promotions should be crafted to suit this demographic. It is also observed from Table IV that both female and male labourers who are lower middle income earners have higher number of online banking non-users. These segments would require much effort and resources to transition online.

TABLE III
PREVALENT FEATURES IN EACH K-MEANS DEMOGRAPHIC AND ONLINE USER CLUSTERS

	Online User	Age Group	Sex	Income Range	Education Level	Region	Job
C1	No	Middle Aged	F	Lower Mid Income	Secondary	North	Retired
C2	No	Matured	F	Lower Mid Income	Secondary	Central	Clerk
C3	No	Middle Aged	F	Lower Mid Income	Secondary	South	Clerk

TABLE IV
PREVALENT FEATURES IN EACH K-MODES DEMOGRAPHIC AND ONLINE USER CLUSTERS

	Online User	Age Group	Sex	Income Range	Education Level	Region	Job
C1	No	Middle Aged	F	Lower Mid Income	Secondary	Central	Clerk
C2	Yes	Middle Aged	M	Mid Income	Secondary	East	Labourer
C3	Yes	Senior Citizens	F	Lower Mid Income	Secondary	North	Retired
C4	No	Senior Citizens	M	Lower Mid Income	Secondary	East	Retired
C5	No	Senior Citizens	F	Low Income	Secondary	East	Retired
C6	No	Matured	F	Lower Mid Income	Secondary	East	Labourer
C7	Yes	Middle Aged	F	Lower Mid Income	University	South	Teacher

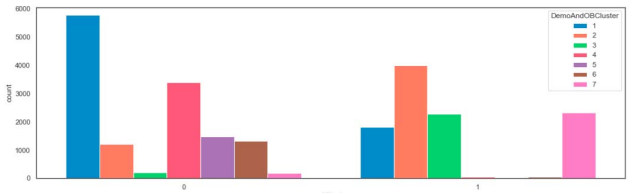

Fig. 2. Non-Online User (0) and Online Banking User (1) based on K-Modes - Demographic and Online User Clusters.

C. Classification

When comparing Table V and Table VI, it was observed that the top performing model was the XGBoost model using the two clusters generated by the K-Modes clustering algorithm: $CLV_{L_{DT}} + ODL_{DT}$ with an accuracy score of 96.1% and re-

TABLE V
RESULTS OF THE DECISION TREE ALGORITHM

Metric	BM_{DT}	**K-Modes** $CLV_{L_{DT}} + ODL_{DT}$	**K-Means** $CLV_{L_{DT}} + ODL_{DT}$
Accuracy	83%	93%	86%
Precision	81%	91%	85%
Recall	80%	92%	84%
F1-Score	80%	92%	84%

TABLE VI
RESULTS OF THE XGBOOST ALGORITHM

Metric	BM_{XGB}	**K-Modes** $CLV_{L_{XGB}} + ODL_{XGB}$	**K-Means** $CLV_{L_{XGB}} + ODL_{XGB}$
Accuracy	88%	96%	90%
Precision	90%	96%	92%
Recall	82%	95%	86%
F1-Score	86%	96%	89%

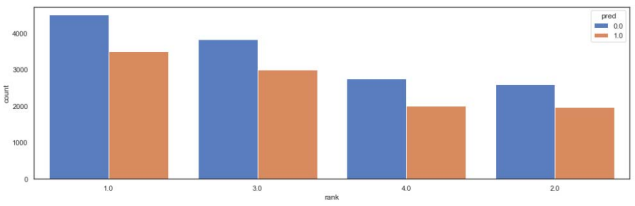

Fig. 3. Non-Online User (0) and Online Banking User (1) Prediction by CLV Rank Cluster.

call score of 95%. All the models with this cluster combination outperformed the other cluster combination models while all benchmark mode accuracy scores for all decision tree models performed the worst. Figure 3 presents the predictions of online user versus non-user for the best performing XGBoost model. It is observed the number one CLV Rank contributes the most to the total number of online banking users. This shows that clients with high CLVs can be targeted as they would be most receptive to bank promotions and marketing techniques.

Based on these results, it was deduced that clustering improves the accuracy of classification models. It is also noted that the majority of evaluation metrics for the cluster combinations based on the K-Modes clustering algorithm outperformed the cluster combination based on the K-Means clustering algorithm. From this observation, it was extrapolated that the K-Modes approach is more suitable for categorical attributes. Similarly, the best type of model, XGBoost, provided the highest accuracy and recall scores.

VI. CONCLUSION

The results of this research highlighted that the two dimensional (demographic and transactional behaviour) focused approach was favourable as the best performing hybrid model, K-Modes clustering algorithm and XGBoost classification model, outperformed the benchmark model with a test accuracy of 96.1% and recall of 95%. This research provided an online banking user propensity model that takes into consideration the loyalty of the customer as well. Further enhancements can be done to model produced by extending the model to monitor and forecast customers' transaction behaviour. This would be beneficial as financial institutions can develop an agile and efficient online platform that can quickly scale up or down to meet customer demand.

REFERENCES

[1] E. Silver Kwendo, M. J. Kirwa, and O. C. Omondi, "An overview of the influence of customer segmentation strategy on performance of commercial banks in Kenya," International Journal of Management and Commerce Innovations, vol. 5, pp. 644-650, 2018.

[2] Y. Feng, X. Wang and L. Li, "The Application Research of Customer Segmentation Model in Bank Financial Marketing," 2019 2nd International Conference on Safety Produce Informatization (IICSPI), 2019, pp. 564-569, doi: 10.1109/IICSPI48186.2019.9095900.

[3] M. Zubair, A. Iqbal, A. Shil, M. Chowdhury, M. A. Moni, and I. Sarker, "An Improved K-means Clustering Algorithm Towards an Efficient Data-Driven Modeling," Annals of Data Science, pp. 1–20, Jun. 2022, doi: 10.1007/s40745-022-00428-2.

[4] E. Umuhoza, D. Ntirushwamaboko, J. Awuah and B. Birir, "Using Unsupervised Machine Learning Techniques for Behavioral-based Credit Card Users Segmentation in Africa," in SAIEE Africa Research Journal, vol. 111, no. 3, pp. 95-101, Sept. 2020, doi: 10.23919/SAIEE.2020.9142602.

[5] R. Rahmadhan and M. Wasesa, "Segmentation using Customers Lifetime Value: Hybrid K-means Clustering and Analytic Hierarchy Process," Journal of Information Systems Engineering and Business Intelligence, 2022.

[6] J. Wu et al., "An Empirical Study on Customer Segmentation by Purchase Behaviors Using a RFM Model and K-Means Algorithm," Mathematical Problems in Engineering, vol. 2020, pp. 1–7, Nov. 2020, doi: 10.1155/2020/8884227.

[7] S. Pradhan, G. Patel, and P. Priya, "Measuring Customer Lifetime Value: Application of Analytic Hierarchy Process in Determining Relative Weights of 'LRFM'", IJAHP, vol. 13, no. 3, Dec. 2021.

[8] N. H. Syani, A. Amirullah, M. B. Saputro, and I. A. Tamaroh, "Classification of potential customers using C4.5 and k-means algorithms to determine customer service priorities to maintain loyalty", J. Soft Comput. Explor., vol. 3, no. 2, pp. 123 - 130, Sep. 2022.

[9] E. Ernawati, S. Baharin, and F. Kasmin, "A review of data mining methods in RFM-based customer segmentation," Journal of Physics: Conference Series, vol. 1869, p. 012085, Apr. 2021, doi: 10.1088/1742-6596/1869/1/012085.

[10] A. Hiziroglu, M. Sisci, H. I. Cebeci, and O. F. Seymen, "An empirical assessment of customer lifetime value models within data mining," Balt. J. Mod. Comput., vol. 6, no. 4, 2018.

[11] F. Zaheri, H. Farughi, H. Soltanpanah, S. Alaniazar, and F. Naseri, "Using multiple criteria decision making models for ranking customers of bank network based on loyalty properties in weighted RFM model," Management Science Letters, vol. 2, pp. 697–704, Apr. 2012, doi: 10.5267/j.msl.2012.01.018.

[12] N. Jha, D. Parekh, M. Mouhoub, and V. Makkar, "Customer Segmentation and Churn Prediction in Online Retail," 2020, pp. 328–334. doi: 10.1007/978-3-030-47358-7_33.

[13] K. S. Dorman and R. Maitra, "An efficient k-modes algorithm for clustering categorical datasets," Statistical Analysis and Data Mining: The ASA Data Science Journal, vol. 15, no. 1, pp. 83–97, 2022, doi: https://doi.org/10.1002/sam.11546.

[14] S. B. Salem, S. Naouali, and Z. Chtourou, "A rough set based algorithm for updating the modes in categorical clustering," International Journal of Machine Learning and Cybernetics, vol. 12, no. 7, pp. 2069–2090, Jul. 2021, doi: 10.1007/s13042-021-01293-w.

[15] P. Cuadros-Solas, S. Valverde, and F. Rodriguez-Fernandez, "A machine learning approach to the digitalization of bank customers: Evidence from random and causal forests," PLoS ONE, vol. 15, Oct. 2020, doi: 10.1371/journal.pone.0240362.