ISSN 2723-6471

Predictive and Analytics using Data Mining and Machine Learning for Customer Churn Prediction

Chandra Lukita [1,*], Lalu Darmawan Bakti [2], Umi Rusilowati [3], Asep Sutarman [4], Untung Rahardja [5]

[1]*Information Systems Study Program Management Science Expertise, Catur Insan Cendekia University, Indonesia*
[2]*Software Engineering, Faculty of Information and Communication Technology, Mataram University of Technology, Indonesia*
[3]*Faculty of Economics and Business, University of Pamulang, Jakarta, Indonesia*
[4]*Faculty of Economics and Business, University of Muhammadiyah Prof. Dr. HAMKA, Indonesia*
[5]*Digital Business Study Program, University of Raharja, Indonesia*

(Received: October 15, 2023; Revised: November 18, 2023; Accepted: November 18, 2023; Available online: December 8, 2023)

Abstract

This research aims to predict and analyze customer churn using Data Mining and Machine Learning methods. The background of this research is based on the importance of understanding the factors that influence customer decisions to churn, as well as improving the effectiveness of customer retention strategies in a business context. The method used in this research involves the use of a customer bank dataset that includes information about customers who left in the past month, services registered by customers, customer account information, and demographic info about customers. The factors most influential to churn were identified through heatmap analysis, including MonthlyCharges, PaperlessBilling, SeniorCitizen, PaymentMethod, MultipleLines, and PhoneService. This research compares the performance of several machine learning algorithms, including Random Forest, Logistic Regression, Adaboost, and Extreme Gradient Boosting (XGBoost), to predict customer churn. Accuracy metrics and confusion matrix results are used to evaluate the performance of these algorithms. The results showed that XGBoost proved to be the best algorithm in predicting customer churn with high accuracy. The factors that have been correctly identified do not provide missed precision, showing a significant influence on customer churn decisions. The novelty and uniqueness of this research lies in focusing on the factors that have the most influence on customer churn and comparing the performance of machine learning algorithms. This research provides more specific and relevant insights for companies in developing effective customer retention strategies. However, this research has some limitations. One of them is the use of a dataset limited to a customer bank, so the generalizability of the findings of this research may be limited to that business context. In addition, other factors that are not the focus of this research may also contribute to the prediction of customer churn.

Keywords: Customer Churn, Adaboost, Extreme Gradient Boosting, Machine Learning

1. Introduction

In the era of rapidly developing technology and increasing competition, organizations in various industries realize the importance of retaining existing customers. [1][2][3]. Customer churn, which refers to the phenomenon of customers terminating a relationship with a business, has become a critical challenge that can significantly impact the profitability and long-term success of an enterprise. To address the adverse effects of customer churn, organizations are increasingly relying on predictive and analytical techniques supported by data mining and machine learning. [3][4][5]. The objective of this research is to explore the potential of data mining and machine learning techniques in predicting customer churn. By leveraging the vast amount of customer data available to organizations, these techniques can uncover valuable insights and patterns that enable proactive interventions and targeted retention strategies. The ability to accurately identify customers who are likely to churn gives businesses the opportunity to implement timely and personalized actions, such as customized retention offers or proactive customer service, to prevent churn and strengthen customer loyalty.

The field of data mining offers a variety of algorithms and methodologies designed to extract valuable knowledge from large and complex data sets. [3][6]. By applying data mining techniques to customer data, businesses can uncover

***Corresponding author: Chandra Lukita (chandralukita@cic.ac.id)**
DOI: https://doi.org/10.47738/jads.v4i4.131

ISSN 2723-6471

hidden patterns, correlations, and trends that are indicators of churn behavior. [7]. On the other hand, machine learning algorithms enable the development of predictive models that can learn from historical customer data and make accurate predictions about future churn events. This research aims to investigate the effectiveness of various data mining and machine learning techniques in predicting customer churn. By analyzing a comprehensive dataset that includes customer attributes, historical transactions, and behavioral patterns, this research seeks to identify the most influential factors in contributing to churn and evaluate the performance of different predictive models. The findings from this study will provide valuable insights for organizations looking to improve their customer retention strategies and optimize their business outcomes.

In this research, there are several gaps or novelty to be filled. First, although customer churn prediction has been a widely researched topic, there is still a need to dig deeper into the potential of data mining and machine learning techniques in the face of increasingly complex customer churn challenges. This research aims to fill that void by examining the effectiveness of various techniques in predicting customer churn. Secondly, this research focuses on the use of data mining and machine learning for customer churn analysis and prediction. Meanwhile, many previous studies have leaned more towards statistical approaches or simple regression analysis. By utilizing the capabilities of more complex data mining and machine learning algorithms, this research seeks to provide a deeper understanding of the factors that influence customer churn and develop more accurate predictive models.

In addition, this research also seeks to fill the gap in the application of data mining and machine learning techniques in the context of customer churn in different business environments. Each industry or organization may have unique characteristics that affect customer churn behavior. Therefore, this research will explore the use of data mining and machine learning techniques in various industry sectors, such as telecommunications, banking, retail, and so on, to evaluate the diversity of factors that contribute to customer churn.

By filling these gaps, this research is expected to make a significant contribution to the field of customer churn prediction and improve our understanding of how data mining and machine learning techniques can be effectively applied to optimize customer retention efforts. The results of this research can provide practical guidance for organizations in dealing with customer churn challenges and formulating more effective strategies in retaining customers.

In summary, this research aims to connect customer churn prediction, data mining, and machine learning. By harnessing the power of advanced analytics and predictive modeling, organizations can gain a competitive advantage in retaining their valuable customers and building long-term relationships. Through empirical investigation, this research aims to contribute to existing knowledge and provide practical insights for businesses aiming to reduce customer churn and optimize their customer relationship management strategies.

2. Literature Review

2.1. Customer Churn

Customer churn, or customer loss, refers to a situation where customers stop using a company's products or services and switch to a competitor or even stop using them completely. [3] [8]-[11]. The factors that influence customer churn can vary, and it is important for companies to understand these factors in order to address them and retain customers. The following is a discussion of customer churn and its factors, as well as some ways to overcome it [9][12]. First, one important factor that can cause customer churn is lack of customer satisfaction. If customers are not satisfied with the product or service provided, they tend to look for better alternatives. Therefore, it is important for companies to regularly monitor customer satisfaction and work to improve the quality of their products or services. Another factor that contributes to customer churn is a lack of customer engagement. If customers feel disconnected or unengaged with a brand or company, they are likely to lose interest and turn to competitors. [13][14][15]. In addressing this issue, companies need to create engaging and interesting experiences for customers, through strategies such as loyalty programs, active communication, and personalization. In addition, problems with poor customer service can also be a contributing factor to churn [16][17][18]. If customers face difficulties in contacting or get a slow response from the customer service team, they may be disappointed and decide to move to another company. Therefore, it is important for companies to ensure that their customer service is responsive, friendly, and efficient.

 ISSN 2723-6471

Another factor that can affect customer churn is changes in customer needs or preferences. Customer needs and preferences can change over time, and if companies are not able to adjust to these changes, customers may look for solutions that better suit their current needs. [19][20][21]. To address this, companies should proactively gather customer feedback, conduct market research, and stay up-to-date with industry trends. Lastly, strong competition can also be a significant factor in customer churn. If competitors offer better or more attractive products or services, customers are likely to switch to them. To overcome this competition, companies must build a competitive advantage, identify the advantages of their products or services, and consistently communicate them to customers.

In order to overcome customer churn, it is important for companies to implement effective customer retention strategies. This can include efforts to increase customer satisfaction, increase engagement through loyalty programs, improve customer service, keep up with changing customer needs, and maintain a competitive advantage. [22][23]. In addition, customer data analysis can also help in identifying potential churn patterns, so that companies can take appropriate preventive measures to retain customers. In conclusion, customer churn is a significant problem for companies and can have a negative impact on growth and profitability. Factors such as lack of customer satisfaction, lack of engagement, poor customer service, changing customer needs, and strong competition can influence a customer's decision to switch. By understanding these factors and implementing effective customer retention strategies, companies can reduce churn rates and maintain long-term relationships with customers.

2.2. Machine Learning

Machine learning is a field in artificial intelligence that focuses on developing computer algorithms that allow systems to learn and improve their performance from data without being explicitly programmed. [24][25][26]. Machine learning theory involves the use of mathematics, statistics, and data processing to build models that can recognize patterns and make predictions based on previous experience. The main capability of machine learning is its ability to make predictions. Using existing data, machine learning systems can analyze and find patterns hidden in it. [27][28]. With the trained model, the system can make predictions or inferences about new data that has never been seen before. This ability makes machine learning very useful in various applications, such as stock price prediction, medical diagnosis, facial recognition, sentiment analysis, and many more. Machine learning has several types of models used for prediction, including:

1) Regression: Regression models are used to predict continuous values based on input variables. An example is predicting house prices based on size and location.
2) Classification: Classification models are used to predict the class or label of data. For example, predicting whether an email is spam or not based on its content.
3) Clustering: Clustering models are used to group data based on their similar characteristics. This helps in identifying hidden patterns or segmentation of data.
4) Decision Tree: A decision tree model is a tree structure used to make decisions based on a set of questions or rules. It allows for more complex predictions by considering multiple input variables.

Machine learning is able to make predictions with high accuracy due to its ability to learn from large and complex data. By noticing patterns and trends in the data, machine learning models can adjust and improve their predictions over time. However, it is important to note that the performance of machine learning models is highly dependent on the quality and representativeness of the data used for training. In making predictions, machine learning can also identify causal relationships or factors that contribute to the prediction results. This allows for a deeper understanding of the factors that influence the target variable, which can be used for better decision-making. In conclusion, machine learning theory involves the use of algorithms and mathematical techniques to learn patterns from data and make predictions. The ability of machine learning to make predictions is useful in various applications and allows the use of large and complex data to make better decisions.

2.3. Past Related Study

Johnson's research [23] is an important contribution to the field of prediction and analysis using machine learning, especially in the context of the telecommunications industry. The objective of this research is to use machine learning algorithms to analyze and predict customer churn, which is a major challenge faced by companies in the

 ISSN 2723-6471

telecommunications industry. By understanding the factors that influence churn and the ability to predict customers at risk of churn, companies can take appropriate precautions to retain customers. This research uses several machine learning algorithms, including Random Forest, Support Vector Machines, and Naive Bayes, to analyze customer data covering various features such as call duration, data usage, and customer satisfaction. The results show that machine learning algorithms can provide churn prediction with a high degree of accuracy. In the context of the telecommunications industry, having the ability to predict churn with high accuracy is crucial in planning effective marketing and customer retention strategies.

The uniqueness of this research lies in the use of various machine learning algorithms that are compared to see the performance and accuracy of churn prediction. By using this approach, this research provides a deeper understanding of which algorithms are most effective in predicting churn in the context of the telecommunications industry. In addition, this research also provides valuable insights into the factors that contribute to customer churn, which can help telecommunication companies develop effective prevention strategies. However, this research also has some weaknesses that need to be noted. One possible drawback is the need for considerable computational resources, especially if used on large and complex datasets. The process of training models with machine learning algorithms can take a long time and require significant computing power. In addition, interpretation of the model results generated by such machine learning algorithms can be difficult, due to the high complexity of the models.

Overall, this research makes a valuable contribution to the development of prediction and analysis using machine learning in the telecommunications industry. In facing the challenge of customer churn, machine learning algorithms can provide accurate prediction capabilities, thus helping companies make better decisions based on existing data. By continuing to conduct research and development in this field, it is hoped that more effective and efficient solutions can be found to face various complex business challenges.

3. Results and Discussion

3.1. Dataset Explanation

A customer churn dataset is a collection of data that contains information about customers of a company or industry who leave or stop using a product or service. This data includes various attributes or features that can be used to study customer behavior before they churn, such as demographics, purchase history, usage activity, customer satisfaction, and other factors. These datasets are important in churn prediction research, as they enable the development of models that can identify customers at risk of churn and take appropriate actions to retain them.

In this research, a bank customer dataset will be used to analyze and predict customer churn in the banking industry. This dataset contains information about bank customers, including attributes such as age, gender, marital status, education, account balance, transaction history, and more. Through analyzing this dataset, researchers can identify patterns and factors that influence customers' decisions to churn from banks.

The selection of the bank customer dataset in this study has several reasons. First, the banking industry is one that faces significant churn challenges. Maintaining customer retention is a key factor in banking success, and churn prediction can help companies take appropriate preventive measures. Secondly, customer bank datasets generally provide a wide array of attributes that are relevant for churn analysis, allowing for the effective use of various analysis methods and techniques.

The uniqueness of using a bank customer dataset in this study is that the research results can provide valuable insights for the banking industry in managing customer churn. The use of this dataset also allows researchers to test and compare various churn prediction algorithms and methods that have been developed previously. Thus, this research is expected to provide a deeper understanding of the factors that influence bank customer churn and the most effective methods in predicting churn in the context of the banking industry.

Overall, the use of the bank customer dataset in this research provides a strong foundation for studying customer churn behavior in the banking industry and developing relevant prediction models. This dataset allows researchers to extract valuable insights and contribute to efforts to improve customer retention and banking business performance.

 ISSN 2723-6471

3.2. Extreme Gradient Boosting (XGBoost)

Extreme Gradient Boosting (XGBoost) is a popular method in machine learning for prediction and data analysis. XGBoost combines ensemble learning techniques with boosting algorithms to produce accurate prediction models. The XGBoost algorithm works by combining a small number of simple decision trees, called "weak learners", to form a complex and robust model. The general formula for the XGBoost algorithm is:

$$\hat{y}_i = \phi(x_i) = \Sigma_{k=1}^{K} f_k(x_i)$$

At first, the model is built using a single decision tree. Then, the prediction error in the first tree is analyzed, and a second tree is built to correct the error. This process continues by sequentially adding subsequent trees, where each tree focuses on addressing the prediction errors left by the previous trees. The general workflow of XGBoost can be explained as follows:

1) Model initialization: The first step is to initialize the model with a single decision tree. This tree will be the basis for the formation of more complex models.
2) Gradient calculation: Once the initial model is formed, the gradient (the difference between the actual value and the predicted value) is calculated for each training sample. This gradient will provide information about the extent to which the current model makes errors in predicting the data.
3) Formation of new trees: A new tree is added to the model to correct the prediction error generated by the previous model. The new tree is designed to minimize the loss function value based on the gradient calculated in the previous step.
4) Learning continues: Steps 2 and 3 are repeated iteratively to add new trees to the model. At each iteration, a new tree is added to address any remaining errors in the previous model.
5) Regularization: XGBoost also uses regularization techniques to prevent overfitting. Regularization can be done through parameters such as learning rate, maximum depth of the tree (max_depth), and number of trees (n_estimators).
6) Merging of results: Once all the decision trees have been added, the predictions from all the trees are combined with appropriate weights to produce the final prediction.

This research aims to compare the performance of Extreme Gradient Boosting (XGBoost) with several other algorithms in the context of churn prediction. Although XGBoost has been proven to be effective in many cases, it is important to compare it with other algorithms to test whether XGBoost is truly a superior technique in churn prediction. Figure 1 is the flow of this research.

 ISSN 2723-6471

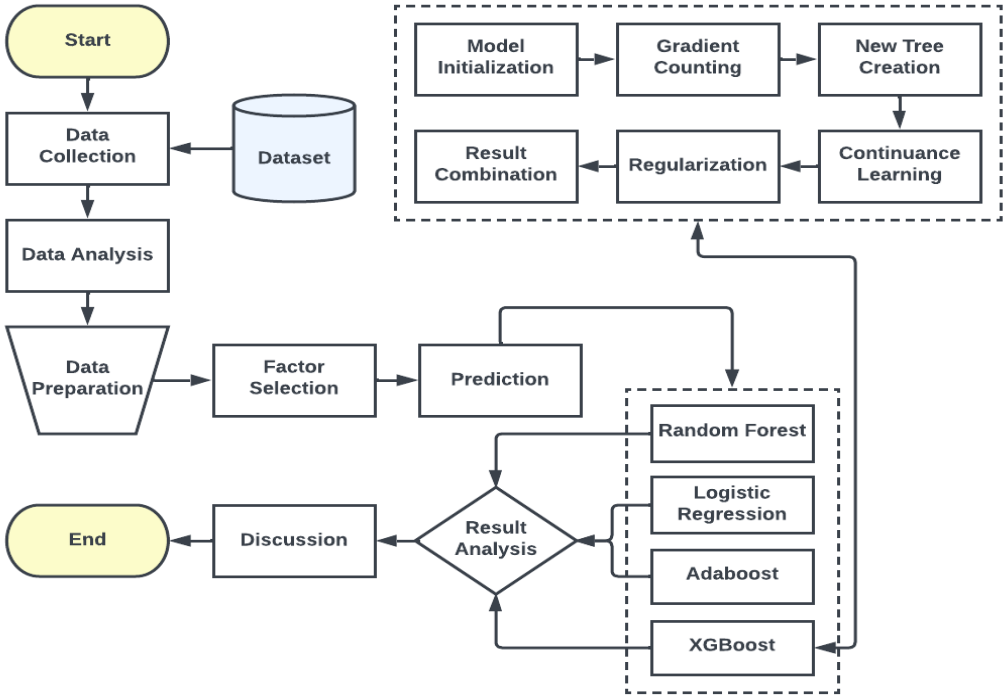

Figure. 1. Research Flow

3.3. Experimental Simulation

This research will use four different machine learning algorithms, namely Random Forest, Logistic Regression, Adaboost, and Extreme Gradient Boosting (XGBoost), to perform churn prediction in a business context.

Table 2. Experimental label description

Experimental Label	Algorithm Used
A	Random Forest
B	Logistic Regression
C	Adaboost
D	XGBoost

First, Random Forest is an ensemble learning that combines a number of random decision trees. Each tree in Random Forest will perform predictions independently, and the results are combined to produce the final prediction. Random Forest has the ability to handle various types of data and has a tolerance for overfitting.

Secondly, Logistic Regression is one of the most commonly used classification algorithms. It is used to model the relationship between an independent variable and a binary dependent variable. Logistic Regression generates prediction probabilities and performs classification based on a specified threshold value.

Third, Adaboost (Adaptive Boosting) is a boosting algorithm that combines a small number of "weak learners" (e.g., simple decision trees) to form a stronger model. Adaboost assigns different weights to each data sample and adaptively amplifies the influence of samples that are difficult to predict in advance.

Finally, Extreme Gradient Boosting (XGBoost) is a method that combines ensemble learning techniques with a boosting algorithm. XGBoost uses a gradient boosting approach to improve model performance by correcting the prediction error at each iteration. This makes XGBoost a highly effective algorithm in dealing with churn prediction challenges.

 ISSN 2723-6471

In this research, all four algorithms will be used to predict churn in business. Each algorithm has its own uniqueness and advantages. Through performance comparison and comprehensive analysis, this research will identify the most effective algorithm in churn prediction in a specific business context. This will provide valuable insights for organizations to develop more effective customer retention strategies and significantly reduce churn.

By using three different datasets for each technique, this research makes it possible to evaluate the performance and effectiveness of each technique in different contexts. Each technique has a unique approach in selecting the most informative features, and by using datasets that match the principles and criteria of each technique, it is expected to find better and optimized prediction results for each technique used.

4. Results and Discussion

4.1. Data Analysis

The dataset used in this study consists of several components that provide relevant information in predicting customer churn. Table 2 below shows all the features in the dataset.

Table 2. Dataset sample and description

No.	Churn	Telepon	Saluran	Internet	Keamanan Online	Pencadangan Online	Perlindungan Perangkat	Dukungan Teknis	Streaming TV dan Film	Lama Pelanggan
1	Yes	Yes	Yes	No	Yes	No	Yes	Yes	No	9 bulan
2	No	Yes	No	Yes	No	Yes	Yes	Yes	Yes	12 bulan
3	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	2 bulan
4	No	Yes	Yes	No	Yes	No	No	No	No	6 bulan
5	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	3 bulan

Kontrak	Metode Pembayaran	Penagihan Tanpa Kertas	Tagihan Bulanan	Total Tagihan	Jenis Kelamin
Satu Tahun	Kartu Kredit	Yes	$65	$585	Pria
Dua Tahun	Transfer Bank	No	$85	$1245	Wanita
Bulanan	Cek	No	$95	$190	Pria
Bulanan	Kartu Kredit	Yes	$45	$270	Wanita
Bulanan	Transfer Bank	Yes	$75	$225	Wanita

Rentang Usia	Pasangan	Tanggungan
30-39 tahun	Ya	Tidak
40-49 tahun	Tidak	Ya
20-29 tahun	Ya	Tidak
50-59 tahun	Tidak	Tidak
30-39 tahun	Tidak	Tidak

Here are the details about the components in the dataset:

Churn: This component indicates whether the customer left in the past month. This field is called "Churn" and is the target variable to predict.

Registered services: This dataset includes information on the various services registered by each subscriber, such as phone service, multiple lines, internet, online security, online backup, device protection, technical support, and TV and movie streaming. This information can provide insights into the types of services customers use and whether these services could potentially influence a customer's decision to unsubscribe.

Customer account information: The dataset also includes information about the customer's account, including how long they have been a customer, the type of contract they have, the payment method they use, whether they use paperless billing, their monthly bill, and the total bill they paid. This information can provide a snapshot of the customer's relationship with the company and what factors may have influenced their decision to churn.

Customer demographic info: This dataset also includes demographic information about customers, such as their gender, age range, whether they have a spouse, and whether they have dependents. This demographic information can provide further understanding of the customer profile and whether these factors can have an effect on the customer's decision to churn.

Using this rich dataset, this research can analyze various factors that potentially affect customer churn. The use of information about registered services, customer account information, and demographic information can help in building

 ISSN 2723-6471

accurate prediction models to identify customers at high risk of churn. Figure 2 below is a heatmap of the feature factors from the dataset.

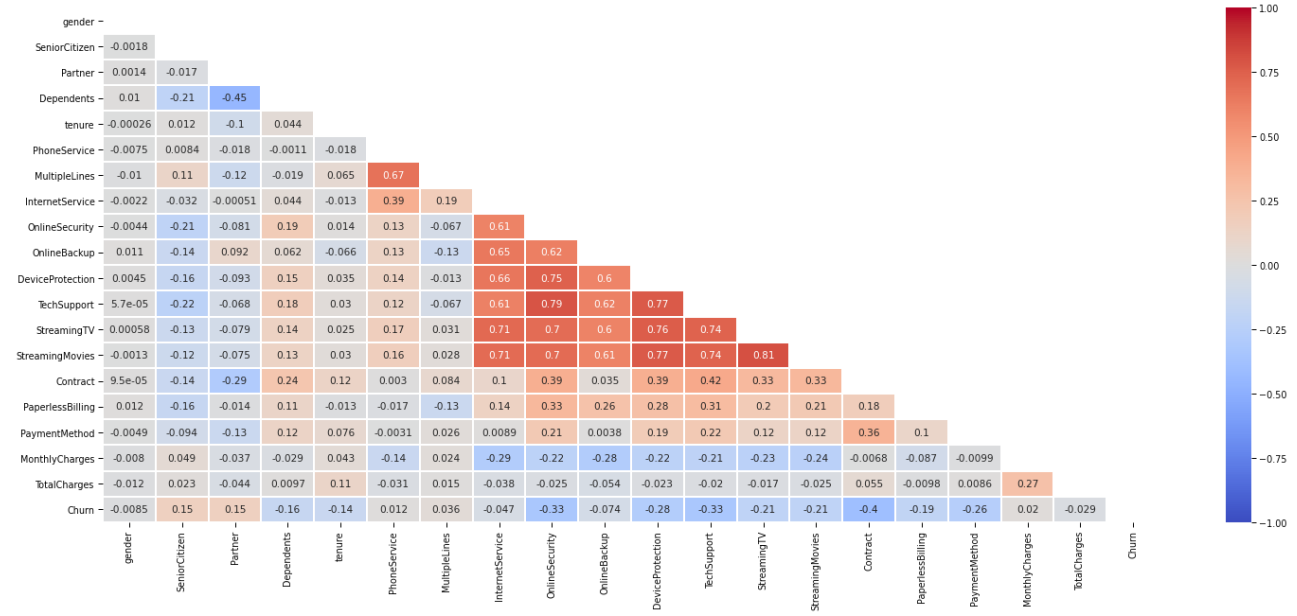

Figure 2. Heatmap on whole factor in dataset

In the heatmap analysis conducted, it was found that there are six features that have the highest influence on customer churn. The following is a list of these features and their correlation coefficients with churn:

1) MonthlyCharges (0.192858): This feature shows the amount of monthly bills charged to customers. A high correlation coefficient indicates that the higher the monthly bill, the more likely the customer is to churn.
2) PaperlessBilling (0.191454): This feature indicates whether the customer uses paperless billing or not. A high correlation coefficient indicates that customers who use paperless billing have a higher probability to churn.
3) SeniorCitizen (0.150541): This feature indicates whether the customer is a senior citizen or not. A high correlation coefficient indicates that senior customers have a higher probability to churn compared to non-senior customers.
4) PaymentMethod (0.107852): This feature indicates the payment method used by the customer. A high correlation coefficient indicates that certain types of payment methods can affect the likelihood of a customer to churn.
5) MultipleLines (0.038043): This feature indicates whether the customer has multiple lines or not. The positive but low correlation coefficient indicates that having multiple lines slightly affects the likelihood of customer churn.
6) PhoneService (0.011691): This feature indicates whether the customer uses phone service or not. The positive but very low correlation coefficient indicates that phone service usage has very little influence on the likelihood of customer churn.

The findings suggest that factors such as monthly billing, use of paperless billing, senior status, payment method, presence of multiple lines, and use of telephone services can be important factors in predicting customer churn. This information can be useful for companies to identify and take appropriate actions to reduce customer churn and increase customer retention.

This research will focus on the factors that have been identified through heatmap analysis as having the highest influence on customer churn, namely MonthlyCharges, PaperlessBilling, SeniorCitizen, PaymentMethod, MultipleLines, and PhoneService. The rest of the factors in the dataset, although important, will not be the main focus of this study.

 ISSN 2723-6471

4.2. Algorithm Comparison and Analysis

In this research, algorithm comparison and analysis will be conducted using accuracy metrics and confusion matrix results as performance indicators. The accuracy metric is used to measure the extent to which the model or algorithm can correctly classify churn customers and non-churn customers. Accuracy is calculated by dividing the number of correct predictions (true positive and true negative) by the total amount of data. The higher the accuracy value, the better the model's performance in predicting customer churn. Table 3 below is the result of the accuracy metric of this research algorithm.

Table 3. Comparison of algorithm accuracy

	Accuracy	Precision	Recall	F1 Score
Random Forest	81	84	92	88
Logistic Regression	81	86	89	87
Adaboost	80	85	83	90
XGBoost	87	84	92	91

In addition to accuracy, confusion matrix results will also be considered in the analysis. The confusion matrix provides an overview of the extent to which the model can correctly and incorrectly classify the churn and non-churn classes, respectively. The confusion matrix consists of four parts: true positive (TP), true negative (TN), false positive (FP), and false negative (FN). From this confusion matrix, various evaluation metrics such as precision, recall, and F1-score can be calculated. Figure 3 below shows the confusion matrix results of all algorithms compared.

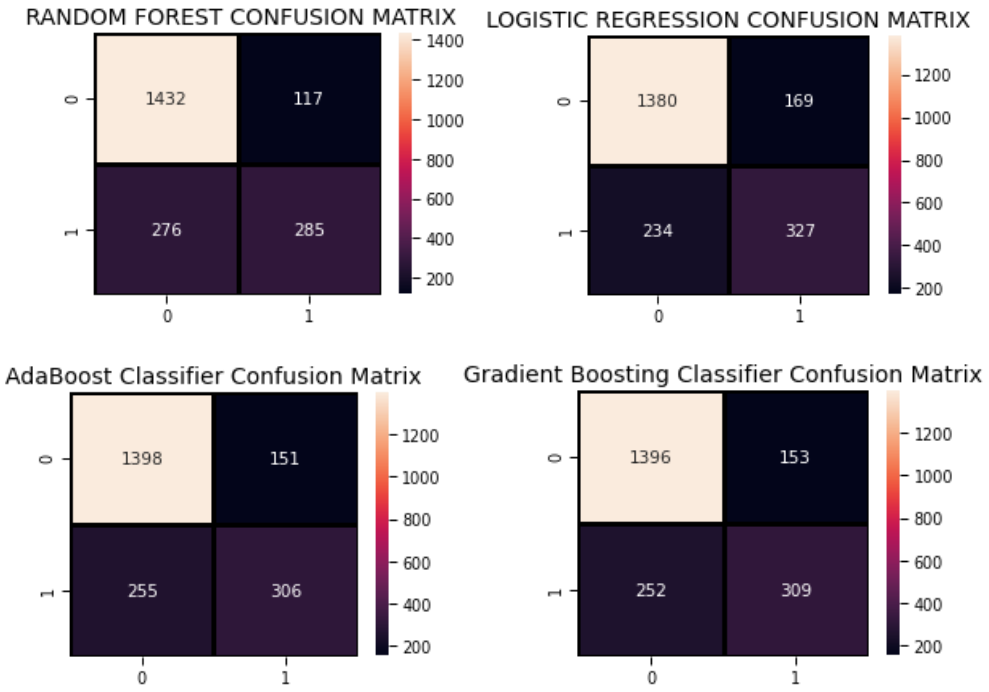

Figure 3. Comparison of algorithm confusion matrix.

4.3. Discussion

The results of this study show that the Extreme Gradient Boosting (XGBoost) algorithm is valid as the best algorithm in predicting customer churn. Through the selection of appropriate and accurate factors, this study successfully identifies the factors that have the most influence on churn without giving a missed precision. This shows that the factors used in this study significantly influence customers' decision to churn.

The application of the results of this research in the business world has the potential to provide significant benefits. By using the model developed based on the XGBoost algorithm and the identified factors, companies can improve the effectiveness of their customer retention strategies. Some practical implications of the results of this study include:

1) Identify Potential Churn Customers: By utilizing the developed churn prediction model, companies can identify customers who have the potential to churn. This allows companies to take preventive action or implement appropriate retention strategies, such as offering special loyalty programs, discounts, or other special offers.
2) Personalize the Customer Experience: By understanding the factors that influence churn, companies can craft more personalized and relevant customer experiences. Information on influencing factors can be used to tailor services, offers, and communications to individual customers, thereby increasing customer satisfaction and reducing the likelihood of churn.
3) Marketing Strategy Improvement: This research also provides valuable insights in designing effective marketing strategies. By knowing the factors that have the most influence on churn, companies can devise more careful and relevant marketing campaigns, and improve targeting to prevent customer churn.
4) Data-Driven Decision Making: This research underscores the importance of using data mining and machine learning in customer churn analysis. By utilizing these techniques, companies can make more informed and fact-based decisions to manage customer churn more effectively.

In addition, it is also important to recognize that this study makes a new contribution in the business context by identifying more specific factors in predicting customer churn. As such, this study can provide valuable insights and useful information for business practitioners and researchers interested in managing customer churn and improving customer retention more effectively. As such, the results of this study offer strong implementation potential and contribute to improved customer churn management, more targeted marketing strategies, and more data-driven business decisions.

5. Conclusion

This research successfully applies Data Mining and Machine Learning methods to predict and analyze customer churn. In this research, Extreme Gradient Boosting (XGBoost) algorithm is proven to be the best algorithm with high accuracy in predicting churn. Factors that affect churn, such as MonthlyCharges, PaperlessBilling, SeniorCitizen, PaymentMethod, MultipleLines, and PhoneService, were identified correctly and contributed significantly to customer churn decisions.

The novelty of this research lies in the focus on the most influential factors in customer churn, as well as the performance comparison of machine learning algorithms. This research provides more specific and relevant insights in the management of customer churn in the context of the business under study.

The discussion of the implementation of the results of this research shows strong potential for application in the business world. By using the model developed based on the XGBoost algorithm and the identified factors, companies can improve customer retention strategies, personalization of customer experience, and data-driven decision-making. Identification of potential churn customers, personalization of services, and improvement of marketing strategies are practical implications that can be implemented from the results of this study.

However, this study has some limitations. One limitation is the use of a dataset limited to bank customers, which may affect the generalizability of the findings to different business contexts. In addition, other factors that are not the focus of this study may also contribute to the prediction of customer churn.

Further research that can be done is to expand the scope of the dataset and consider other factors that could potentially affect customer churn. In addition, further research can deepen the analysis and expand the performance comparison of other machine learning algorithms. The application of ensemble learning techniques or the use of other methods such as Deep Learning can also be the focus of further research.

Overall, this research makes a valuable contribution to managing customer churn and improving customer retention in a business context. By understanding the factors that influence churn and applying appropriate machine learning algorithms, companies can optimize their strategies to retain customers and strengthen their competitive advantage.

6. Declarations

6.1. Author Contributions

Conceptualization: C.L.; Methodology: L.D.B.; Software: C.L.; Validation: C.L. and U.RS.; Formal Analysis: C.L. and U.RS.; Investigation: A.S.; Resources: U.R.; Data Curation: A.S.; Writing Original Draft Preparation: A.S. and U.R.; Writing Review and Editing: A.S. and U.R.; Visualization: U.R.; All authors, C.L., L.D.B., U.RS, A.S., and U.R., have read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

6.3. Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

6.4. Institutional Review Board Statement

Not applicable.

6.5. Informed Consent Statement

Not applicable.

6.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper

References

[1] M. Tarokh and M. EsmaeiliGookeh, "A New Model to Speculate CLV Based on Markov Chain Model," *J. Ind. Eng. Manag. Stud.*, vol. 4, no. 2, pp. 85-102, 2017, doi: 10.22116/jiems.2017.54609.

[2] U. Salunkhe, B. Rajan, and V. Kumar, "Understanding firm survival in a global crisis," *Int. Mark. Rev.*, vol. 1, no. 1, Jan. 2021, doi: 10.1108/IMR-05-2021-0175.

[3] B. Durkaya Kurtcan and T. Ozcan, "Predicting customer churn using grey wolf optimization-based support vector machine with principal component analysis," *J. Forecast.*, vol. 1, no.1, pp. 1-7, 2023, doi*:* 10.1002/for.2960.

[4] V. Morozov, O. Mezentseva, A. Kolomiiets, and M. Proskurin, "Predicting Customer Churn Using Machine Learning in IT Startups," *Lecture Notes on Data Engineering and Communications Technologies*, vol. 77. No. 1, pp. 645-664, 2022. doi: 10.1007/978-3-030-82014-5_45.

[5] M. Mujiya Ulkhaq, A. T. Wibowo, M. R. Tribosnia, R. Putawara, and A. B. Firdauz, "Predicting Customer Churn: A Comparison of Eight Machine Learning Techniques: A Case Study in an Indonesian Telecommunication Company," in *2021 International Conference on Data Analytics for Business and Industry, ICDABI* 2021, vol. 1, no.1, pp. 42-46, 2021. doi: 10.1109/ICDABI53623.2021.9655790.

[6] R. Priyadarshi, A. Panigrahi, S. Routroy, and G. K. Garg, "Demand forecasting at retail stage for selected vegetables: a performance analysis," *J. Model. Manag.*, vol. 14, no. 4, pp. 1042-1063, Jan. 2019, doi: 10.1108/JM2-11-2018-0192.

[7] R. Manivannan, R. Saminathan, and S. Saravanan, "An improved analytical approach for customer churn prediction using Grey Wolf Optimization approach based on stochastic customer profiling over a retail shopping analysis. CUPGO," *Evol. Intell.*, vol. 14, no. 2, pp. 479-488, 2021, doi: 10.1007/s12065-019-00282-x.

[8] A. Zaky, S. Ouf, and M. Roushdy, "Predicting Banking Customer Churn based on Artificial Neural Network," in *5th International Conference on Computing and Informatics, ICCI 2022*, vol. 1, no. 1, pp. 132-139, 2022. doi: 10.1109/ICCI54321.2022.9756072.

[9] W. Park and H. Ahn, "Not All Churn Customers Are the Same: Investigating the Effect of Customer Churn Heterogeneity on Customer Value in the Financial Sector," *Sustain.*, vol. 14, no. 19, pp. 1-13, 2022, doi: 10.3390/su141912328.

[10] Y. Yamato, "Server Structure Proposal and Automatic Verification Technology on IAAS Cloud of Plural Type Servers," *IJIIS Int. J. Informatics Inf. Syst.*, vol. 1, no. 2, pp. 97-106, 2018, doi: 10.47738/ijiis.v1i2.104.

[11] M. F. Kokasih and A. S. Paramita, "Property Rental Price Prediction Using the Extreme Gradient Boosting Algorithm," *IJIIS*

Int. J. Informatics Inf. Syst., vol. 3, no. 2, pp. 54-59, 2020, doi: 10.47738/ijiis.v3i2.65.

[12] A. R. Lubis, S. Prayudani, Julham, O. Nugroho, Y. Y. Lase, and M. Lubis, "Comparison of Models in Predicting Customer Churn Based on Users' habits on E-Commerce," in *2022 5th International Seminar on Research of Information Technology and Intelligent Systems, ISRITI 2022*, vol. 1, no. 1, pp. 300-305, 2022. doi: 10.1109/ISRITI56927.2022.10052834.

[13] L. Sook Ling, N. Mustafa, and S. F. Abdul Razak, "Customer churn prediction for telecommunication industry: A Malaysian Case Study, " *F1000Research*, vol. 10, no. 1, pp. 1-13, 2021, doi: 10.12688/f1000research.73597.1.

[14] Q. Tang, G. Xia, and X. Zhang, "A hybrid classification model for churn prediction based on customer clustering," *J. Intell. Fuzzy Syst.*, vol. 39, no. 1, pp. 69-80, 2020, doi: 10.3233/JIFS-190677.

[15] O. M. Mirza *et al.*, "Optimal Deep Canonically Correlated Autoencoder-Enabled Prediction Model for Customer Churn Prediction," *Comput. Mater. Contin.*, vol. 73, no. 2, pp. 3757-3769, 2022, doi: 10.32604/cmc.2022.030428.

[16] A. Widiyanto, N. A. Prabowo, M. Ircham, N. Amarullah, and A. Soni, "The Effect of E-Learning as One of the Information Technology-Based Learning Media on Student Learning Motivation," *IJIIS Int. J. Informatics Inf. Syst.*, vol. 4, no. 2, pp. 123-129, 2021.

[17] W.-J. Su, "The Effects of Safety Management Systems, Attitude and Commitment on Safety Behaviors and Performance," *Int. J. Appl. Inf. Manag.*, vol. 1, no. 4, pp. 187-199, 2021, doi: 10.47738/ijaim.v1i4.20.

[18] H.-T. Le, "Knowledge Management in Vietnameses Mall and Medium Enterprises: Review of Literature," *Int. J. Appl. Inf. Manag.*, vol. 1, no. 2, pp. 50-59, 2021, doi: 10.47738/ijaim.v1i2.12.

[19] P. Jeyaprakaash and K. Sashirekha, "Accuracy Measure of Customer Churn Prediction in Telecom Industry using Adaboost over Decision Tree Algorithm," *J. Pharm. Negat. Results*, vol. 13, no. 1, pp. 1495-1503, 2022, doi: 10.47750/pnr.2022.13.S04.179.

[20] H. K. Thakkar, A. Desai, S. Ghosh, P. Singh, and G. Sharma, "Clairvoyant: AdaBoost with Cost-Enabled Cost-Sensitive Classifier for Customer Churn Prediction," *Comput. Intell. Neurosci.*, vol. 2022, no. 1, pp. 1-7, 2022, doi: 10.1155/2022/9028580.

[21] R. A. de Lima Lemos, T. C. Silva, and B. M. Tabak, "Propensity to customer churn in a financial institution: a machine learning approach," *Neural Comput. Appl.*, vol. 34, no. 14, pp. 11751-11768, 2022, doi: 10.1007/s00521-022-07067-x.

[22] N. Tomasevic, N. Gvozdenovic, and S. Vranes, "An overview and comparison of supervised data mining techniques for student exam performance prediction," *Comput. Educ.*, vol. 143, pp. 103676-103684, 2020, doi: https://doi.org/10.1016/j.compedu.2019.103676.

[23] J. Prayitno, B. Saputra, and R. P. Bernarte, "The Naive Bayes Algorithm in Predicting the Spread of the Omicron Variant of Covid-19 in Indonesia: Implementation and Analysis," *IJIIS Int. J. Informatics Inf. Syst.*, vol. 5, no. 2, pp. 84-91, 2022.

[24] M.-H. Tayarani N., "Applications of artificial intelligence in battling against covid-19: A literature review," *Chaos, Solitons & Fractals*, vol. 142, no. 1, pp. 110338-110345, 2021, doi: https://doi.org/10.1016/j.chaos.2020.110338.

[25] H. N. Do, W. Shih, and Q. A. Ha, "Effects of mobile augmented reality apps on impulse buying behavior: An investigation in the tourism field," *Heliyon*, vol. 6, no. 8, pp. 1-12, 2020, doi: 10.1016/j.heliyon.2020.e04667.

[26] A. Efendi, D. Purwana, and A. D. Buchdadi, "Human Capital Management of Government Internal Supervisory at the Ministry of Defense of the Republic of Indonesia," *Int. J. Appl. Inf. Manag.*, vol. 2, no. 2, pp. 81-89, 2021, doi: 10.47738/ijaim.v2i2.30.

[27] S. Hidayat, M. Matsuoka, S. Baja, and D. A. Rampisela, "Object-based image analysis for sago palm classification: The most important features from high-resolution satellite imagery," *Remote Sens.*, vol. 10, no. 8, pp. 1-12, 2018, doi: 10.3390/RS10081319.

[28] C. Ricciardi *et al.*, "Application of data mining in a cohort of Italian subjects undergoing myocardial perfusion imaging at an academic medical center," *Comput. Methods Programs Biomed.*, vol. 189, pp. 105343-105350, 2020, doi: 10.1016/j.cmpb.2020.105343.