JURNAL ILMIAH KOMPUTER GRAFIS, Vol.17, No.1, Juli 2024, pp. 168-178
p-ISSN : 1979-0414(print), e-ISSN : 2621-6256 (online)
DOI : https://doi.org/10.51903/pixel.v17i1.2006

The Implementation of a Logistic Regression Algorithm and Gradient Boosting Classifier for Predicting Telco Customer Churn

Angga Adiansya[1], Zaenal Abidin[2]
angga.240399@students.unnes.ac.id[1]
Universitas Negeri Semarang
Teknik Informatika
Kampus UNNES Sekaran Gunungpati Kota Semarang

ARTICLE INFO

Received 23 April 2024
Accepted 27 Mei 2024
Published 24 Juli 2024

ABSTRACT

This research aims to predict customer churn in a telecommunications company using Logistic Regression (LR) and Gradient Boosting Classifier (GBC) algorithms. Customer churn poses a significant challenge as acquiring new customers is costlier than retaining existing ones. The dataset from Kaggle comprises 7043 records and 21 attributes. The process includes data pre-processing, cleaning, transformation, and normalization using a Min-Max Scaler. The data is split into features (X) and target (y), then divided into training and testing sets with an 80:20 ratio. Both models were trained and evaluated using a confusion matrix. Results show that the GBC model outperforms the LR model, with an accuracy of 83% compared to LR's 81%. This study demonstrates the effectiveness of GBC in predicting customer churn.

Keywords:* *Customer Churn, Logistic Regression, Gradient Boosting Classifier.

1. INTRODUCTION

As time progresses, technology has advanced significantly, particularly in telecommunications. The telecommunications sector has become one of the most vital aspects of this era [1]. The competition in the telecommunications sector is becoming increasingly fierce, so companies must retain their existing customers to prevent them from switching to competitors [2]. Customers are valuable assets. Customers, often referred to as clients, can easily switch to competitors if dissatisfied with the service. Such customers are referred to as churned customers [3]. Customer churn, or customer attrition, is detrimental for telecommunications companies, considering the high cost of customer acquisition and the negative impact on long-term revenue [4]. The telecommunications industry experiences an average churn rate of over 30% [5]. Meanwhile, acquiring a new customer costs 5-10 times more than retaining an existing one [6]. Therefore, addressing churn is a primary focus and concern for telecommunications companies, as it can significantly impact their revenue and business sustainability [7]. In this increasingly competitive context, retaining existing customers has become crucial for telecommunications companies to ensure sustainable growth [8]. However, the main challenge in understanding and addressing churn is the complexity of the factors influencing customers' decisions to switch [9].

However, the main challenge in understanding and addressing churn is the complexity of the factors influencing customers' decisions to switch to learning and make predictions by identifying patterns that must be explicitly visible in computer programs [5]. Two algorithms that can be utilized are Logistic Regression (LR) and Gradient Boosting Classifier (GBC). LR is chosen for its ability to provide clear interpretations of the factors influencing churn decisions [10], At the same time, GBC is selected for its capability to handle complex models and improve prediction accuracy [11].

The primary objective of this study is to develop and compare the performance of both methods in the context of the telecommunications industry. This research will focus on the implementation of LR and GBC algorithms to predict customer churn and compare the accuracy levels of these two algorithms [12],[6].

2. LITERATURE REVIEW

In telecommunications, customer churn prediction has become a widely researched topic. Several previous studies have explored the use of various machine-learning techniques to predict customer churn [13],[14]. Different machine learning models applicable for predicting customer churn include Support Vector Machines, Decision Trees, Regression Models, Neural Networks, Clustering, and Bayesian Models [13].

Geetha, et al. [15] propose using a random forest classifier and support vector machine algorithms to predict customer churn. This approach involves 21 customer activity attributes, which are then input into the algorithm for churn prediction. This research emphasizes the importance of data collection, preprocessing, and classifying raw data into churn and non-churn customers to facilitate effective churn prediction in the telecommunications sector.

Li and Zhou [16] propose a user segmentation and piecewise regression approach to identify relevant features and build different churn prediction models for each customer segment. Additionally, this study discusses the challenges faced in customer retention strategies, emphasizing the importance of identifying vulnerable customer groups and implementing effective retention measures.

Kavitha, et al. [1] Several machine learning methods, including the random forest algorithm, logistic regression, and XGBoost, were compared for customer churn prediction and classification. This research emphasizes the importance of feature selection and engineering in enhancing classification performance, ensuring the selection of relevant variables for accurate prediction.

3. METHODS

In this study, two machine learning algorithms are used to predict customer churn in telecommunications companies: LR and GBC. The research focuses on comparing the classification accuracy of LR and GBC. The process begins with inputting the dataset, performing data preprocessing, splitting the data into training and testing sets, and then classifying using LR and GBC. The stages of this process are illustrated in Figure .

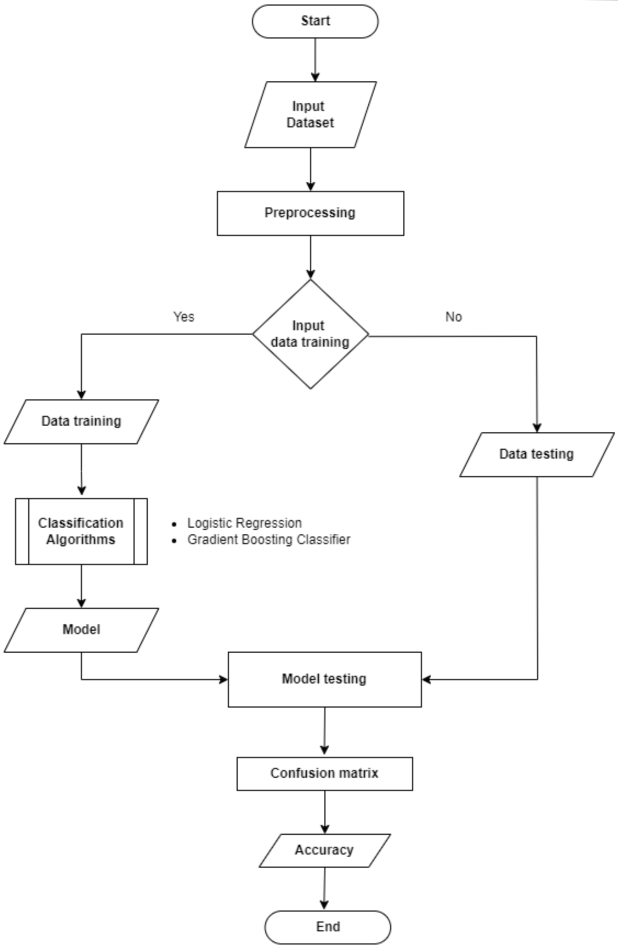

Figure 1. Research model

3.1 Dataset

In this study, the dataset is sourced from Kaggle Datasets. The dataset can be downloaded from the link https://www.kaggle.com/datasets/blastchar/telco-customer-churn, containing a total of 7043 records and 21 attributes. A sample of the dataset is shown in Table 1.

Table 1. Description of each data attribute.

Variable	Description	Type Data
Customerid	A unique identifier for each customer.	Categorical
Gender	The gender of the customer	Categorical
Seniorcitizen	Whether the customer is a senior citizen.	Numerical
Partner	Whether the customer has a partner.	Categorical
Dependents	Whether the customer has dependents.	Categorical
Tenure	The number of months the customer has been with the company.	Numerical
Phoneservice	Whether the customer has phone service.	Categorical
Multiplelines	Whether the customer has multiple lines.	Categorical
Internet service	The type of internet service the customer has.	Categorical
Online Security	Whether the customer has online security.	Categorical
Online backup	Whether the customer has online backup.	Categorical

Device protection	Whether the customer has device protection.	Categorical
Tech support	Whether the customer has tech support.	Categorical
Streamingtv	Whether the customer has streaming TV.	Categorical
Streamingmovies	Whether the customer has streaming movies.	Categorical
Contract	The type of contract the customer has.	Categorical
Paperlessbilling	Whether the customer has paperless billing.	Categorical
Paymentmethod	The method of payment the customer uses.	Categorical
Monthlycharges	The monthly charges for the customer's service.	Numerical
Totalcharges	The total charges incurred by the customer.	Categorical
Churn	Whether the customer has churned.	Categorical

3.2 Pre-processing Data

Data pre-processing is crucial in data analysis and machine learning model development, as data quality directly affects prediction outcomes. Data pre-processing includes data cleaning, variable transformation, and splitting the data into appropriate subsets for model training and testing.

1. Data cleansing
   The initial stage in data pre-processing is data cleaning, which involves identifying and handling missing values and removing duplicate data [17]. In the Telco Customer Churn dataset, the 'customerID' column is removed as it is irrelevant to the churn prediction analysis. Additionally, missing values in the 'TotalCharges' column are filled with the median of that column to avoid bias in the model.
2. Data transformation
   Converting categorical data into numerical form, categorical variables must be converted into numeric form [18]. This is done by encoding categorical variables into dummy variables. For example, the 'gender' column is converted into 0 and 1 to represent Female and Male categories. Similarly, other variables such as 'Partner', 'Dependents', and 'PhoneService' are converted into numeric values. Categorical variables with more than two categories, such as 'InternetService' and 'PaymentMethod', are transformed into dummy variables using one-hot encoding techniques.
3. Data Normalization
   Data normalization is adjusting the scale of features to be within the same range, typically from 0 to 1, to improve the performance and convergence of machine learning models [19].
4. Data Split
   The final step in pre-processing is splitting the data into training and testing subsets. The data is divided into an 80:20 ratio, where 80% of the data is used to train the model, and the remaining 20% is used to test the model. The data is randomly split to ensure that each subset is representative of the overall data distribution [20].

3.3 Classification Algorithm

Classification is an essential part of machine learning that uses classification to organize data into specific categories or classes. This technique benefits various applications, from pattern recognition to customer churn prediction. In this stage, two classification algorithms are employed: LR and GBC, each with advantages.

3.4 Classification using logistic regression

The logistic regression algorithm predicts the probability of a binary event (two classes). This model is well-suited for cases where the dependent variable is dichotomous, such as customer churn prediction. LR employs the logit function to link independent variables to the class probability. The advantages of LR include easy interpretability and relatively fast computational efficiency [10, 17]. using the logistic function logistic regression algorithm formula in the Equation (1 .

$$Logit[p(x)] = \frac{1}{1 + e^{-(\beta 0 + \beta 1 X1 + \beta 2 x2 + \cdots . + \beta k x k)}} \tag{1}$$

1. Probability calculation
   The logistic regression model determines the likelihood of a binary result. $Logit[p(x)]$ using the logistic function.
2. Rounding to a binary value
   The calculated probabilities are then rounded to a binary value of 0 or 1 based on a threshold of 0.5 [21]. If $Logit[p(x)] > 0.5$, the prediction is 1. While $Logit[p(x)] < 0.5$, then the prediction is 0.

3.5 Classification using gradient boosting classifier

The gradient boosting classifier is an ensemble technique that combines multiple weak models, such as decision trees, to form a robust predictive model. This algorithm works by sequentially adding new models to correct the errors of the previous models. GBC is known for its high accuracy and flexibility in handling complex data. However, this algorithm also requires longer training times and is prone to overfitting if the parameters are not carefully tuned [22]. gradient boosting classifier algorithm formula in Equation (2.

$$F_m(x) = F_{m^{-1}}(x) + n.h_m(x) \tag{2}$$

1. The first model
   $F_m(x)$ is initialized by predicting the target's mean value or mode. At each iteration m a new model $h_m(x)$ is added to correct errors from the previous model.
2. New model fittings
   New model $h_m(x)$ It was built to minimize the remaining loss from the previous model by calculating the gradient.
3. Final Prediction
   A combination of all the weak models was added during the iteration.

$$F_m(x) = F_0(x) + n \sum_{m=1}^{M} h_m(x) \tag{3}$$

3.5 Model Testing

Model evaluation is a crucial step in the machine learning process. The confusion matrix is an essential evaluation tool in analyzing classification models, used to assess the accuracy of the model's predictions on the test data. This matrix provides details on the number of correct and incorrect predictions for each class, consisting of four main components: True Positive (TP), True

Negative (TN), False Positive (FP), and False Negative (FN) [23]. The accuracy calculation uses Equation (4.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

TP and TN indicate the correct predictions for the positive and negative classes. At the same time, FP and FN show the number of incorrect predictions for the positive and negative classes. A confusion matrix allows for calculating various evaluation metrics such as accuracy, precision, recall, and F1-score, providing a comprehensive insight into the model's performance [24].

4. RESULTS AND DISCUSSION

This research uses LR and GBC algorithms to predict customer churn on the Telco Customer Churn dataset from Kaggle. After data collection, the next step is data preprocessing, which includes cleaning, variable transformation, and splitting the data into training and testing subsets. Then, the algorithm models are evaluated using a confusion matrix. In this study, the variable ‘churn’ is the primary target for prediction. Meanwhile, other attributes such as ‘gender,’ ‘tenure,’ ‘partner,’ ‘total charges, and others will be used as predictors. The dataset retrieval process can be seen in Figure.

	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	...
488	4472-LVYGI	Female	0	Yes	Yes	0	No	No phone service	DSL	Yes	...
753	3115-CZMZD	Male	0	No	Yes	0	Yes	No	No	No internet service	...
936	5709-LVOEQ	Female	0	Yes	Yes	0	Yes	No	DSL	Yes	...
1082	4367-NUYAO	Male	0	Yes	Yes	0	Yes	Yes	No	No internet service	...
1340	1371-DWPAZ	Female	0	Yes	Yes	0	No	No phone service	DSL	Yes	...
3331	7644-OMVMY	Male	0	Yes	Yes	0	Yes	No	No	No internet service	...
3826	3213-VVOLG	Male	0	Yes	Yes	0	Yes	Yes	No	No internet service	...
4380	2520-SGTTA	Female	0	Yes	Yes	0	Yes	No	No	No internet service	...
5218	2923-ARZLG	Male	0	Yes	Yes	0	Yes	No	No	No internet service	...
6670	4075-WKNIU	Female	0	Yes	Yes	0	Yes	Yes	DSL	No	...
6754	2775-SEFEE	Male	0	No	Yes	0	Yes	Yes	DSL	Yes	...

11 rows × 21 columns

Figure 2. Upload dataset Telco

The next step is the data pre-processing stage, where the initial process involves data cleaning. Irrelevant data that does not contribute to the churn prediction analysis can be removed. An example of the cleaned dataset can be seen in Figure.

	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity
0	Female	0	Yes	No	1	No	No phone service	DSL	No
1	Male	0	No	No	34	Yes	No	DSL	Yes
2	Male	0	No	No	2	Yes	No	DSL	Yes
3	Male	0	No	No	45	No	No phone service	DSL	Yes
4	Female	0	No	No	2	Yes	No	Fiber optic	No
...	...	...	...	...	...	...	...	...	...
7038	Male	0	Yes	Yes	24	Yes	Yes	DSL	Yes
7039	Female	0	Yes	Yes	72	Yes	Yes	Fiber optic	No
7040	Female	0	Yes	Yes	11	No	No phone service	DSL	Yes
7041	Male	1	Yes	No	4	Yes	Yes	Fiber optic	No
7042	Male	0	No	No	66	Yes	No	Fiber optic	Yes

7043 rows × 20 columns

Figure 3. The cleaned dataset

The next step is exploratory data analysis of the target 'churn' numerical and categorical variables. This helps in understanding data distribution, identifying patterns and factors influencing churn, and discovering correlations between variables. Correlations between numerical and categorical variables can be seen in

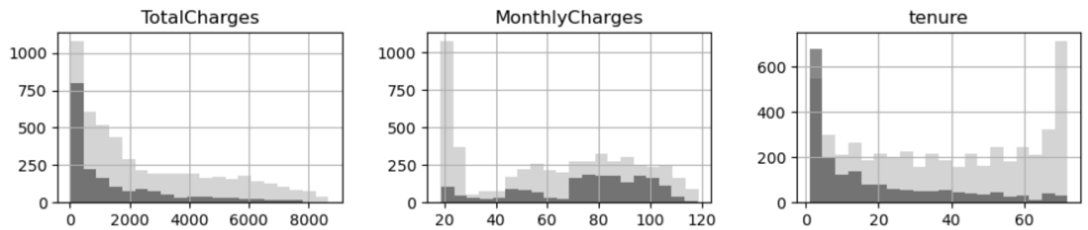

Figure 4 and Figure 5.

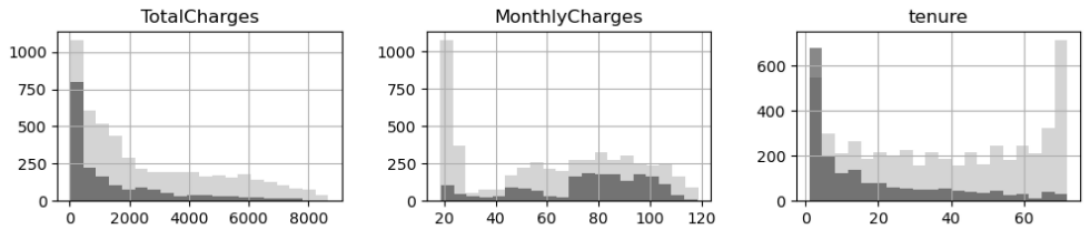

Figure 4. Correlation of numerical variables with targets

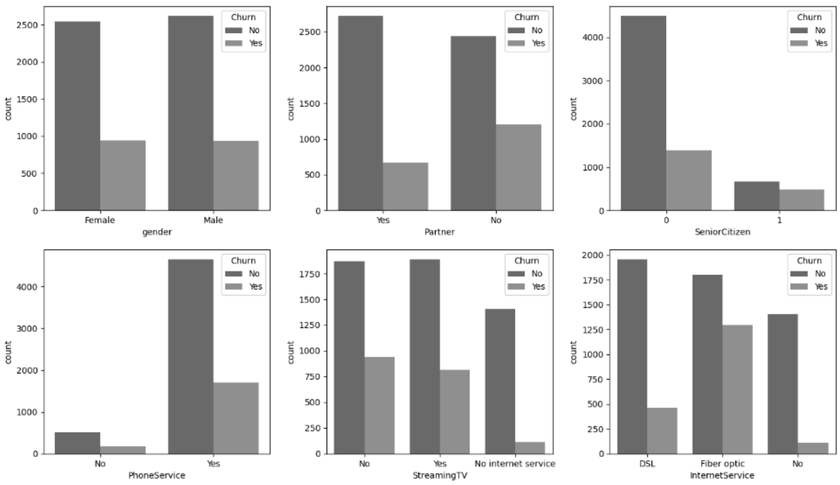

Figure 5. Correlation of categorical variables with targets

The next stage is data transformation, where categorical data is converted into numeric types. This process facilitates handling the dataset when used in the learning model. An example of the processed data can be seen in Table 2.

Table 2. Data transformation

Atribut	Value	Transformation
Gender	Female	0
	Male	1
Partner	Yes	1
	No	0
Dependents	Yes	1
	No	0

Next, data normalization is performed using a min-max scaler to standardize the values by mapping the data to a range of 0-1. The results of the normalization are shown in Table 3.

Table 3. The dataset after normalization.

No	Gender	Tenure	MonthlyCharges	TotalCharges	..
0	1	0.000000	0.115423	0.001275	..
1	0	0.464789	0.385075	0.215867	..
2	0	0.014085	0.354229	0.010310	..
…	…	…	…	…	..
7042	0	0.915493	0.869652	0.787641	..

The next step involves splitting the data into testing and training data an 80:20 ratio. The first step before beginning the classification model process is to separate the preprocessed data into two parts: features (X) and target (y). The "Churn" variable is placed in the target (y), which is the variable to be predicted or classified, while the other variables are included in the features (X). The subsequent step is to build classification models using the LR and GBC algorithms and compare the accuracy results of both algorithms.

The classification results using the LR and GBC algorithms are evaluated with a confusion matrix, which shows the number of correct and incorrect predictions for each class. The classification results based on the confusion matrix can be seen in Table 4 and
Table 5.

Table 4. Confusion matrix LR

	No churn	Churn
No churn	3707	423
Churn	661	834

Based on the confusion matrix calculations shown in Table 4, the computations using Equation (4 Yield the following results.

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

$$= \frac{3707 + 834}{3707 + 661 + 834 + 423} \times 100\% = 81\%$$

Table 5. Confusion matrix GBC

	No churn	Churn
No churn	3784	346
Churn	619	876

Based on the confusion matrix calculations shown in
Table 5, the computations using Equation (4 yield the following results.

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

$$= \frac{3784 + 876}{3784 + 619 + 876 + 346} \times 100\% = 83\%$$

This study focuses on comparing the classification accuracy of LR and GBC. The comparison of classification results between the two algorithms can be seen in Table 6.

Table 6. Comparison of LR and GBC Algorithm Accuracy.

Algorithm	Accuracy
Logistic regression	81%
Gradient boosting classifier	83%

Based on Table 6, The evaluation results show that the GBC model outperforms the LR model. The confusion matrix for GBC indicates an accuracy of 83%, highlighting its effectiveness in identifying potential churn customers.

4. Conclusion

This study aims to develop and compare the performance of two machine learning algorithms, LR and GBC, in predicting customer churn in the telecommunications industry. The research results indicate that both models have significant capabilities in predicting churn, with LR showing an accuracy of 81% and GBC demonstrating an accuracy of 83%. LR offers ease of interpretability and computational efficiency, while GBC exhibits high accuracy and can handle complex data. Model evaluation using metrics provides a comprehensive view of the models' performance.

The results of this study demonstrate that the appropriate combination of machine learning techniques can provide effective solutions to customer churn, a significant challenge for telecommunications companies. The study also highlights the importance of the data pre-processing stage and the selection of appropriate evaluation metrics to achieve accurate results

References

[1] V. Kavitha, G. H. Kumar, S. M. Kumar, and M. Harish, "Churn prediction of customer in telecom industry using machine learning algorithms," *International Journal of Engineering Research & Technology (2278-0181),* vol. 9, no. 05, pp. 181-184, 2020.

[2] N. Sjarif, N. Azmi, H. Sarkan, S. Sam, and M. Osman, "Predicting churn: how multilayer perceptron method can help with customer retention in telecom industry," in *IOP Conference Series: Materials Science and Engineering*, 2020, vol. 864, no. 1: IOP Publishing, p. 012076.

[3] A. Amin, F. Al-Obeidat, B. Shah, A. Adnan, J. Loo, and S. Anwar, "Customer churn prediction in telecommunication industry using data certainty," *Journal of Business Research,* vol. 94, pp. 290-301, 2019.

[4] W. H. Khoh, Y. H. Pang, S. Y. Ooi, L.-Y.-K. Wang, and Q. W. Poh, "Predictive churn modeling for sustainable business in the telecommunication industry: optimized weighted ensemble machine learning," *Sustainability,* vol. 15, no. 11, p. 8631, 2023.

[5] A. K. Ahmad, A. Jafar, and K. Aljoumaa, "Customer churn prediction in telecom using machine learning in big data platform," *Journal of Big Data,* vol. 6, no. 1, pp. 1-24, 2019.

[6] I. Ullah, B. Raza, A. K. Malik, M. Imran, S. U. Islam, and S. W. Kim, "A churn prediction model using random forest: analysis of machine learning techniques for churn prediction and factor identification in telecom sector," *IEEE access,* vol. 7, pp. 60134-60149, 2019.

[7] N. I. Mohammad, S. A. Ismail, M. N. Kama, O. M. Yusop, and A. Azmi, "Customer churn prediction in telecommunication industry using machine learning classifiers," in *Proceedings of the 3rd international conference on vision, image and signal processing*, 2019, pp. 1-7.

[8] A. Manzoor, M. A. Qureshi, E. Kidney, and L. Longo, "A review on machine learning methods for customer churn prediction and recommendations for business practitioners," *IEEE Access*, 2024.

[9] T. Zhang, S. Moro, and R. F. Ramos, "A data-driven approach to improve customer churn prediction based on telecom customer segmentation," *Future Internet*, vol. 14, no. 3, p. 94, 2022.

[10] M. Günay and T. Ensarı, "Predictive churn analysis with machine learning methods," in *2018 26th Signal Processing and Communications Applications Conference (SIU)*, 2018: IEEE, pp. 1-4.

[11] K. Ebrah and S. Elnasir, "Churn prediction using machine learning and recommendations plans for telecoms," *Journal of Computer and Communications,* vol. 7, no. 11, pp. 33-53, 2019.

[12] M. R. Ismail, M. K. Awang, M. N. A. Rahman, and M. Makhtar, "A multi-layer perceptron approach for customer churn prediction," *International Journal of Multimedia and Ubiquitous Engineering,* vol. 10, no. 7, pp. 213-222, 2015.

[13] A. Bhattarai, E. Shrestha, and R. P. Sapkota, "Customer churn prediction for imbalanced class distribution of data in business sector," *Journal of Advanced College of Engineering and Management,* vol. 5, pp. 101-110, 2019.

[14] B. Huang, M. T. Kechadi, and B. Buckley, "Customer churn prediction in telecommunications," *Expert Systems with Applications,* vol. 39, no. 1, pp. 1414-1425, 2012.

[15] V. Geetha, A. Punitha, A. Nandhini, T. Nandhini, S. Shakila, and R. Sushmitha, "Customer churn prediction in telecommunication industry using random forest classifier," in *2020 International Conference on System, Computation, Automation and Networking (ICSCAN),* 2020: IEEE, pp. 1-5.

[16] W. Li and C. Zhou, "Customer churn prediction in telecom using big data analytics," in *IOP Conference Series: Materials Science and Engineering*, 2020, vol. 768, no. 5: IOP Publishing, p. 052070.

[17] A. F. Dewi and R. Pratiwi, "Analisis regresi logistik biner pada pengaruh harga, kualitas pelayanan dan promosi terhadap kepuasan pelanggan dalam menggunakan jasa layanan grab di kabupaten lamongan," *Inferensi,* vol. 4, no. 2, pp. 77-84, 2021.

[18] M. R. Khan, J. Manoj, A. Singh, and J. Blumenstock, "Behavioral modeling for churn prediction: Early indicators and accurate predictors of custom defection and loyalty," in *2015 IEEE International Congress on Big Data,* 2015: IEEE, pp. 677-680.

[19] A. Chouiekh, "Deep convolutional neural networks for customer churn prediction analysis," *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI),* vol. 14, no. 1, pp. 1-16, 2020.

[20] M. Y. Matdoan, "Pemodelan regresi robust least trimmed square (LTS)(studi kasus: faktor-faktor yang mempengaruhi penyebaran penyakit malaria di indonesia)," *Euclid,* vol. 7, no. 2, pp. 77-85, 2020.

[21] D. W. Hosmer Jr, S. Lemeshow, and R. X. Sturdivant, *Applied logistic regression*. John Wiley & Sons, 2013.

[22] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of statistics,* pp. 1189-1232, 2001.

[23] L.-W. Wei, C.-M. Huang, H. Chen, C.-T. Lee, C.-C. Chi, and C.-L. Chiu, "Adopting the I 3–R 24 rainfall index and landslide susceptibility for the establishment of an early warning model for rainfall-induced shallow landslides," *Natural Hazards and Earth System Sciences,* vol. 18, no. 6, pp. 1717-1733, 2018.

[24] A. Tharwat, "Classification assessment methods," *Applied computing and informatics,* vol. 17, no. 1, pp. 168-192, 2021.