Contents lists available at www.journal.unipdu.ac.id

Register

Journal Page is available to https://journal.unipdu.ac.id/index.php/register/

Research article

Customer Churn Prediction Using the RFM Approach and Extreme Gradient Boosting for Company Strategy Recommendation

Mohammad Isa Irawan [a,*], *Nadhifa Afrinia Dwi Putris* [b], *Noryanti binti Muhammad* [c]

[a,b] *Department of Mathematics, Institut Teknologi Sepuluh Nopember, Jl. Raya ITS Sukolilo, Surabaya, 60111, Indonesia*
[c] *Centre for Mathematical Sciences, Universiti Malaysia Pahang, Lebuh Persiaran Tun Khalil Yaakob, Pahang, 26300, Malaysia*
email: [a,*] *mii@its.ac.id*, [b] *nadhifaafrinia25@gmail.com*, [c] *noryanti@ump.edu.my*
* *Correspondence*

ARTICLE INFO

Article history:
Received August 16, 2023
Revised February 11, 2024
Accepted November 30, 2024
Available online Desember 25, 2024

Keywords:
Customer Churn
Customer
Companies
XGBoost

Please cite this article in IEEE style as:
M. I. Irawan, N. A. D. Putris, and N. Muhammad, "Customer Churn Prediction Using The RFM Approach and Extreme Gradient Boosting for Company Strategy Recommendation," *Register: Jurnal Ilmiah Teknologi Sistem Informasi*, vol. 10, no. 2, pp. 127-140, 2024.

ABSTRACT

Customers are vital assets in the growth and sustainability of business organizations. However, customers may discontinue their engagement with a company and switch to competitors' products or services for various reasons. This event referred to as customer churn. Losing customers significantly impacts a company's revenue, often resulting in financial decline. Churn events, which are subject to dynamic monthly changes, are further influenced by intense competition and rapid technological advancements. Analyzing customer characteristics is crucial to understanding customer behavior, with metrics such as recency, frequency, monetary (RFM) serving as key indicators of subscription and transaction patterns. The Extreme Gradient Boosting method is applied to address the challenge of classifying churn and non-churn customers. The prescriptive analytics process is carried out to identify the features most influential in prediction outcomes, enabling the formulation of strategic recommendations to mitigate churn problems. The integration of RFM analysis with the XGBoost method provides optimal results, particularly in the third segmentation, achieving an accuracy of = 0.98833, precession = 0.98768, recall = 0.98899, and f1-score = 0.98833. The prescriptive analytics process highlights three critical features, namely city factor, GMV generation, and total customer transaction generation. This findings demonstrate that the segmentation characteristics, data representation, and behavioral approach with RFM analysis have an effect on improving the performance of the model in churn prediction.

1. Introduction

In today's business era, a key requirement for a company's growth is cultivating a loyal customer base. Customers, whether individuals or groups, purchase or utilize the company's products and services. A large customer base typically increases business revenue. As of January 2023, approximately 835 companies were listed on the Indonesia Stock Exchange (IDX) [1]. Furthermore, in 2022, the number of Micro, Small, and Medium Enterprises (MSMEs) in Indonesia reached approximately 9.137 million, according to the Ministry of Cooperatives and SMEs of the Republic of Indonesia [2].

Today, global technological advancements are driving increasingly intense competition among companies striving to enhance the quality of their services and products. The large number of competitors forces businesses, including Micro, Small and Medium Enterprises (MSMEs), to compete aggressively to attract and retain as many customers as possible. The diversity and sheer volume of customers present significant challenges for companies [3]. Each customer has unique expectations, needs, behaviors, and income profiles, making it crucial for companies to understand these characteristics. Failure to address this aspect increases the risk of customers leaving and switching to competitors' products or services, a phenomenon known as customer churn [4].

The churn phenomenon draws significant attention across all business sectors, as it can result in financial losses and threaten a company's long-term sustainability amid intense competition [5]. To mitigate churn, companies can implement strategies focused on customer retention by leveraging customer segmentation based on behavioral patterns [6]. One widely used method for analyzing customer behavior is recency, frequency, monetary (RFM) analysis. RFM analysis involves examining customer behavior through three key variables, namely recency, which tracks the most recent customer activity; frequency, which reflects the number of transactions or interactions; and monetary, which represents the total value of customer transactions [7].

Chen and Guestrin introduced the Extreme Gradient Boosting (XGBoost) method in 2016 as an open-source project, providing an efficient, fast, and scalable machine learning system [8]. XGBoost leverages the outputs of previous classification models, sequentially combining them while accounting for relevant errors. This iterative process continues until a model with high performance and minimal error is achieved.

Previous research on customer churn prediction was conducted by Herawati et al., who utilized the Fuzzy Iterative Dichotomiser 3 method to examine the impact of fuzzy curve representation, Fuzziness Control Threshold (FCT) value, Leaf Decision Threshold (LDT) value, and training data size on the number of rules generated and their accuracy. Their findings indicated that the trapezoidal curve representation yielded the highest accuracy and the largest number of rules. However, experiments with this representation experienced overfitting due to a low FCT value, which inflated accuracy [9]. Research on churn has also been conducted by Shresta and Shakya. In their study, they used the XGBoost method to examine two different datasets. The resulting accuracy on the general dataset is 96.25%, while the accuracy on the Nepalese telecommunications industry-specific dataset is 97%. Their study highlighted the role of machine learning in customer segmentation and its effect on churn prediction [10]. Mena et al. conducted another study that compared churn prediction using an LSTM model with and without Recency, Frequency, Monetary (RFM) features, as well as static features. The results demonstrated that the LSTM model incorporating RFM features achieved a higher AUC level than the LSTM model with static features alone, achieving AUC levels of 0.779 and 0.775, respectively [11]. Zhang and Zhang applied the GWO-attention-ConvLSTM model for customer churn prediction and satisfaction analysis within customer relationship management system [12]. Additionally, Poudel et al. investigated churn prediction in the telecommunications industry using a tabular machine-learning model [13], while Usman-Hamza et.al. proposed a novel heterogeneous multi-layer stacking ensemble method for the same sector [14].

Building on previous studies, this research employs RFM analysis and the XGBoost method to evaluate customer characteristics based on subscription and transaction behavior, aiming to predict churn and non-churn customers using available data. Additionally, prescriptive analytics is applied to recommend targeted company strategies. The research seeks to provide insights into customer segmentation based on behavioral characteristics, predict customer classification, and offer specific strategic recommendations for the company. The research process involves several steps: data collection, data preprocessing, feature engineering, RFM analysis, oversampling, model development using the XGBoost method, prescriptive analytics, and finally, drawing conclusions and providing recommendations. The findings can serve as valuable input for companies to enhance customer retention strategies.

This research article is organized into three main chapters: introduction, materials and methods, results and discussion, and conclusion. The Introduction outlines the background of the study, a review of the state of the art summarizing previous related research, and the objectives of the study. The Materials and Methods section describes the fundamental theory behind this research and the methodology employed. Results and discussion explain the results and discussion of this research. Conclusion contains a brief conclusion of this research.

2. Materials and Methods

2.1 Customer Churn

Customer churn refers to the phenomenon where customers become inactive or discontinue their relationship with a company due to various conditions. Customers are invaluable assets for companies, and losing them can negatively impact revenue and threaten the company's long-term sustainability. In business, there are two primary categories: contract and non-contract businesses. In contract businesses, customer churn can occur due to behaviors such as contract cancellation by active customers, failure to renew subscriptions after the subscription period ends, or the company discontinuing services for specific reasons. Conversely, in non-contract businesses, churn may be identified through behaviors like officially terminating customer relationships or customers disregarding the company's presence. This issue makes it challenging for companies to identify and classify customers as experiencing churn [4].

2.2 RFM Analysis

RFM analysis is a method used to examine behavior on specific contexts, often applied for customer segmentation. It incorporates three key variables: Recency (R), Frequency (F), and Monetary (M). Recency refers to the time elapsed since the customer's last interaction with the company. A lower recency value indicates a more recent interaction. Frequency measures the number of interactions or behaviors a customer exhibits within a specified time frame, with a higher frequency value signifying greater customer engagement. Monetary represents the total transaction money for each customer within a certain period. A higher monetary value reflects a greater total transaction amount for the customer [15]. We applied four stages in the RFM analysis, which are outlined as follows [7]:

Step I: Calculate the value of RFM variables

The first step in conducting RFM analysis is to calculate the value for each RFM variable. The recency variable is determined by comparing the last day all customers behaved and the last day a customer behaved. The equation for calculating the recency variable is presented in Equation 1:

$$recency = max\ days\ of\ all\ customers - \ specific\ customer's\ max\ days \quad (1)$$

Frequency variable is the sum of all recorded behavioral activities or formulated as Equation 2:

$$frequency = \sum_{i=1}^{n} behavioral\_activity\_feature_i \quad (2)$$

where n = number of behavioral activities. In calculating the monetary variable, it depends on the total transaction money of a behavior as shown in Equation 3:

$$monetary = \sum_{i=1}^{m} money\_feature_i \quad (3)$$

Step II: Convert variable values into a scale score

The second step in RFM analysis involves converting variable values into a standardized score scale. It is crucial to adapt the scoring scale based on the company's specific data distribution, requirements, and business characteristics. Several methods can be employed to convert the variable values obtained in the previous step into scores, such as using a quartile approach or a tailored method based on the relevant business context. The quartile approach groups RFM variable values into quartiles based on the data distribution. The mathematical formula for calculating quartiles Qi is provided in Equation 4:

$$Q_i = X_{\frac{i(n+1)}{4}} \quad (4)$$

Each quartile represents a different segment, determined based on the relative position of the RFM variable values. Table 1 is presented to determine the recency, frequency, and monetary scores based on the quartile approach.

Table 1. Convert Value of R, F, M

Location of recency, frequency, and monetary value	R Score	F and M Score
$recency, frequency, monetary \leq Q_1$	r	1
$Q_1 < recency, frequency, monetary \leq Q_2$	r-1	2
...	…	…
$Q_p < recency, frequency, monetary$	1	r

where p = last quartile and r = last score value.

Step III: Summation of RFM Scores

The RFM score is determined by adding up the RFM variable scores of each customer. This approach provides a more structured understanding of customer behavior, allowing customers into several segments according to company policy as illustrated in Equation 5.

$$RFM = R + F + M \quad (5)$$

Step IV: Customer Segmentation

The RFM value will be accumulated with all customers and then customer segmentation will be carried out. The division of customer segmentation is shown in Table 2.

Table 2. Customer Segmentation

Location of RFM value	Segmentation
$RFM \leq Threshold_1$	s
$Threshold_1 < RFM \leq Threshold_2$	s-1
...	…
$Threshold_q < RFM$	1

where q = last threshold and s = last segmentation. Threshold values, shown in Table 2, are set according to company needs, company policy, and customer characteristics.

2.3 Extreme Gradient Boosting (XGBoost)

Extreme Gradient Boosting or XGBoost is a tree-based machine learning algorithm designed to reduce and mitigate overfitting through a systematic approach to building a regression tree structure. The principle adopted by XGBoost is to utilize the outcomes of previous classification models, sequentially combining them while addressing relevant errors. This iterative process constructs a stronger model that minimizes errors, resulting in more accurate predictions [16].

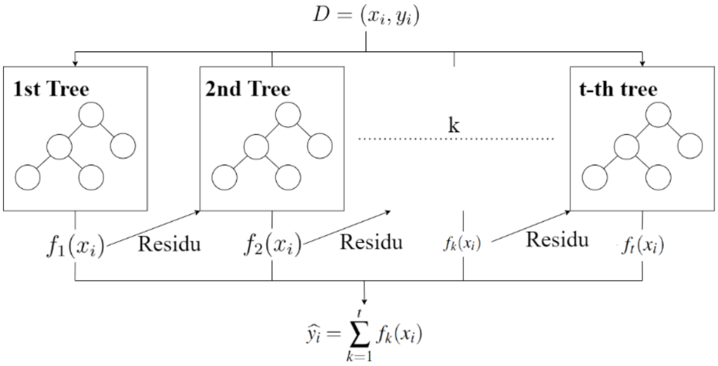

Fig. 1. Illustration of the XGBoost Method

Fig. 1 shows how XGBoost works with dataset D with n = examples, m = features, and $\hat{y}_i$ is described in Equation 6:

$$\hat{y}_i = \sum_{k=1}^{t} f_k(x_i)\,, f_k \in \mathcal{F} \tag{6}$$

where $\hat{y}_i$=final tree model prediction, f_k = k-th tree, x_i = i-th data feature, and $\mathcal{F} = \{f(x) = w_{q(x)}\}(q: \mathbb{R}^m \to T, w \in \mathbb{R}^T)$. In training the model, it is necessary to determine the objective function to build the tree by minimizing Equation 7:

$$\mathcal{L}(\emptyset) = \sum_{i=1} l(y_i, \hat{y}_i) + \sum_{i=1} \Omega(f_k) \tag{7}$$

where $\mathcal{L}$ = loss function measures the difference between the prediction and the target and Ω = represents the regularization function of the kth tree. The regularization function can control the complexity of the model to avoid overfitting. The loss function and regularization function are described as follows Eq. 8 and Eq. 9:

$$l(y_i, \hat{y}_i) = y_i log(\hat{y}_i) + (1 - y_i) log(1 - \hat{y}_i) \tag{8}$$

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda w^2 \tag{9}$$

where γ = for pruning tree, T = number of leaves, λ = regularization parameter, and w = output value. There is a term similarity score for each leaf node which is denoted as S and shown in Equation 10. Similarity score shows the similarity value between leaves.

$$S = \frac{\left(\sum_{i \epsilon I_j}(y_i - \hat{y}_i)\right)^2}{\sum_{i \epsilon I_j}(y_i - \hat{y}_i) + \lambda} \tag{10}$$

where I_j = set of leaf nodes j. There is a term gain which is denoted as G to indicate the information from the leaf. If the gain value is maximized, then the gain is chosen to be the threshold. Gain is formulated in Equation 11:

$$G = \left[\frac{\left(\sum_{i \epsilon I_L}(y_i - \hat{y}_i)\right)^2}{\sum_{i \epsilon I_L}(y_i - \hat{y}_i) + \lambda} + \frac{\left(\sum_{i \epsilon I_R}(y_i - \hat{y}_i)\right)^2}{\sum_{i \epsilon I_R}(y_i - \hat{y}_i) + \lambda} - \frac{\left(\sum_{i \epsilon I}(y_i - \hat{y}_i)\right)^2}{\sum_{i \epsilon I}(y_i - \hat{y}_i) + \lambda}\right] \tag{11}$$

where I_L = set of left sample and I_R = set of right sample. To obtain a new predicted value, the output value is calculated, which is denoted as w. The output value is formulated in Equation 12:

$$w^*{}_j = \frac{\sum_{i\epsilon I}(y_i - \hat{y}_i)}{\sum_{i\epsilon I}(y_i - \hat{y}_i) + \lambda} \tag{12}$$

The output value does not yet describe the predicted value of each leaf. Therefore, the output value is converted into the predicted value of each leaf which is denoted as $P'(y_i)$ using a logistic function like Equation 13 and Equation 14 [17]:

$$P'(y_i) = \frac{e^{logit(\hat{y}_i)}}{1 + e^{logit(\hat{y}_i)}} \tag{13}$$

$$logit(\hat{y}) = log\left(\frac{\hat{y}}{1 - \hat{y}}\right) \tag{14}$$

In statistics, it is said to be log with an Euler number base. The calculation of the new predicted value is formulated as Equation 15:

$$logitP'(\hat{y}_i) = logit(\hat{y}) + (learningrate)\left(w^*{}_j\right) \tag{15}$$

The learning rate controls how much each tree contributes to the final prediction of the model. The prediction value generated in Equation 15 is still in $logitP'$ form, so it is converted into $P'(y)$ form using Equation 13.

2.4 Performance Calculation

We utilized four functions to calculate the performance of the model: True Positive (TP), which represents positive data correctly predicted as positive data; True Negative (TN), which represents negative data correctly predicted as negative ; False Positive (FP), which represents negative data incorrectly predicted as positive; and False Negative (FN) which represents positive data incorrectly predicted as negative. Additionally, four commonly used evaluation metrics, accuracy, precession, recall, and f1-score are defined in Equation 16, Equation 17, Equation 18, and Equation 19 [18]:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{16}$$

$$Precission = \frac{TP}{TP + FP} \tag{17}$$

$$Recall = \frac{TP}{TP + FN} \tag{18}$$

$$F1Score = 2\left(\frac{Recall * Precission}{Recall + Precission}\right) \tag{19}$$

2.5 Feature Importance

Feature importance is a technique used to evaluate the significance of input features within a constructed model. Features with higher importance values indicate a greater influence on the model's ability to predict specific variables. In this study, the Gini index was applied to calculate the initial stage of determining feature importance. Suppose there are c features with X1, X2, X3,...,Xc , Equation 20 shows the Gini index equation [18]:

$$GI_m = 1 - \sum_{k=1}^{K} \sum_{k \neq k'} p_{m,k} p_{m,k'} \tag{20}$$

where K = number of categories, $p_{m,k}$ = probability of k-th category, and $p_{m,k'}$ = the probability of the other category represented by k'. In each sub-tree there are nodes that affect the feature importance value. So, we used IV formula as the value of feature importance in Equation 21:

$$IV_{j,m} = p_m GI_m - p_l GI_l - p_r GI_r \tag{21}$$

where p = probability every node, l and r = new nodes after branching. All features are normalized against all features in M models based on Equation 22:

$$IV_j = \frac{\sum_{m \in M} IV_{j,m}}{\sum_{i=1}^{c} IV_i} \tag{22}$$

3. Results and Discussion

We used data from 24,194 customers obtained from PT KPI, a software company providing services to MSMEs. The pre-processed data was utilized to create new features such as 'gmv', 'transaction_amount', 'avg_recommendation_score', promo_amount', price_amount', 'gmv-10', 'order-10', 'gmv-20', 'order-20', 'gmv-30', and 'order-30'. Subsequently, we conducted RFM analysis to generate the 'RFM' feature, which identifies customer characteristics related to subscription and transaction behavior. Customers were then divided into threedistinct segments based on their RFM values. To optimize classification results, we applied the XGBoost method after performing oversampling using the Synthetic Minority Oversampling Technique (SMOTE) on a combination of the newly created features and additional features, namely 'city' and 'field,' as per the composition of each segment. Finally, predictive modeling was performed for each segment and for the unsegmented dataset.

3.1 RFM Analysis

We analyzed the data using recency, frequency, and monetary (RFM) analysis to understand customer behavior, there are several processes, including:

Step I: Calculate the value of RFM variables.

The recency variable is calculated by subtracting the maximum last recorded time of all customers from the last recorded time of the specific customer. The frequency variable is determined by counting the total number of behavior records for each customer. The monetary variable is derived from the total transaction value for each customer. The features used to obtain the recency, frequency, and monetary variables are presented in Table 3.

Table 3. Acquisition of Recency, Frequency, and Monetary variables

Behavior	Variable	Feature
Subscription	Recency	End_time
	Frequency	Type
	Monetary	Price
Transaction	Recency	Date
	Frequency	Transaction_amount
	Monetary	GMV

Step II: Convert variable values into a scale score

We applied the quartile approach in assigning scores of one to four as shown in Table 4 and Table 5.

Table 4. Conversion of RFM Transaction Variable Value

Recency	R	Frequency	F	Monetary	M
$recency \leq 139$	4	$frequency \leq 3$	1	$monetary \leq 410500$	1
$139 < recency \leq 141$	3	$3 < frequency \leq 24$	2	$410500 < monetary \leq 4068000$	2
$141 < recency \leq 156$	2	$24 < frequency \leq 222$	3	$4068000 < monetary \leq 30127250$	3
$156 < recency$	1	$222 < frequency$	4	$30127250 < monetary$	4

Table 5. Conversion of RFM Subscription Variable Value

Recency	R	Frequency	F	Monetary	M
$recency \leq 22$	4	$frequency \leq 1$	1	$monetary \leq 55500$	1
$22 < recency \leq 66$	3	$1 < frequency \leq 1$	2	$55500 < monetary \leq 55500$	2
$66 < recency \leq 125$	2	$1 < frequency \leq 1$	3	$55500 < monetary \leq 111000$	3
$125 < recency$	1	$1 < frequency$	4	$111000 < monetary$	4

Step III: Summation of RFM Scores

We summed all converted RFM variables to obtain the final RFM values. The distribution of customer RFM scores for subscription behavior, is illustrated in Fig. 2, while the distribution for transaction behavior is shown in Fig. 3. Subsequently, the RFM scores for all behaviors were aggregated. Customer who engaged in only one type of behavior were excluded from the analysis.

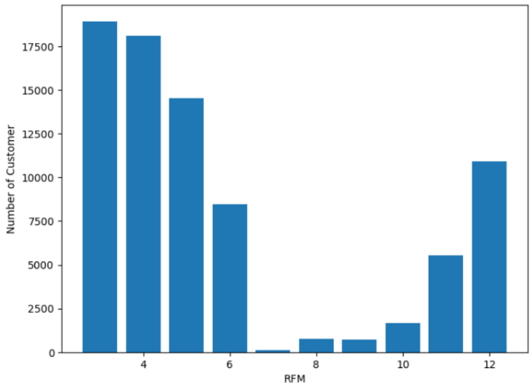

Fig. 2. The Distribution of Customer RFM Score Data on Subscription Behavior

Fig. 2 indicates that the dominant customer interaction in subscription behavior is infrequent. This conclusion is based on the RFM analysis, where a score of 3 contains the highest number of customers.

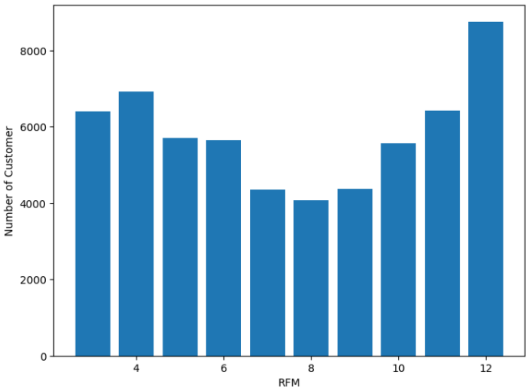

Fig. 3. The Distribution of Customer RFM Score Data on Transaction Behavior

Meanwhile, Fig. 3 reveals that an RFM value of 12 represents the largest number of customers, indicating a high level of interaction in transaction behavior. This suggests that while the majority of the company's customers subscribe infrequently, they engage in transactions more frequently.

Step IV: Customer Segmentation

We segmented the customer data into three groups: the first segment includes customers with RFM scores of 19, 20, 21, 22, 23, and 24; the second segment includes scores of 12, 13, 14, 15, 16, 17, 18, and the third segment includes scores of 6, 7, 8, 9, 10, 11. A comparison of customer segmentation results with unsegmented data is presented in Table 6.

Table 6. Comparison of Customer Segmentation Data

Composition	Number of Data	Number of Class Labels
First Segmentation	13751	*Churn*: 1721 *Non-churn*: 12030
Second Segmentation	3636	*Churn*: 2919 *Non-churn*: 717
Third Segmentation	6807	*Churn*: 6724 *Non-churn*: 83
Not Segmented	24194	*Churn*: 11364 *Non-churn*: 12830

3.2 Modelling

We implemented the XGBoost method to classify the customer with potential churn within each segmentation and compared the results to those obtained from unsegmented data. The model building process began with parameter tuning, employing Randomized Search with 10 iterations, Stratified Kfold Cross Validation 5 folds, and the objective function is binary: logistic as shown in Table 7.

Table 7. Parameter Tuning

Parameters	1st Segment	2nd Segment	3rd segment	No Segment and No RFM	No Segment and With RFM
Learning rate	0.3	0.3	0.3	0.3	0.1
Max depth	7	5	7	3	7
Min child weight	3	5	3	5	5
Gamma	0	0.5	0	0	0.1
N estimators	500	500	500	100	100
Subsample	0.8	0.8	0.8	0.8	1
Colsample bytree	1	0.5	1	0.5	1

The learning rate parameter controls the contribution or weight assigned to each tree in the ensemble model. Max depth specifies the maximum depth of each tree, while min child weight determines the minimum sample weight required for each leaf. The gamma parameter is used to set the minimum loss reduction, and n estimators represent the number of CART trees constructed. Subsample defines the proportion of training samples used, and colsample bytree specifies the number of features utilized. Based on the tuned parameters outlined in Table 7, trials were conducted to calculate the accuracy, precision, recall, and f1score.

Table 8. Average Evaluation Results

Composition	Accuracy	Precision	Recall	F1Score
First Segmentation	0.89564	0.90269	0.88703	0.89473
Second Segmentation	0.87581	0.87115	0.88215	0.87659
Third Segmentation	0.98833	0.98768	0.98899	0.98833
No Segment (No RFM)	0.88489	0.91202	0.83553	0.87209
No Segment (With RFM)	0.92523	0.94625	0.89150	0.91802

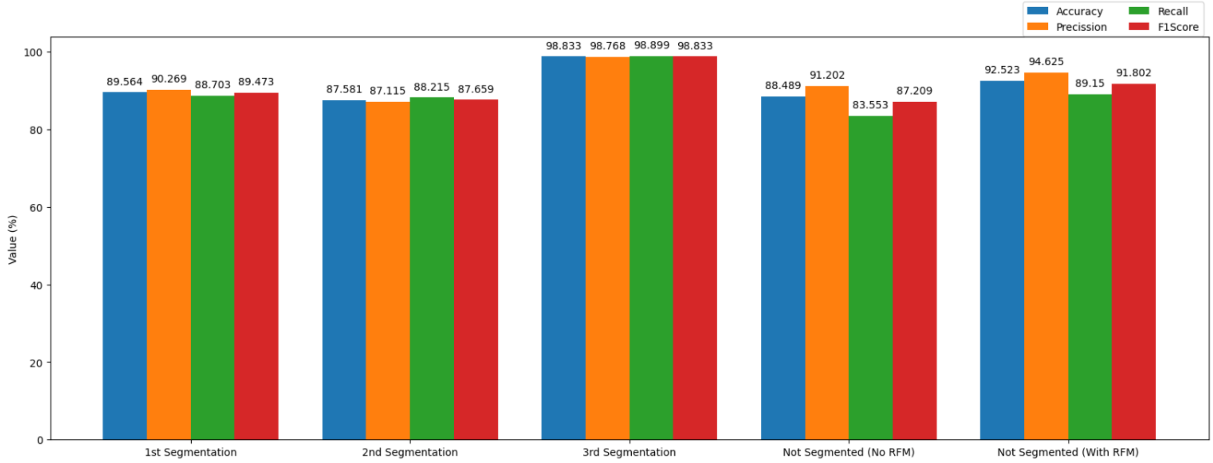

Fig. 4. Comparison of Average Evaluation Results on Overall Composition

Table 8 shows that the performance metrics of the third segmentation model are higher than those of the other segments. This can be attributed to the quality of the data and the distinct characteristics of the segmentation. The third segmentation exhibits a clearer and more easily identifiable pattern, whereas the first and second segmentations have more complex characteristics, making it challenging for the model to recognize their patterns. The superior performance of the first and third segmentations, compared to unsegmented data, is due to the more focused representation and consistent contributions of segmented data. This is evident from the high evaluation metrics achieved when using unsegmented RFM features. These results highlight the effectiveness of RFM in enhancing model performance, particularly when applying the XGBoost method.

Previous research by Sharesta and Shakya on churn prediction using the XGBoost method produced results consistent with the findings of this study. Sharesta and Shakya's research demonstrated high accuracy using the XGBoost method across two different datasets [10]. Similarly, this study achieved high accuracy across the overall composition, further highlighting the strong performance of the XGBoost method in classification tasks. This research incorporated the recommendations of Sharesta and Shakya by applying customer segmentation to the XGBoost model. Customer segmentation was performed using a behavioral approach with RFM analysis. The implementation of this recommendation revealed that the XGBoost model performs more effectively when applied to individual customer segments.

3.3 Prescriptive Analysis

In this research, perspective analytics was used to analyze the key features influencing the performance of the XGBoost method. Identifying important features for each model enables the development of strategic recommendations or solutions tailored to customers segmentation and overall data divisions. Fig. 5, Fig. 6, and Fig. 7 illustrate the acquisition of important features in each segmentation.

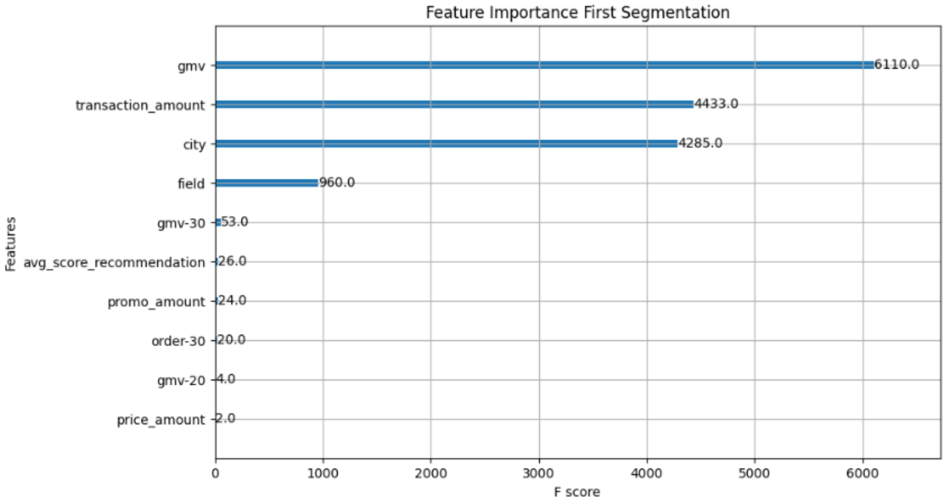

Fig. 5. Feature Importance of First Segmentation

Fig. 5 illustrates the important feature gains from the first segmentation modelThe gmv feature ranks highest, with a feature importance value of 6110, followed by the transaction_amount feature at 4433 and the city feature at 4285.

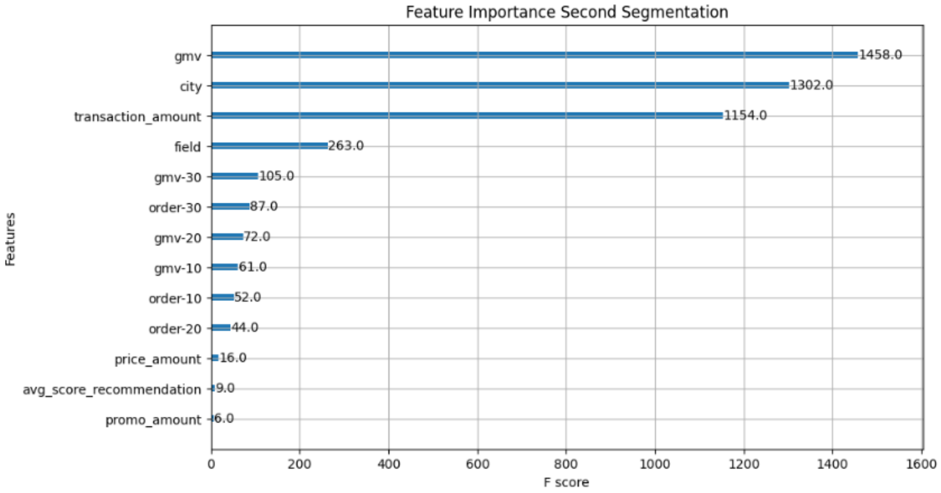

Fig. 6. Feature Importance of Second Segmentation

Fig. 6 illustrate the key features from the second segmentation model. The gmv feature ranks highest, with a feature importance value of 1458, followed by the city feature at 1302 and the transaction_amount feature at 1154.

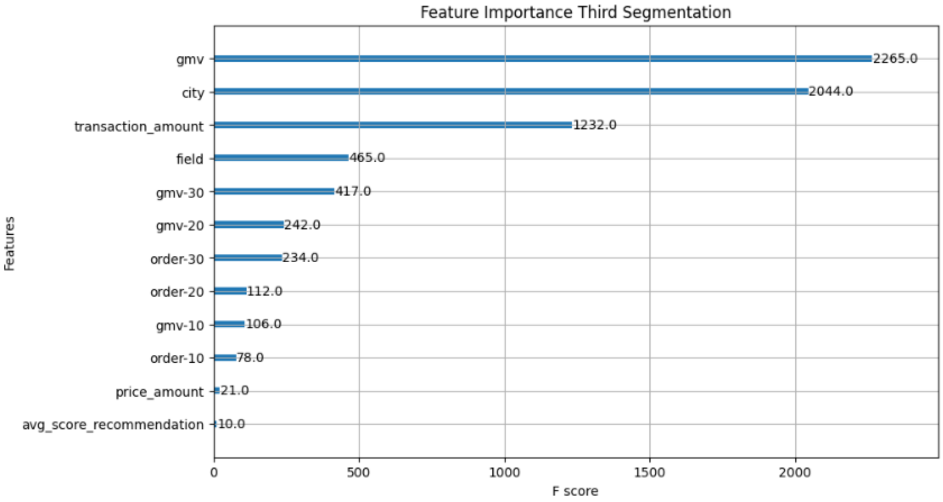

Fig. 7. Feature Importance of Third Segmentation

When Fig. 7 displays the important feature acquisition of the third segmentation model, the gmv feature occupies the first level with a feature importance value of 2265, followed by the city feature with a value of 2044 and the transaction_amount feature with a value of 1232. We observed that all segmentations identified the same three most important features, although their order and values varied. These features are the 'city' feature, the 'gmv' feature, and the ‘transaction_amount' feature. Thus, the 'city' feature significantly contributes to the model's ability to predict potential churn or non-churn customers. Prescriptive analytics can offer strategies for companies to further investigate cities that have high churn rates. It could be that there is a problem with the quality of service in that city, or there are specific customer needs in that city that have not been met.

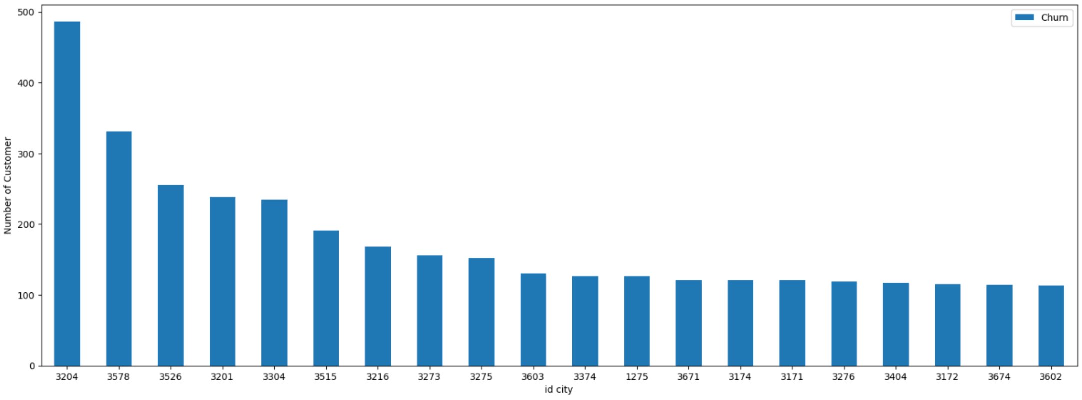

Fig. 8. Data Distribution of 20 Cities with Highest Churn

Fig. 8 illustrates the distribution of customer data for the 20 with the highest churned customers. The visualization highlights that cities with id 3204 should receive special attention from the company, as they have a significant number of churned customers, followed by cities with id 3578, among others. For instance, the company could target advertisements or offer special promotions in these cities, such as subscription discounts, reward vouchers, or customized solutions to meet specific customer needs and encourage continued use of the application. Additionally, localized training and support in these cities could enhance customer trust and satisfaction. Another potential strategy is to form partnerships with local businesses in high churn areas, creating mutual benefits while expanding the company's reach and offering greater value to customers in each city.

Next, the prescriptive analytics process focuses on the 'gmv' feature, which represents the transaction revenue of each customer. If the 'gmv' feature is found to have high importance in the classification of churn customers, it suggests that the GMV level of customers is strongly associated with their likelihood of churning.

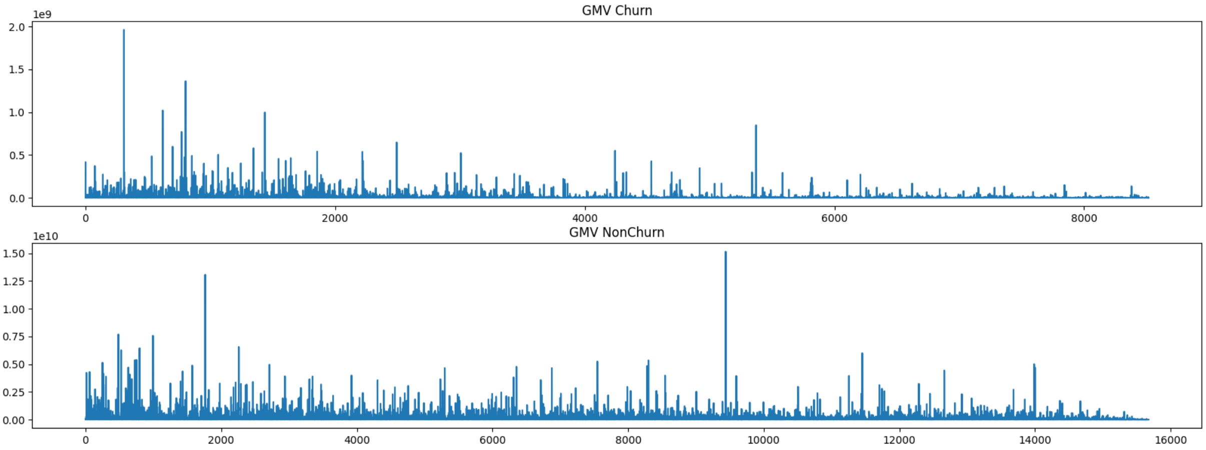

Fig. 9. Distribution of Customer GMV Data

Fig. 9 presents the GMV data distribution graph of all churned and non-churned customers. The X-axis represents the customer index, while the Y-axis shows the GMV acquisition for each customer. It can be seen that the GMV acquisition of customers identified as churn has a low value when compared to customers identified as non-churn. This indicates that customers with low GMV acquisition will potentially churn.

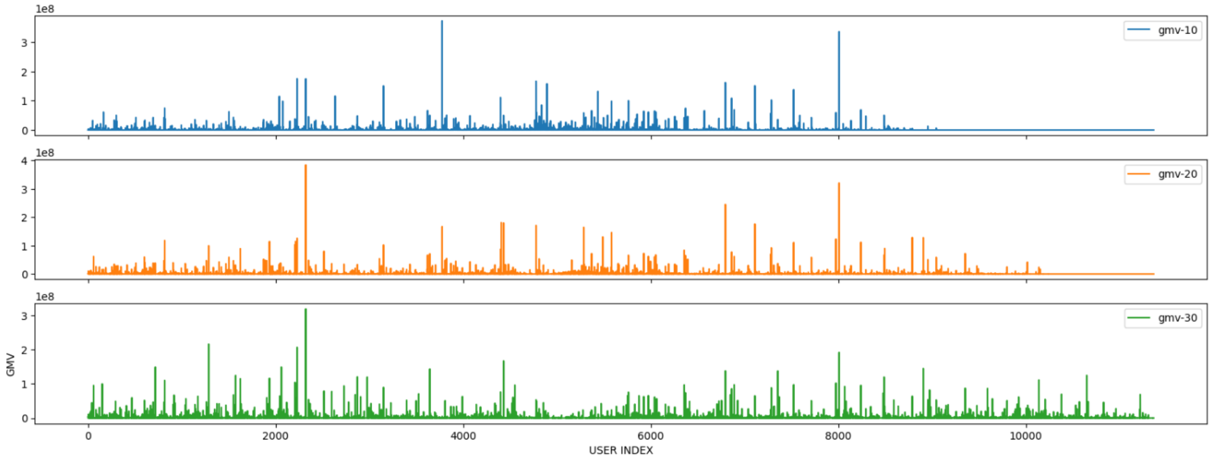

Fig. 10. Distribution of Churn Customer GMV Development Data

Fig. 10 displays the distribution of GMV growth data of customers during the last 30 days before their subscription expires. The customer index is represented as the X-axis, while the GMV value is represented as the Y-axis. The customer X-axis points on all three graphs are consistent. To interpret the visualization, focus on the example highlighted in the middle of the graph. The same X-axis point is compared across the GMV graphs for other time periods. Each graph corresponds to a specific time frame. The blue graph covers the last 10 days before the subscription time runs out. The orange graph represents the 10 to 20 days prior. While the green graph covers the 20 to 30 days before the subscription expiration. The distribution of GMV data for each time period is different. Among the 11364 churned customers, 3254 showed a decrease in GMV when getting closer to the last subscription time and 872 customers who experienced an increase in GMV when getting closer to the last subscription time. While the rest of the customers have zero values at GMV 30, GMV 20, and GMV 10. So there is no need to mention it. Thus, GMV has the effect that the more GMV acquisition decreases, the more customers have the potential to churn.

Prescriptive analytics can recommend strategies for companies to offer specially designed packages or services aimed at increasing customer GMV. Another approach could be implementing a reward program to encourage customer engagement and boost transaction volumes. For example, rewards for customers who have reached a specific GMV. Additionally, the company could provide business training programs to enhance customers' knowledge of business management.

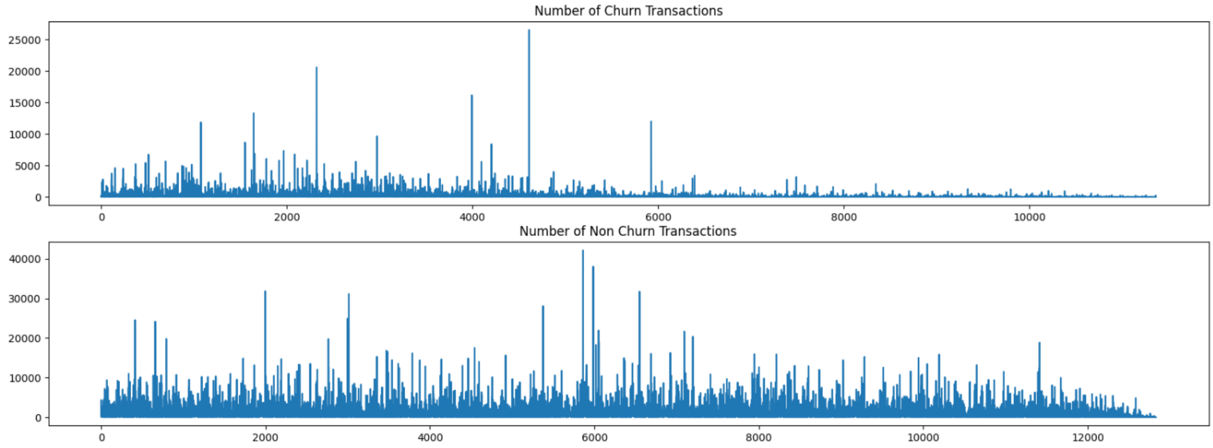

Fig. 11. Distribution of Customer Transaction Data

Fig. 11 presents the distribution graph of total transaction data for all churned and non-churned customers. The X axis represents the customer index, while the Y axis shows the total number of transactions per customer. It can be seen that the total transactions of customers identified as churn that reach a value above 25000 are only one customer. This suggests that customers with lower transaction volumes are more likely to churn.

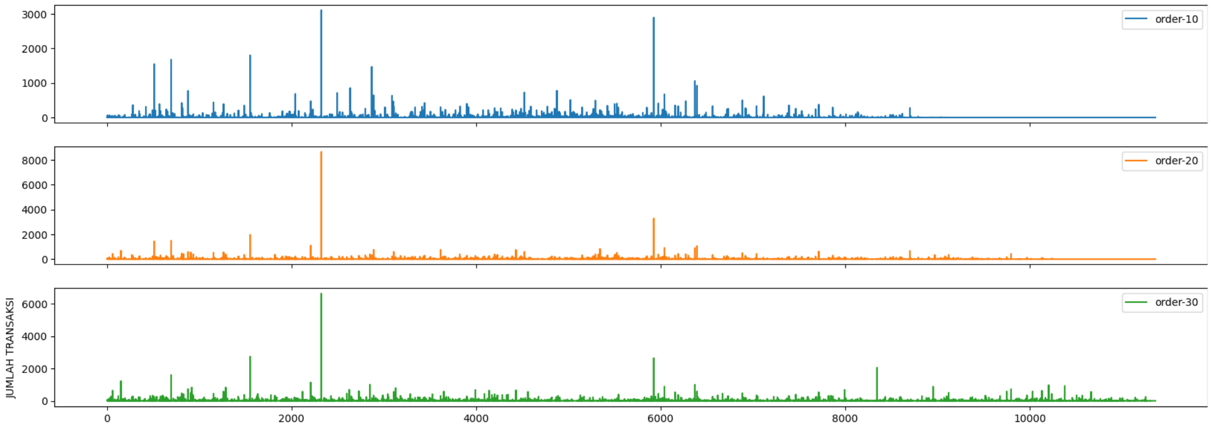

Fig. 12. Distribution of Data on the Development of the Number of Churn Customer Transactions

Fig.12 illustrates the distribution graph of total customer transaction data during the last 30 days before their subscription expires. The distribution of the data can be seen to have almost the same level as the distribution of GMV data. The 'transaction_amount' feature is strongly correlated with GMV acquisition. In the graphs, the X-axis represents the customers index, and the Y-axis denotes the number of transactions per customer. The customer X-axis points across all three graphs, are the same. To interpret the visualization, focus on the example highlighted in the middle of the graph. Each graph depicts total transactions over a specific time frame. The blue graph shows the graph in the last 10 days before the subscription time runs out. The orange graph shows a graph of the last 10 to 20 days before the subscription time runs out. Meanwhile, the green graph is a graph in the last 20 to 30 days before the subscription time runs out. The distribution of transaction data differs across these periods. Among the 11364 customers who experience churn, there are 3375 customers who experience a decrease in total transactions as the subscription time approaches and 896 customers who experience an increase in total transactions as the subscription time approaches. The rest of the customers have zero values in order 30, order 20, and order 10. Therefore, there is no need to mention it. As a result, if a customer experiences churn, the total transactions within the last 30 days will decrease.

Strategic recommendations based on the transaction_amount' feature include building customer communities through online or offline platforms. These communities can serve as spaces for customers to share experiences, tips, or inspiration for running their business. By engaging in these communities, customers may feel more connected and supported, which could motivate them to keep using the app and increase their number of transactions. Additionally, companies could offer business consulting services to help customers enhance their operational efficiency and sales strategies. During these consultations, companies could assist customers in analyzing their transaction data, identify growth opportunities, and provide advice on better inventory management, product pricing, and marketing strategies.

4. Conclusion

We proposed an Extreme Gradient Boosting (XGBoost) method in this study, leveraging customer behavior analysis based on RFM metrics to predict the classification of churn or non-churn customers. Customers were divided into three segments based on RFM values. The first segment included 13751 customers, the second segment 3636 customers, and the third segment 6807 customers. The third segmentation test achieved the highest metric evaluation, with accuracy = 0.98833, precision = 0.98768, recall = 0.98899, and f1score = 0.98833. It was concluded that segmentation characteristics, data representation, and behavioral approach using RFM significantly influenced the model's performance. We recommend strategies for companies based on the impact of city, GMV acquisition, and total customer transactions. For further research, we suggest using datasets from various companies with larger and more uniform customer distributions, analyzing additional behaviors, such as customer perceptions of applications, and comparing the XGBoost method with other boosting technique methods, such as Catboost and LightGBM.

Author Contributions

M. I. Irawan: Conceptualization, formal analysis, funding acquisition, methodology, project administration, review, investigation, validation. N. A. D. Putris: Investigation, methodology, software, resources, supervision, writing, data curation, visualization, programming, editing. N. B. Muhammad: Project administration, review, investigation.

Acknowledgment

The authors gratefully acknowledge financial support from the Institut Teknologi Sepuluh Nopember for this work under the Publication Writing and IPR Incentive Program (PPHKI) 2023 project scheme.

Declaration of Competing Interest

We declare that we have no conflict of interest.

References

[1] B. Indonesia. “Profil Perusahaan Tercatat.” BEI. https://www.idx.co.id/id/perusahaan-tercatat/profil-perusahaan-tercatat/ (accessed Jan. 21, 2023).

[2] K. UKM. “SATU DATA KUMKM TERINTEGRASI.“ KEMENKOPUKM. https://satudata.kemenkopukm.go.id/kumkm_dashboard/ (accessed Jan. 20, 2023)

[3] Hsiao-Ting Tseng, ”Customer-centered data power: Sensing and responding capability in big data analytics“, Journal of Business Research, Volume 158, March 2023, 113689. [Online serial]. Available: https://doi.org/10.1016/j.jbusres.2023.113689. [Accessed Nov 10, 2024]

[4] A. Perišić, D. Jung, and M. Pahor, “Churn in the Mobile Gaming Field: Establishing Churn Definitions and Measuring Classification Similarities,” Expert Systems with Applications, vol. 191, April, 2022. [Online serial]. Available: https://www.sciencedirect.com/science/article/pii/S09574174210 15852. [Accessed Jan. 20, 2023]

[5] A. Wicaksono, A. Anita, and T. Padilah, “Uji Performa Teknik Klasifikasi untuk Memprediksi Customer Churn,” Bianglala Informatika, vol. 9, no. 1, 2021. [Online serial]. Available: https://ejournal.bsi.ac.id/ejurnal/index.php/Bianglala/article/view/9992. [Accessed Jan. 21, 2023]

[6] I. Kabasakal, “Customer Segmentation Based On Recency Frequency Monetary Model: A Case Study in E-Retailing,” Bilişim Teknolojileri Dergisi, vol. 13, no. 1, January, 2020. [Online serial]. Available: https://www.researchgate.net/publication/338961311_Customer_Segmentation_Based_On_Recency_Frequency_Monetary_Model_A_Case_Study_in_E-Retailing. [Accessed Jan. 21, 2023]

[7] K. Liu, *Supply Chain Analytics: Concepts, Techniques and Applications*, 1st ed. Swiss: Palgrave Macmillan Cham, 2022. [E-book] Available: https://doi.org/10.1007/978-3-030-92224-5.

[8] T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August, 2016. [Online serial]. Available: https://doi.org/10.1145/2939672.2939785. [Accessed Jan. 26, 2023]

[9] M. Herawati, I. Wibowo, and I. Mukhlash, “Prediksi Customer Churn Menggunakan Algoritma Fuzzy Iterative Dichotomiser 3,” Limits: Journal of Mathematics and Its Applications, vol. 13, no. 1, 2016. [Online serial]. Available: http://dx.doi.org/10.12962/j1829605X.v13i1.1913. [Accessed Feb. 07, 2023]

[10] S. Shrestha and A. Shakya, “A Customer Churn Prediction Model using XGBoost for the Telecommunication Industry in Nepal,” Procedia Computer Science, vol. 215, 2022. [Online serial]. Available: https://doi.org/10.1016/j.procs.2022.12.067. [Accessed Jan. 16, 2023]

[11] C. Mena, A. Caigny, K. Coussement, K. Bock, and S. Lessmann, “Churn prediction with sequential data and deep neural networks. a comparative analysis,” Computer Science arXiv, September, 2019. [Online serial]. Available: https://doi.org/10.48550/arXiv.1909.11114

[12] H. Zhang and W. Zhang, “Application of GWO-attention-ConvLSTM model in customer churn prediction and satisfaction analysis in customer relationship management”. Heliyon, September 4, 2024 [Online serial]. Available: https://doi.org/10.1016/j.heliyon.2024.e37229 [Accessed Nov 20, 2024]

[13] S.S. Poudel, S. Pokharel, M. Timilsina, ”Explaining customer churn prediction in telecom industry using tabular machine learning models”, Machine Learning with Applications, June 24, 2024 [online serial]. Available: https://doi.org/10.1016/j.mlwa.2024.100567 [Accessed Nov 15, 2024]

[14] F.E. Usman-Hamza, L.F. Capretz , A.O. Balogun , H.A. Mojeed , R.T. Amosa , S. A. Salihu, A.G. Akintola, and M.A. Mabayoje, "Sampling-based novel heterogeneous multi-layer stacking ensemble method for telecom customer churn prediction ", Scientific African, May 3, 2024 [Online serial]. Available: https://doi.org/10.1016/j.sciaf.2024.e02223 [Accessed No 18, 2024]

[15] G. Meiselwitz, *Social Computing and Social Media. Communication and Social Communities*, 1st ed. Jerman: Springer Cham, 2019. [E-book] Available: https://doi.org/10.1007/978-3-030-21905-5.

[16] S. Yulianti, O. Soesanto, and Y. Sukmawaty, "Penerapan Metode Extreme Gradient Boosting (XGBOOST) pada Klasifikasi Nasabah Kartu Kredit," Journal of Mathematics: Theory and Applications, 2022. [Online serial]. Available: https://doi.org/10.31605/jomta.v4i1.1792. [Accessed Jan. 18, 2023]

[17] V. Bewick, L. Cheek, J. Ball, "Statistics review 14: Logistic regression," Critical care, vol. 9, no. 112, January, 2005. [Online serial]. Available: https://doi.org/10.1186/cc3045. [Accessed May. 24, 2023]

[18] B. Pratiwi, A. Handayani, and S. Sarjana, "Pengukuran Kinerja Sistem Kualitas Udara Dengan Teknologi Wsn Menggunakan Confusion Matrix," Jurnal Informatika Upgris, vol. 6, no. 2, December, 2020. [Online serial]. Available: https://doi.org/10.26877/jiu.v6i2.6552. [Accessed Jan. 26, 2023]