Green Intelligent Systems and Applications

Research Article
Volume 3(1), 2023, 22-34
https://doi.org/10.53623/gisa.v3i1.249

Machine Learning Predictive Models Analysis on Telecommunications Service Churn Rate

Teuku Alif Rafi Akbar*, Catur Apriono

Department of Electrical Engineering, Faculty of Engineering, Universitas Indonesia, Kampus Baru UI Depok, Jawa Barat, 16424, Indonesia

*Correspondence: teuku.alif@ui.ac.id

SUBMITTED: 19 April 2023; REVISED: 29 May 2023; ACCEPTED: 31 May 2023

ABSTRACT: Customer churn frequently occurs in the telecommunications industry, which provides services and can be detrimental to companies. A predictive model can be useful in determining and analyzing the causes of churn actions taken by customers. This paper aims to analyze and implement machine learning models to predict churn actions using Kaggle data on customer churn. The models considered for this research include the XG Boost Classifier algorithm, Bernoulli Naïve Bayes, and Decision Tree algorithms. The research covers the steps of data preparation, cleaning, and transformation, exploratory data analysis (EDA), prediction model design, and analysis of accuracy, F1 Score, receiver operating characteristic (ROC) curve, and area under the ROC curve (AUC) score. The EDA results indicate that the contract type, length of tenure, monthly invoice, and total bill are the most influential features affecting churn actions. Among the models considered, the XG Boost Classifier algorithm achieved the highest accuracy and F1 score of 81.59% and 74.76%, respectively. However, in terms of efficiency, the Bernoulli Naïve Bayes and Decision Tree algorithms outperformed XG Boost, with AUC scores of 0.7469 and 0.7468, respectively.

KEYWORDS: Customer churn; pedictive models; accuracy; F1 score; ROC curve; AUC

1. Introduction

Customer churn refers to the action taken when a customer stops using a service or accessing content from a service provider. Essentially, it involves breaking a contract with a company. This action can have detrimental effects on service companies that strive to compete and satisfy their customers. Identifying potential churn actions is crucial for maintaining good customer service and ensuring the revenue of a service provider. Several factors can contribute to churn actions, including customer dissatisfaction resulting from poor service experiences, problematic products, lack of communication platforms, and the absence of customer loyalty programs. On average, the telecommunications industry experiences an annual churn rate of 20-40%. To maintain revenue, it is more cost-effective for companies to focus on retaining existing customers rather than acquiring new ones, as the latter can be five to ten times more expensive.

Churn rate modeling and analysis play a vital role in the telecommunications industry [1–2]. Creating a predictive churn model involves multiple steps, such as data collection,

understanding, pre-processing, learning, model design, development, validation, and evaluation [3, 4]. The effective use of training and testing data simplifies the process, ensuring the accuracy and effectiveness of the designed model. However, due to the unbalanced nature of churn data, with most customers belonging to the non-churn class, traditional machine learning methods struggle to achieve accurate classification rates. This challenge highlights the importance of customer relationship management and marketing processes in minimizing churn amid intense competition and rapid developments in telecommunications services. Customer churn can be categorized into two types: voluntary and involuntary. Voluntary churn occurs when customers willingly participate in the churn action, while involuntary churn is a consequence of delayed billing leading to the termination of a subscription. Contract terminations can result in churn from customers. Among these two types, companies often face greater difficulty in retaining customers who churn voluntarily. Understanding the reasons behind contract terminations and identifying service deficiencies can be challenging for companies.

Figure 1 illustrates the classification of customer churn into various groups [4]. Within voluntary churn, further division can be made into two sub-categories: incidental churn and deliberate churn. Incidental churn occurs when changes in a customer's life circumstances force them to terminate a contract with a company. On the other hand, deliberate churn happens when customers choose to switch to competitors or adopt new technology that offers high-quality services at competitive prices, catering to their needs [5]. Technological advancements, price competition, influence from friends and family, and experiments conducted by many individuals are among the many factors contributing to churn. Intense competition often leads to customers switching services from one company to another [6].

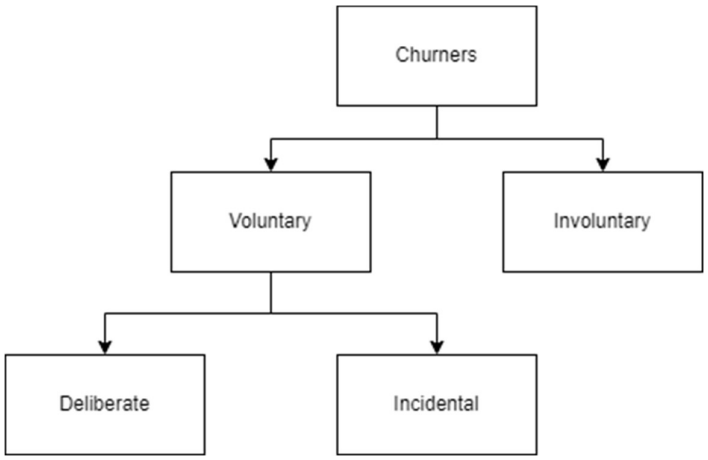


Figure 1. General Classification of Types of Churn that Occur in the Telecommunications Sector[4].

Many studies have explored churn prediction models using various classification methods. The accuracy value often serves as the primary parameter for evaluating the performance of these models. However, it is crucial to consider the specific data and features being analyzed, as well as the necessary data processing and information related to the available features and modeling techniques. Machine learning methods offer predictive models that can be adapted to different datasets [7–13]. Many of these studies employ accurate meta-heuristic algorithms. Some researchers have focused on improving sample data through more efficient

pre-processing methods, such as incorporating social features in data extraction and selecting appropriate algorithms [14].

Several studies have examined predictive models using comparable datasets [3, 7, 8, 15–29]. These studies indicate that the SVM algorithm often achieves the highest accuracy value. Another approach involves dimension reduction through feature selection before implementing classifiers, and some researchers have employed stratified splitting to improve training results on the classifier [30, 31]. These approaches have resulted in increased prediction accuracy and improved performance efficiency. With numerous studies conducted in the past five years and beyond, there is still ample room for improvement in every machine learning algorithm used for predicting customer churn across various industries. Given the wide range of machine learning algorithms available, this research aims to compare common algorithms as a starting point, providing an opportunity to consider various methods. Each algorithm possesses unique characteristics in terms of its prediction system, and this study aims to simplify the available options for predicting customer churn and further development. The dataset used primarily consists of categorical and numerical features. The analysis of the dataset employs common machine learning algorithms, as shown in Table 1.

Table 1. Comparision between 10 classifier model used in this research.

Model	Advantages between Common Classifier Models
Logistic Regression	Uses statistical approach to analyze with one or more independent variables to find the best fitting model that describes the relationship between variables [32]
Random Forest	a supervised machine learning model that solves two-group classification issues using classification techniques. An SVM model can classify incoming text after receiving sets of labeled training data for each category [24].
Support Vector Machine	a supervised machine learning model that solves two-group classification issues using classification techniques. An SVM model can classify incoming text after receiving sets of labeled training data for each category [33, 34].
K-Nearest Neighbor (n=4)	KNN perform discriminant analysis when reliable parametric estimates of probability densities are unknown or difficult to determine. One of the most fundamental classification method and simple to use [35].
Decision Tree	Most important feature is the capability of capturing descriptive decisionmaking knowledge from the supplied data, since it has the ability to its ability to use different feature subsets and decision rules at different stages of classification [36, 37].
Bernoulli Naive Bayes	It is based on probability models that incorporate strong independence assumptions [32].
Discriminant Analysis	a reliable classification technique that supports dimension reduction whether or not data normalcy is assumed. It has closed-form solutions that are simple to compute, naturally multiclass, have a good track record in reality, and don't require tuning of any hyperparameters. [33].
ADA Boost	In order to improve the effectiveness of binary classifiers, an ensemble learning technique called "meta-learning" was first developed. AdaBoost employs an iterative methodology to improve weak classifiers by learning from their errors and replacing them with stronger ones. [35].
Gradient Boost	Known as ADA Boost with Weighted Minimization, this technique can reduce loss—that is, the discrepancy between the training example's actual class value and the predicted class value. Although understanding the method for decreasing the classifier's loss is not necessary, it works similarly to gradient descent in a neural network [38].
XG Boost	A fast and accurate solution to a variety of data science issues can be found in the parallel tree strengthening technique XGBoost (based off Gradient Boosting), also called GBDt or GBM. Beyond thousands of examples, it is possible to solve problems with the same algorithm that operates in core distributed environments like Hadoop, SGE and MPI [39].

This research focuses on churn rate prediction modeling using a dataset obtained from Kaggle. The aim was to compare and analyze different algorithms to gain insights and contribute to the identification of customer churn. By designing predictive models and implementing various classification techniques, this research aimed to provide valuable findings. The algorithms considered for comparison include Logistic Regression, Random Forest Classifier, Support Vector Machine (SVM), Bernoulli Naive Bayes Classifier, K-Nearest Neighbor Classifier, Decision Tree Classifier, ADA Boost Classifier, Gradient Boost Classifier, and XG Boost Classifier. The effectiveness and accuracy of these models are evaluated using parameters such as accuracy value, average F1 score, ROC curve, and AUC score. These metrics help determine the performance of each model in predicting customer churn. Overall, this research describes predictive modeling techniques specifically tailored to address customer churn issues in the telecommunications sector.

2. Materials and Methods

Figure 2 illustrates the methodology employed in this research. Prior to analyzing and comparing machine learning algorithms using the dataset, several key stages are essential for feature clarification and further processing. The initial stage involves identifying the dataset obtained from open-source data, specifically focusing on telecommunications service providers. The subsequent stage is initialization, which includes identifying dataset features, data types, correlations, addressing empty or invalid data, transforming values in columns into numeric values, and performing data normalization [40, 41].

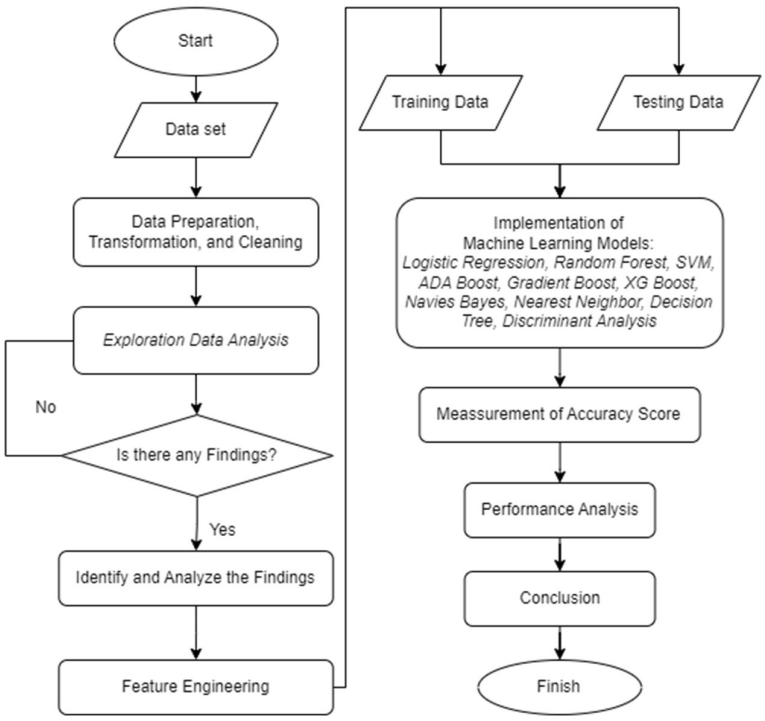


Figure 2. Proposed research methodology.

Figure 3 depicts the features present in the dataset. It is crucial to identify any missing values within the dataset. The subsequent step involves determining the correlation values between the features in the dataset. At this stage, it is observed that several factors, such as

different types of contracts, the availability of online security systems, and the presence of tech support, are positively correlated with churn. On the other hand, services like online security, streaming TV, and online backup services exhibit a negative correlation with churn. The exploratory data analysis step aims to uncover patterns within the data and develop potential hypotheses related to churn actions. This step involves learning about various aspects, including the distribution of individual variables.

```
RangeIndex: 7043 entries, 0 to 7042
Data columns (total 21 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   customerID        7043 non-null   object
 1   gender            7043 non-null   object
 2   SeniorCitizen     7043 non-null   int64
 3   Partner           7043 non-null   object
 4   Dependents        7043 non-null   object
 5   tenure            7043 non-null   int64
 6   PhoneService      7043 non-null   object
 7   MultipleLines     7043 non-null   object
 8   InternetService   7043 non-null   object
 9   OnlineSecurity    7043 non-null   object
 10  OnlineBackup      7043 non-null   object
 11  DeviceProtection  7043 non-null   object
 12  TechSupport       7043 non-null   object
 13  StreamingTV       7043 non-null   object
 14  StreamingMovies   7043 non-null   object
 15  Contract          7043 non-null   object
 16  PaperlessBilling  7043 non-null   object
 17  PaymentMethod     7043 non-null   object
 18  MonthlyCharges    7043 non-null   float64
 19  TotalCharges      7043 non-null   object
 20  Churn             7043 non-null   object
dtypes: float64(1), int64(2), object(18)
```

Figure 3. Features and data type.

After the data preparation stage, various machine learning models are implemented. The data is divided into training and testing sets. The performance of the models is evaluated using the confusion matrix, which provides information about different parameters, such as the F1 score and model accuracy. In this research, the accuracy and F1 score equations described in Equation (1) and Equation (2) are considered [26]. The confusion matrix consists of four assessment sections: True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN). The definitions of each part of the confusion matrix are as follows:

- True positives: Number of churns made that are correctly predicted to be actual churns.
- True negatives: Number of non-churns that are correctly predicted as non-churns.
- False positives: The amount of churn made is incorrectly predicted as non-churn.
- False negatives: The number of non-churns committed is incorrectly predicted as real churn.

$$Accuracy = \frac{TP+TN}{TP+FP+FN+T} \quad (1)$$

$$F1\ Score = 2 \cdot \frac{Precision \cdot \text{Recall}}{Precision+Reca} \quad (2)$$

The efficiency of the designed model can be assessed using the parameters of the receiver operating characteristic curve (ROC) and the area under the ROC curve (AUC). The ROC curve illustrates the performance of a classification model across different classification thresholds. On the other hand, the AUC score is a measurement method that quantifies the two-dimensional area under the overall ROC curve. The higher the AUC score, the better the model is at differentiating between positive and negative classes. A perfect AUC score of '1' indicates that the classifier can accurately distinguish between all 'positive' and 'negative' class points. Conversely, if the classifier predicts all negatives as positive and vice versa, the AUC score is '0'. The ROC curve utilizes parameters such as true positive rate (TPR) and false positive rate (FPR). It represents the relationship between TPR and FPR at various classification thresholds. The AUC score provides a comprehensive measure of performance across all possible classification thresholds in the model.

3. Results and Discussion

The dataset utilized in this research is sourced from open-source data available on Kaggle, specifically the IBM sample dataset, which contains information on customer data and program retention [41]. The dataset encompasses personal information of customers, services associated with the utilized program, churn actions performed by customers, and customer demographics. In total, the dataset comprises 21 features, with a recorded count of 7,044 customers. Figure 4 illustrates the distribution of the various services utilized by customers.

After exploring data related to personal data, the identified points are as follows:

- 50.5% of customers are male, and 49.5% of customers are female
- 16.2% registered as elderly residents
- 48.3% of customers have spouses, while there are around 30% of customers have dependents

The more identified information is as follows:

- 60% of customers choose to do paperless billing
- 33.6% of all customers use electronic checks as a payment method
- 55% of customers choose a month-to-month contract in subscribing to services. The rest are 1-year or 2-year contracts.
- The distribution of subscribers more (around 800 customers) in the first month, but some customers for approximately 72 months (around 500 customers)
- Customers using a 2-year contract tend to last longer in subscribing to services (around 72 months), while customers using a monthly contract (month-to-month) tend to subscribe for approximately 1-2 months.

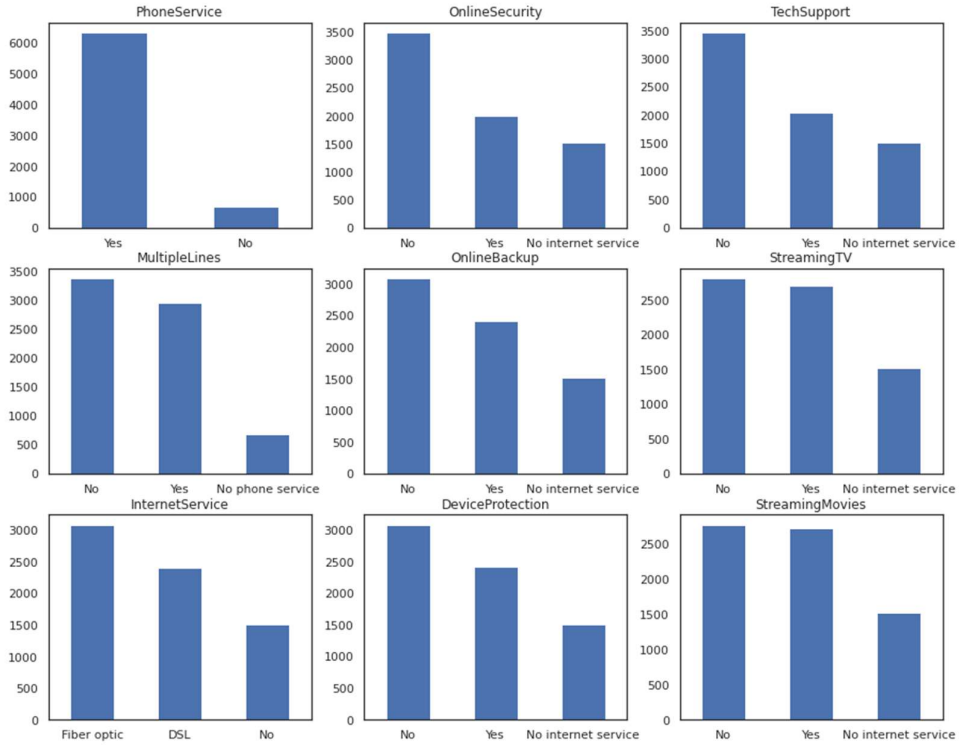


Figure 4. Distribution of services used by customers.

Figure 5 illustrates the distribution of customer churn rates, with 73.4% of customers not conducting churn. To balance the number of customers who choose to churn and those who do not, alternative methods are necessary. One approach is the use of stratified sampling/splitting or multilevel data separation to prevent an excessive number of false negatives, which can lead to lower accuracy in predicting churn actions [30, 31, 42, 43]. Stratified sampling helps achieve a balanced dataset and reduces skewness. Another method is cross-validation, which provides a stable dataset and reliable estimates of model performance. It can also be used to compare different models and training algorithms, as well as to determine optimal model parameters [44]. For example, by randomly sampling non-churners and balancing the number of non-churners with churners before splitting the data into training and testing sets for each classifier method, the data can be balanced. However, it should be noted that poor data splitting can result in inaccurate and highly variable model performance [45, 46].

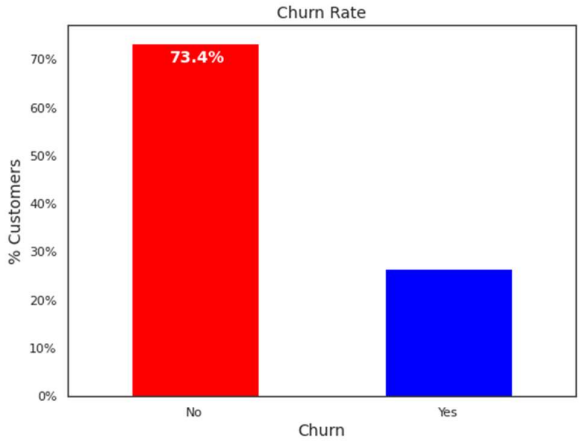


Figure 5. Customer churn rate distribution.

The followings are findings obtained in the features exploration process and their relationship to the churn rate that occurs, among others:

1. Many subscribers who use fiber optic services to provide internet are out of contract. On the other hand, customers using DSL churn less frequently.
2. Customers without internet service have very low churn rates.
3. Citizens with senior status have nearly double the churn rate than younger populations.
4. Longer the service provider, the more likely the customer will not churn.
5. The rate of customers churning is higher when the total cost or bill is lower.
6. A higher percentage of customers churn when their fees or monthly bills are high.

After identification, exploration, observation, and analysis, the models were applied to determine the accuracy value and F1 score. The data was divided into 80% for training and 20% for testing. Table 2 presents the results for each model, indicating that XG Boost has the highest F1 score and accuracy score, with values of 74.76% and 81.59%, respectively. This outcome can be attributed to the slow learning and implementation of parallel trees, which scan gradient values and utilize the partial sum of these gradient values to evaluate the quality of data splitting across all possibilities. Other results demonstrate that the accuracy value and F1 score are comparable across all models. This suggests that the data has not yet reached optimal conditions. It is possible that the data has not fully recovered from skewness and lacks a proper

distribution to create more accurate predictive models. Optimal data quality encompasses factors such as accuracy, completeness, data consistency, integrity, duplication, and timeliness [47]. A good and optimal dataset should take into account these metrics. The used dataset has a limitation in terms of the uneven size between churners and non-churners, leading to skewness and potential outliers. These outliers could result from data entry errors, measurement errors, or natural occurrences. Further studies can be conducted to identify and minimize outliers in this dataset, thus optimizing the performance of machine learning algorithms. As the dataset employed in this study has not been previously researched by other scholars, it represents a unique contribution and has not been utilized in other studies thus far.

Table 2. Model accuracy results and F1-score from machine learning models.

Model	Results (%)	
	F1 Score	Accuracy Score
Logistic Regression	74.15	80.89
Random Forest	73.59	81.24
Support Vector Machine	74.19	81.24
K-Nearest Neighbor (n=4)	69.19	78.96
Decision Tree	74.17	79.89
Bernoulli Naive Bayes	68.28	70.85
Discriminant Analysis	73.80	80.53
ADA Boost	73.92	81.02
Gradient Boost	74.22	81.16
XG Boost	74.76	81.59

Implementing various machine learning algorithms reveals that customers with a 2-month contract with the company can reduce churn. Figure 6 illustrates the feature correlation in model design and analysis using logistic regression algorithms. Variables such as total billing, monthly contracts, internet services via fiber optic cables, and the senior status of customers have a significant impact on churn.

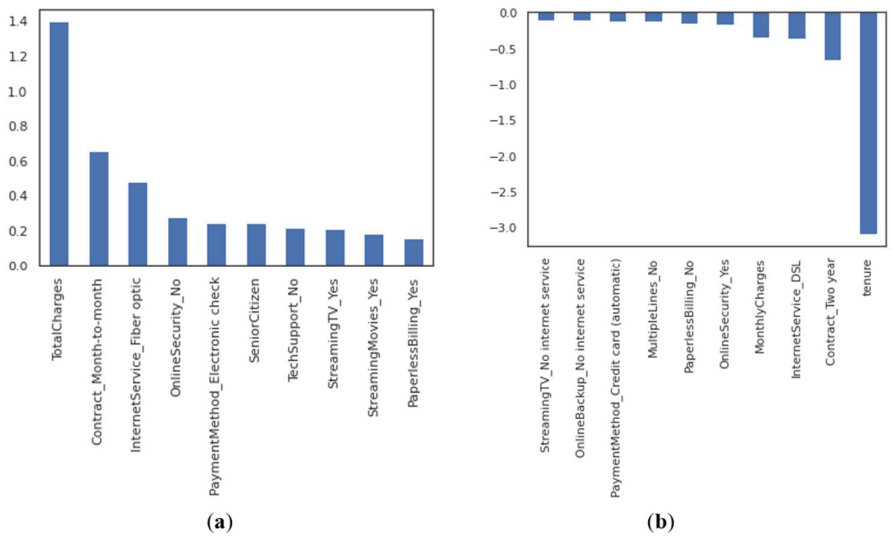


Figure 6. The plot features correlation on model design and analysis with logistic regression algorithms: **(a)** Variables with positive correlation values on features relationships with churn; **(b)** Variables with negative correlation values on features relationships with churn.

Figure 7 displays the feature importance in the dataset according to the XG Boost Algorithm. It highlights the significance of monthly and total bills in determining customer churn actions. In terms of customer services, the obtained significance values tend to fall in the category of minor importance to the churn action's effect.

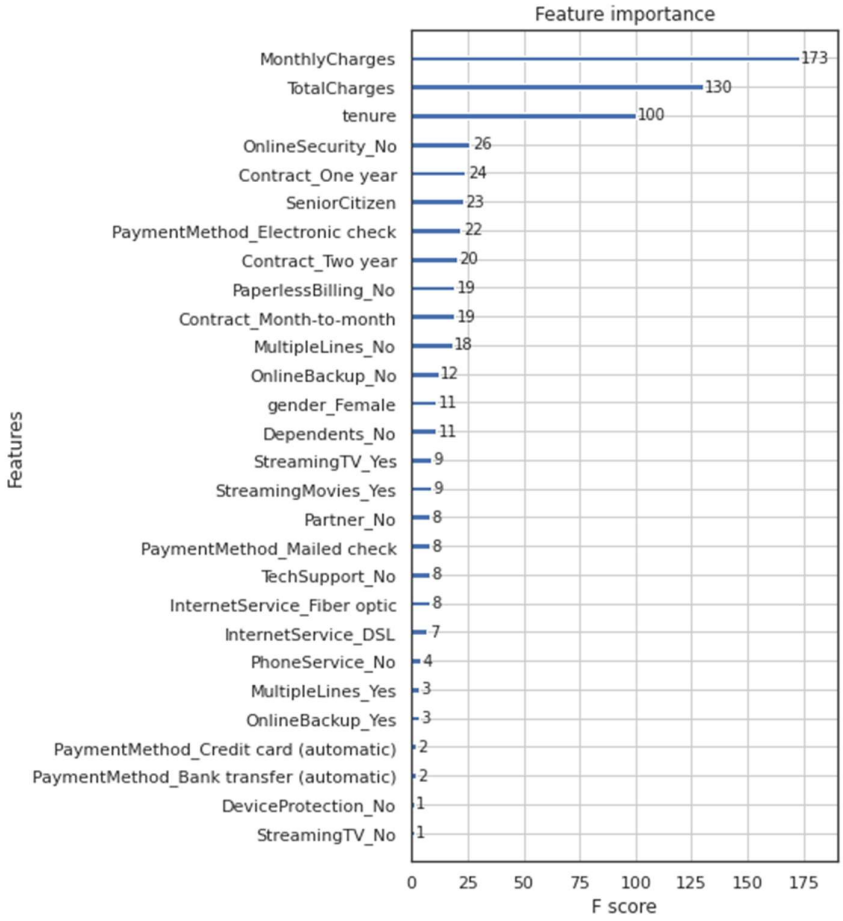
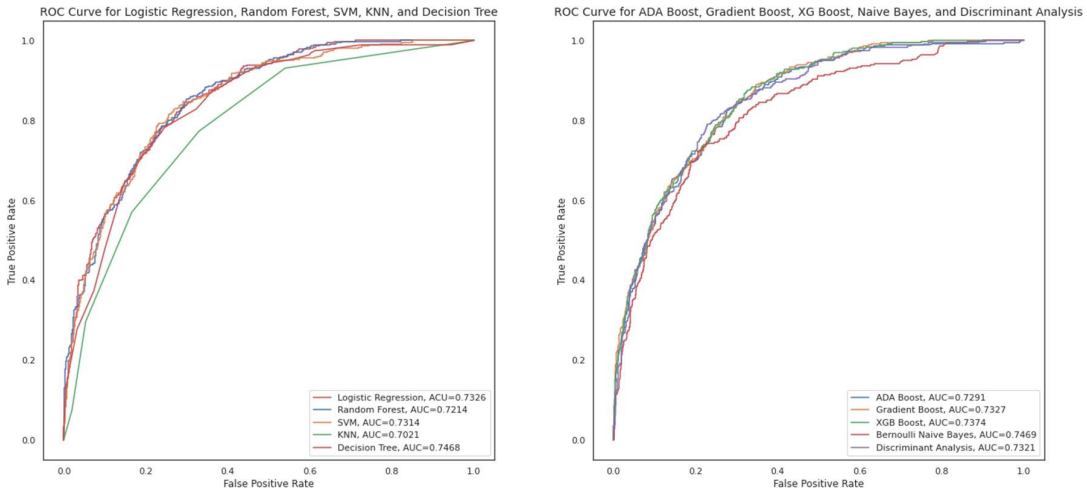


Figure 7. Feature importance in the data set based on the XG Boost Algorithm.

Figure 8 illustrates the ROC curve and AUC score obtained from the implementation of machine learning algorithms in predictive models for churn actions. The AUC score represents the level of separability between features under the ROC curve, which is a probability curve reflecting the performance of a classifier model at different thresholds. It indicates how well the model can differentiate between classes [48]. A higher AUC indicates that the model is more accurate at classifying the '0' classes as '0' and the '1' classes as '1'. For example, the higher the AUC, the better the model is at differentiating between true and false boolean answers. Based on the results of the ROC curve and AUC score, the Bernoulli Naïve Bayes and Decision Tree algorithms show the highest efficiency with nearly identical values. Although XG Boost has a higher F1 score and accuracy value compared to Decision Tree and Bernoulli Naïve Bayes, in terms of predictive model efficiency, XG Boost scores relatively lower than the two models.



(**a**) (**b**)

Figure 8. ROC curve and AUC score from machine learning algorithm implementation on predictive models to predict churn actions: **(a)** Logistic Regression, Random Forest, SVM, K-Nearest Neighbor, and Decision Tree; **(b)** ADA Boost, Gradient Boost, XG Boost, Bernoulli Naive Bayes, and Discriminant Analysis.

4. Conclusions

The type of contract, tenure, monthly bill, and total bill are the features that have the most significant influence on customer churn actions. Among the customer services, fiber optic cable service has the highest impact on churn, while DSL service has a minimal effect. Customer attributes such as gender and seniority status do not significantly affect churn tendencies. The results indicate that the XG Boost Classifier algorithm performs the best, achieving an accuracy value of 81.59% and an F1 Score of 74.76%. In terms of efficiency, the Bernoulli Naïve Bayes and Decision Tree algorithms show AUC scores of 0.7469 and 0.7468, respectively. The ROC curve for these models is better than that of XG Boost, despite XG Boost having the highest accuracy and F1 score.

Acknowledgments

This research was supported by the Department of Electrical Engineering, Faculty of Engineering, Universitas Indonesia.

Competing Interest

The authors declare no financial or non-financial competing interest.

References

[1] Xu, T.; Ma, Y.; Kim, K. (2021). Telecom Churn Prediction System Based on Ensemble Learning Using Feature Grouping. *Applied Sciences, 11, 4742.* https://doi.org/10.3390/APP11114742.

[2] Germann, F.; Lilien, G.L.; Moorman, C.; Fiedler, L.; Groβmaβ, T. (2020). Driving Customer Analytics From the Top. *Customer Needs and Solutions, 7*, 43-61. https://doi.org/10.1007/S40547-020-00109-2.

[3] Nhu, N.Y.; Van Ly, T.; Truong Son, D.V. (2022). Churn Prediction in Telecommunication Industry Using Kernel Support Vector Machines. *PLoS ONE, 17*, e0267935. https://doi.org/10.1371/journal.pone.0267935.

[4] The Telco Churn Management Handbook (accessed on 22 February 2023) Available online: https://books.google.co.id/books?hl=en&lr=&id=M_uuQx7vMngC&oi=fnd&pg=PA1&dq=Mattison+R.+Churn+Taxonomy.+In:+The+telco+churn+management+handbook.+Oakwood+Hills,+IL:+Xit+Press&ots=QHcczOeJRa&sig=If_VOjYpMoa-pZyOVMMXZbvaF58&redir_esc=y#v=onepage&q&f=false.

[5] Domingos, E.; Ojeme, B.; Daramola, O. (2021). Experimental Analysis of Hyperparameters for Deep Learning-Based Churn Prediction in the Banking Sector. *Computation, 9*, 34. https://doi.org/10.3390/COMPUTATION9030034.

[6] Ahmed, H.M.S. (2019). The Impact of Customer Churn Factors (CCF) on Customer's Loyalty. *International Journal of Customer Relationship Marketing and Management, 10*, 48-70. https://doi.org/10.4018/IJCRMM.2019010104.

[7] Panchal, M.N.; Anala, D.; Pandit, A. (2020). Churn Prediction Using Supervised Machine Learning Algorithms - Impact of Oversampling. *International Research Journal of Engineering and Technology, 7*, 1014-1019.

[8] Kaya, E.; Dong, X.; Suhara, Y.; Balcisoy, S.; Bozkaya, B.; Pentland, A. S. (2018). Behavioral Attributes and Financial Churn Prediction. *EPJ Data Science, 7*, 41. https://doi.org/10.1140/EPJDS/S13688-018-0165-5.

[9] Thakkar, H.K.; Desai, A.; Ghosh, S.; Singh, P.; Sharma, G. (2022). Clairvoyant: AdaBoost with Cost-Enabled Cost-Sensitive Classifier for Customer Churn Prediction. *Computational Intelligence and Neuroscience, 2022*, 9028580. https://doi.org/10.1155/2022/9028580.

[10] Semeraro, G.; Vassilakis, C.; Saias, J.; Rato, L.; Gonçalves, T. (2022). An Approach to Churn Prediction for Cloud Services Recommendation and User Retention. *Information, 13*, 227. https://doi.org/10.3390/INFO13050227.

[11] de Lima Lemos, R.A.; Silva, T.C.; Tabak, B.M. (2022). Propension to Customer Churn in a Financial Institution: A Machine Learning Approach. *Neural Computing and Applications*, *34*, 11751–11768. https://doi.org/10.1007/S00521-022-07067-X/FIGURES/10.

[12] Panjasuchat, M.; Limpiyakorn, Y. (2020). Applying Reinforcement Learning for Customer Churn Prediction. *Journal of Physics: Conference Series*, *1619*, 012016. https://doi.org/10.1088/1742-6596/1619/1/012016.

[13] Hu, X.; Yang, Y.; Chen, L.; Zhu, S. (2020). Research on a Customer Churn Combination Prediction Model Based on Decision Tree and Neural Network. 2020 IEEE 5th International Conference on Cloud Computing and Big Data Analytics (ICCCBDA 2020), pp. 129–132. https://doi.org/10.1109/ICCCBDA49378.2020.9095611.

[14] Oskarsdottir, M.; Bravo, C.; Verbeke, W.; Sarraute, C.; Baesens, B.; Vanthienen, J. (2016). A Comparative Study of Social Network Classifiers for Predicting Churn in the Telecommunication Industry. Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2016), pp. 1151–1158. https://doi.org/10.1109/ASONAM.2016.7752384.

[15] [Zhao, Y.; Li, B.; Li, X.; Liu, W.; Ren, S. (2005). Customer Churn Prediction Using Improved One-Class Support Vector Machine. *Lecture Notes in Computer Science*, *3584,* 300–306. https://doi.org/10.1007/11527503_36/COVER.

[16] A Support Vector Machine Approach for Churn Prediction in Telecom Industry. (accessed on 22 February 2023) Available online: https://www.researchgate.net/publication/264534919_A_Support_Vector_Machine_Approach_for_Churn_Prediction_in_Telecom_Industry.

[17] Ebrah, K.; Elnasir, S.; Ebrah, K.; Elnasir, S. (2019). Churn Prediction Using Machine Learning and Recommendations Plans for Telecoms. *Journal of Computer and Communications*, *7*, 11, 33–53. https://doi.org/10.4236/JCC.2019.711003.

[18] Ajitha, P.; Sivasangari, A.; Gomathi, R.M.; Indira, K. (2020). Prediction of Customer Plan Using Churn Analysis for Telecom Industry. *Recent Advances in Computer Science and Communications*, *13*, 926–929. https://doi.org/10.2174/2213275912666190410114104.

[19] Lu, N.; Lin, H.; Lu, J.; Zhang, G. (2014). A Customer Churn Prediction Model in Telecom Industry Using Boosting. *IEEE Transactions on Industrial Informatics*, *10*, 2, 1659-1665. https://doi.org/10.1109/TII.2012.2224355.

[20] Verbekea, W.; Martensa, D.; Muesc, C.; Baesensa, B. (2011). Building comprehensible customer churn prediction models with advanced rule induction techniques. *Expert Systems with Applications*, *38*, 3, 2354-2364. https://doi.org/10.1016/j.eswa.2010.08.023.

[21] Gerpott, T. J.; Rams, W.; Schindler, A. (2001). Customer Retention, Loyalty, and Satisfaction in the German Mobile Cellular Telecommunications Market. *Telecommun Policy*, *25*, 4, 249–269. https://doi.org/10.1016/s0308-5961(00)00097-5.

[22] Wei, C. P.; Chiu, I. T. (2002) Turning Telecommunications Call Details to Churn Prediction: A Data Mining Approach. *Expert Syst Appl*, *23*, 2, 103–112. https://doi.org/10.1016/s0957-4174(02)00030-1.

[23] Jain, H., Khunteta, A.; Srivastava, S. (2020). Churn Prediction in Telecommunication Using Logistic Regression and Logit Boost. *Procedia Computer Science*, *167*, 101–112. https://doi.org/10.1016/J.PROCS.2020.03.187.

[24] Ullah, I.; Raza, B.; Malik, A. K.; Imran, M.; Islam, S. U.; Kim, S. W. (2019). A Churn Prediction Model Using Random Forest: Analysis of Machine Learning Techniques for Churn Prediction and Factor Identification in Telecom Sector. *IEEE Access*, *7*, 60134–60149. https://doi.org/10.1109/ACCESS.2019.2914999.

[25] Geiler, L.; Affeldt, S.; Nadif, M. (2022). A Survey on Machine Learning Methods for Churn Prediction. *International Journal of Data Science and Analytics*, *14*, 3, 217–242. https://doi.org/10.1007/S41060-022-00312-5.

[26] AL-Najjar, D.; Al-Rousan, N.; AL-Najjar, H. (2022). Machine Learning to Develop Credit Card Customer Churn Prediction. *Journal of Theoretical and Applied Electronic Commerce Research*, *17*, 4, 1529–1542. https://doi.org/10.3390/jtaer17040077.

[27] Tanha, J.; van Someren, M.; Afsarmanesh, H. (2017). Semi-Supervised Self-Training for Decision Tree Classifiers. *International Journal of Machine Learning and Cybernetics*, *8*, 1, 355–370. https://doi.org/10.1007/s13042-015-0328-7.

[28] Ahmed, M.; Afzal, H.; Siddiqi, I.; Amjad, M. F.; Khurshid, K. (2018). Exploring Nested Ensemble Learners Using Overproduction and Choose Approach for Churn Prediction in Telecom Industry. *Neural Comput. Appl.*, *32*, 8, 3237–3251. https://doi.org/10.1007/s00521-018-3678-8.

[29] Hudaib, A.; Dannoun, R.; Harfoushi, O.; Obiedat, R.; Faris, H. (2015). Hybrid Data Mining Models for Predicting Customer Churn. *International Journal of Communications, Network and System Sciences*, *8*, 5, 91–96. https://doi.org/10.4236/IJCNS.2015.85012.

[30] Lynn, P. (2019).The Advantage and Disadvantage of Implicitly Stratified Sampling. *Methods, Data, Analyses*, *13*, 2, 253–266. https://doi.org/10.12758/mda.2018.02.

[31] May, R. J.; Maier, H. R.; Dandy, G. C. (2010). Data Splitting for Artificial Neural Networks Using SOM-Based Stratified Sampling. *Neural Networks*, *23*, 2, 283–294. https://doi.org/10.1016/j.neunet.2009.11.009.

[32] Tsangaratos, P.; Ilia, I. (2016). Comparison of a Logistic Regression and Naïve Bayes Classifier in Landslide Susceptibility Assessments: The Influence of Models Complexity and Training Dataset Size. *Catena*, *145*, 164–179. https://doi.org/10.1016/j.catena.2016.06.004.

[33] Alkan, A.; Günay, M. (2012). Identification of EMG Signals Using Discriminant Analysis and SVM Classifier. *Expert Systems with Applications*, *39*, 1, 44–47. https://doi.org/10.1016/J.ESWA.2011.06.043.

[34] Li, X.; Wang, L.; Sung, E. (2008). AdaBoost with SVM-Based Component Classifiers. *Engineering Applications of Artificial Intelligence*, *21*, 5, 785–795. https://doi.org/10.1016/j.engappai.2007.07.001.

[35] Tian, J.; Morillo, C.; Azarian, M. H.; Pecht, M. (2016). Motor Bearing Fault Detection Using Spectral Kurtosis-Based Feature Extraction Coupled with K-Nearest Neighbor Distance Analysis. *IEEE Transactions on Industrial Electronics*, *63*, 3, 1793–1803. https://doi.org/10.1109/TIE.2015.2509913.

[36] Bhukya, D. P.; Ramachandram, S. (2010). Decision Tree Induction: An Approach for Data Classification Using AVL-Tree. *International Journal of Computer and Electrical Engineering*, 660–665. https://doi.org/10.7763/IJCEE.2010.V2.208.

[37] Song, Y. Y.; Lu, Y. (2015). Decision Tree Methods: Applications for Classification and Prediction. *Shanghai Archives of Psychiatry*, *27*, 2, 130. https://doi.org/10.11919/J.ISSN.1002-0829.215044.

[38] Blagus, R.; Lusa, L. (2017). Gradient Boosting for High-Dimensional Prediction of Rare Events. *Computational Statistics and Data Analysis*, *113*, 19–37. https://doi.org/10.1016/j.csda.2016.07.016.

[39] Chen, Z.; Jiang, F.; Cheng, Y.; Gu, X.; Liu, W.; Peng, J. (2018). XGBoost Classifier for DDoS Attack Detection and Analysis in SDN-Based Cloud. *Proceedings - 2018 IEEE International Conference on Big Data and Smart Computing,* 251–256. https://doi.org/10.1109/BIGCOMP.2018.00044.

[40] Ahmad, A. K.; Jafar, A.; Aljoumaa, K. (2019). Customer Churn Prediction in Telecom Using Machine Learning in Big Data Platform. *Journal of Big Data*, *6*, 1, 1–24. https://doi.org/10.1186/S40537-019-0191-6.

[41] Telco Customer Churn (11.1.3+). (accessed on 22 February 2023) Available online: https://community.ibm.com/community/user/businessanalytics/blogs/steven-macko/2019/07/11/telco-customer-churn-1113.

[42] Joseph, V. R. (2022). Optimal Ratio for Data Splitting. *Statistical Analysis and Data Mining*, *15*, 4, 531–538. https://doi.org/10.1002/SAM.11583.

[43] Stratified Sampling: You May Have Been Splitting Your Dataset All Wrong | Towards Data Science. (accessed on 28 May 2023) Available online: https://towardsdatascience.com/stratified-sampling-you-may-have-been-splitting-your-dataset-all-wrong-8cfdd0d32502.

[44] Ağbulut, Ü.; Gürel, A. E.; Biçen, Y. (2021). Prediction of Daily Global Solar Radiation Using Different Machine Learning Algorithms: Evaluation and Comparison. *Renewable and Sustainable Energy Reviews*, *135*, 110114. https://doi.org/10.1016/J.RSER.2020.110114.

[45] Roshan, V.; Stewart, J. H. M.; Joseph, R.; Stewart, H. M. (2022). Optimal Ratio for Data Splitting. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, *15*, 4, 531–538. https://doi.org/10.1002/SAM.11583.

[46] Joseph, V. R.; Vakayil, A. (2022). Split: An Optimal Method for Data Splitting. *Technometrics*, *64*, 2, 166–176. https://doi.org/10.1080/00401706.2021.1921037.

[47] Batini, C.; Cappiello, C.; Francalanci, C.; Maurino, A. (2009). Methodologies for Data Quality Assessment and Improvement. *ACM Computing Surveys (CSUR)*, *41*, 3, 16, 1-52. https://doi.org/10.1145/1541880.1541883.

[48] Hoo, Z. H.; Candlish, J.; Teare, D. (2021). What Is an ROC Curve?. *Emergency Medicine Journal*, *34*, 6, 357–359. http://doi.org/10.1136/emermed-2017-206735.