Analysis Implementation of the Ensemble Algorithm in Predicting Customer Churn in Telco Data: A Comparative Study

Renny Puspita Sari*, Ferdy Febriyanto, Ahmad Cahyono Adi
Department of Information System, Faculty of Mathematic and Natural Sciences, Tanjungpura University, Indonesia
E-mail: rennysari@sisfo.untan.ac.id

Keywords: churn, ensemble algorithm, adaboost, gradient boost, XGBoost, CatBoost, LightGBM

Received: April 13, 2023

Globalization and technological advancements in the telecommunication industry have led to a significant rise in the number of operators, leading to intense market competition. This sector has become crucial in developed countries, and companies strive to increase profits by acquiring new customers, up-selling existing ones, and extending the retention period of current clients. In the traditional method of defect prediction, a single classifier is used to build a model on a pre-labeled dataset. However, this approach has limitations in predicting defects accurately under certain circumstances. To overcome these limitations, boosting is applied to combine multiple weak classifiers and create a robust classification model. Among many algorithms used for churn prediction, ensemble techniques have demonstrated greater accuracy than simpler approaches. This study aims to overcome these limitations by experimenting with five ensemble algorithms, including Adaboost, Gradient Boost, XGBoost, CatBoost, and LightGBM. The results indicate that XGBoost outperforms other techniques and is the most suitable algorithm to build the predictive model. Additionally, the study achieves higher accuracy by performing a Grid Search CV hyper-parameter setting with XGBoost, resulting in an accuracy of 81.2%.

Povzetek: Študija je primerjala pet ansambelskih algoritmov za napovedovanje prekinitve naročniškega razmerja. Rezultati kažejo, da je XGBoost najboljši algoritem z natančnostjo 81,2 %.

1 Introduction

The telecommunication industry's globalization and advancements have resulted in an exponential increase in operators, leading to heightened market competition [1]. Over the past two decades, the telecommunications sector has emerged as a critical industry in developed countries [2]. Due to the availability of extensive data, data mining has become essential for prediction and analysis in this industry. One primary application of data mining is predicting churners, which helps increase customer retention and profitability. Data mining techniques are commonly used in telecom to monitor customer churn behavior. Customers anticipate excellent services at reasonable prices. If they are dissatisfied, they will quickly switch to another telecom network. In such a competitive market, companies must discover innovative approaches to forecast possible customer churn to thrive. Customer churn refers to the percentage of customers who have stopped utilizing a company's products or services during a specific period [3].

To maximize profits in this competitive era, companies have proposed various strategies, including acquiring new customers, up-selling existing customers, and increasing the retention period of existing customers. Among these strategies, customer retention is the least expensive. To adopt this strategy, companies must reduce potential customer churn, which occurs when customers switch service providers due to dissatisfaction with the consumer service and support system. Dissatisfaction, increased costs, poor quality, lack of features, and privacy concerns are among the reasons why customers may churn. To address this problem, it is necessary to forecast which customers are at risk of churning [4,5,6].

The telecommunications industry is expanding rapidly thanks to various technologies, and different companies offer varying data communication services with different levels of quality. To combat churn, companies offer various attractive services to retain their customers. Data mining technologies, such as Naïve Bayes, decision trees, neural networks, and logistic regression algorithms, are used to predict churn. An accurate prediction model is essential for correctly identifying customer churn and is critical in making retention decisions [7]. The most effective customer churn prediction model can identify churners and guide decision-makers to generate maximum profit [8,9].

There are differences between the use of algorithms in determining customer churn on telco data. These differences are influenced by variables, types of data, and the amount of data that varies. Therefore, previous studies have differences in determining the best algorithm for customer churn analysis. In this study, a literature study and experiment approach was used to prove the best algorithm that can be used in customer churn analysis. This study combines two methods: the experimental method as the primary method and the literature study method as a comparison method so that the results of this study produce a comparison of the experiments that have been carried out using the worst possibility and the consequences of pre-existing research.

Churn analysis on telco data is a crucial process for any company operating in the telecommunications industry. The main aim of this analysis is to predict the likelihood of customers abandoning a company's services. By identifying customers likely to churn, companies can take proactive measures to retain them and prevent revenue loss. This analysis involves using machine learning techniques and data analysis methods to identify the factors influencing a customer's decision to stop using the company's services.

Churn analysis on telco data can benefit companies in several ways. Firstly, it can help improve customer retention rates, as companies can take timely action to retain customers at risk of churning. This can lead to increased customer loyalty and higher revenue. Secondly, it can help increase customer satisfaction by identifying and addressing the factors that lead to customer dissatisfaction. This can lead to an overall improvement in the quality of services provided by the company. Lastly, it can reduce the cost of acquiring new customers, as retaining existing customers is often cheaper than acquiring new ones.

In conclusion, churn analysis on telco data is a critical process for companies operating in the telecommunications industry. It can help improve customer retention, increase customer satisfaction, and reduce the cost of acquiring new customers. By leveraging the right algorithmic approach, companies can gain valuable insights into customer behavior and make informed decisions to retain their customers. There is still scope for further research to explore new techniques and approaches to churn analysis and strengthen existing ones.

1.1 An ensemble algorithm

The conventional approach to forecasting defects involves utilizing a solitary classifier, such as the naïve Bayes classifier, decision trees, or a multilayer perceptron, to establish a predictive model based on a dataset that has already been labeled. However, particular classifiers may not be effective in predicting certain defects in specific situations. To tackle this problem, ensemble learning combines the advantages of several classifiers to enhance the identification of defects in the dataset. In recent years, various researchers have demonstrated through empirical evidence that ensemble methods yield greater classification accuracy than individual classifiers. [10]. Boosting is an approach to combining several weak classifiers to create a robust classifier. The first algorithm designed for binary classification to enhance accuracy was AdaBoost, now considered a practical technique for various types of boosting in machine learning. However, AdaBoost has an inherent drawback of being a cost-insensitive boosting algorithm, which limits its applications in situations where the costs of misclassification errors need to be treated differently. This study aims to explore ways to overcome this limitation. [11].

Previous studies have indicated that an effective churn prediction model should efficiently use a large volume of historical data to identify churners accurately. However, existing models have several limitations that prevent efficient and accurate churn prediction. The telecom sector generates large amounts of data that may contain missing values, resulting in poor and inaccurate prediction outputs. The churn prediction model being proposed combines clustering and classification algorithms. Its performance is assessed on various datasets used for churn prediction. The evaluation involves using accuracy, precision, recall, and f-measure metrics. The research goals are to pinpoint problems in previous studies and create a more efficient model for predicting customer churn and accurately identify potential churners and offer retention strategies to them. The experimental results show that the proposed churn prediction model achieves higher accuracy and performs better in predicting churn [11]

The research presents a new approach to prediction using a hybrid ensemble model that combines various classifiers. The proposed model is tested on two datasets related to telecom and demonstrates high accuracy. However, the model is specifically designed for particular datasets rather than being generalized. Additionally, the article introduces another methodology for churn prediction in fund management services that utilize ensemble learning and introduces a new weighting mechanism to handle imbalanced cost sensitivity when dealing with financial data. The model uses data from different companies and can be enhanced through other learning techniques [12,13]. Ensemble learning refers to the process of creating and combining multiple learners to achieve better results than what could be achieved with a single algorithm. The method utilizes machine learning algorithms to produce weak predictive outcomes based on features extracted from different data projections. These results are then combined using diverse voting methods to achieve better performance. [14,15,16].

Özer Çelik et al. found that deep learning techniques are beneficial for analyzing vast quantities of data. In contrast, ensemble machine learning techniques are more suitable for smaller datasets to achieve superior prediction outcomes. Additionally, the study discovered that the Cox Regression approach effectively identifies highly dispersed independent features in the dataset. However, the primary constraint of this study was how the dataset was adapted, where separate datasets were employed for machine learning and deep learning techniques. Utilizing a single dataset for machine learning and another for deep learning would have been more effective. [18].

Deng et al. predicted customer churn using Catboost, LightGBM, and RandomForest (RF), which produced the best algorithm, Random Forest, with 92% accuracy. Random Forest exhibited the best performance in this experiment, followed by Lightgbm. The dataset used in this experiment had many samples with over 80 features, making it susceptible to over-fitting. However, Random Forest employs an integrated algorithm, and its accuracy surpasses most unique algorithms. Due to the incorporation of two randomness factors, Random Forest is less prone to over-fitting and has a degree of resistance to noise, which results in a strong performance on the testing set. The model can identify the correlation between customer attributes, service attributes, customer spending data, and loss and provide specific mathematical formulas or rules to determine customer churn probability. By leveraging this information, customer churn can be effectively minimized through improved customer service. [19].

Thakkar et al. utilized AdaBoost with a Cost-Enabled Cost-Sensitive Classifier to predict customer churn. While various classifiers and boosting techniques, such as deep learning (DL) algorithms, have been recommended to address customer churn, traditional classification algorithms rely on an error-based framework that prioritizes improving the classifier's accuracy rather than cost sensitivity. In real-world scenarios, misclassification errors are unequal, but conventional classification algorithms treat them as such. However, DL algorithms are computationally intensive and time-consuming. To overcome these challenges, the study proposes a new class-dependent cost-sensitive boosting algorithm called AdaBoostWithCost, which seeks to minimize the cost of churn. The research assesses the proposed algorithm and demonstrates that it consistently outperforms the discrete AdaBoost algorithm in telecom churn prediction. The main objective of the AdaBoostWithCost classifier is to significantly decrease false harmful errors and misclassification costs compared to the AdaBoost algorithm. [20].

Ahmad et al. predicted customer churn using four models: Decision Tree, Random Forest, Gradient Boosted Machine Tree "GBM," and Extreme Gradient Boosting "XGBOOST." The model developed a churn prediction model using four different models: Decision Tree, Random Forest, Gradient Boosted Machine Tree "GBM," and Extreme Gradient Boosting "XGBOOST." Their work involved utilizing machine learning techniques on a big data platform and introducing a novel feature engineering and selection approach. To evaluate model performance, they adopted the Area Under Curve (AUC) standard measure, which yielded a value of 93.3%. Another significant contribution of the study was incorporating customer social networks into the prediction model through Social Network Analysis (SNA) feature extraction, which improved model performance from 84% to 93.3% against the AUC standard. The researchers worked on a large dataset created by transforming big raw data provided by SyriaTel telecom company and tested the model using the Spark environment. The best results were achieved by implementing the XGBOOST algorithm for classification in this churn predictive model. [21].

Based on the literature review that has been done, it is evident that customer churn analysis has become a crucial research topic in the telecommunication industry, given the significant impact of customer churn on company revenue and growth. Various studies have used machine learning algorithms, particularly ensemble algorithms, to analyze customer churn and predict customer behavior. However, there has yet to be a consensus on the best ensemble algorithm sequence for customer churn analysis. Therefore, this study aims to fill the gap by comparing and evaluating the performance of several commonly used ensemble algorithms, including XGBoost, AdaBoost, CatBoost, LightGBM, and Gradient Boost, on telco data, using a set of existing variables, including those with the worst possibility. The study aims to determine the best-performing algorithm sequence and provide insights into the factors contributing to churn prediction accuracy.

Furthermore, this study aims to strengthen previous research using ensemble algorithms to perform customer churn analysis using a machine-learning approach. Previous studies have shown promising results in using ensemble algorithms to predict customer churn. Still, the results are often limited by the choice of variables, the algorithm sequence used, and the evaluation metrics applied. This study intends to address these limitations by comprehensively evaluating various ensemble algorithms using telco data and a set of existing variables. This study hopes to contribute to the growing body of literature on customer churn analysis and provide practical insights for telecommunication companies to improve their customer retention strategies.

2 Material and method

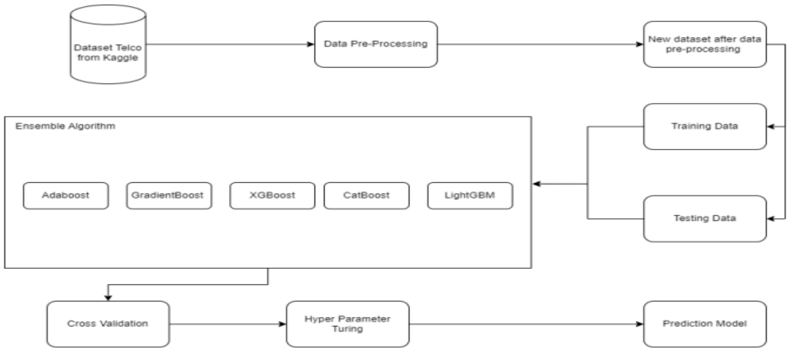

Figure1: Overall methodological framework.

The methodology used to create the prediction model in this research is illustrated in Figure 1. The individual steps involved in the process are elaborated on in the following sections. The dataset utilized in this study is sourced from the website www.kaggle.com and pertains to customer data from Telco. The data consists of 1409 entries, each with a variety of attributes such as the data contains information on customers who have terminated their services within the past month (listed in the "Churn" column), the services that each customer has subscribed to, such as phone, internet, online security, and streaming TV and movies; details on the customer's account, including their length of tenure, contract, payment method, monthly charges, and total charges; and demographic information about the customers, including their gender, age range, and whether they have partners and dependents. This task aims to conduct data preprocessing and eliminate any incomplete, noisy, or untrustworthy data from the system. This process is critical in creating a forecasting model to identify customer churn behavior. The "Python Pandas Library" was utilized to accomplish this goal. In this data, attribute disposal is carried out, namely, the customer ID, which is not used in the process analysis. The attributes of each data are described in Table 1.

2.1 Experiment and evaluation

Each model described above underwent training and testing through a 5-fold cross-validation method, which is an honest approach where each observation from the dataset has an equal opportunity to become part of the train and test sets [22]. Due to the small dataset, this approach is considered appropriate. The models were evaluated based on standard metrics for text classification, including accuracy (i.e., the proportion of correctly identified customer churn), recall (ratio of actual customer churn determined), precision (proportion of correctly identified customer churn cases), and F1-score (harmonic mean between recall and accuracy). The results for all metrics are reported in percentages, with higher values indicating better performance [23].

dataset. The correlation between variables in the telco dataset to find customer churn refers to the relationship or association between variables in the dataset that can be used to detect or identify customer churn behavior in the telecommunications industry. Telco datasets typically include information on phone calls, data usage, subscriptions, and customer costs, and these variables can have a significant correlation with customer churn behavior. In correlation analysis, a correlation coefficient measures the strength of the relationship between two variables. The correlation coefficient ranges from -1 to 1, with 1 indicating.

Table 1: The dataset atrribute

No	Attribute	Description	Representation
1	customerID	The unique string for the customer	Combine alphabeth, symbol, and number with 8 characters
2	Gender	Whether the customer is a male or female	String
3	SeniorCitizen	Whether the customer is a senior citizen or not	Boolean
4	Partner	Whether the customer is a partner or not	Boolean
5	Dependents	Whether the customer has a dependent or not	Boolean
6	Tenure	Number of months the customer has stayed with the company	Number
7	PhoneService	Whether the customer has a phone service or not	Boolean
8	MultipleLines	Whether the customer has multiple line or not	String
9	InternetService	Customer’s internet service provider (DSL, Fiber Optic or no)	String
10	OnlineSecurity	Whether the customer has online security or not	String

3 Result and discussion

3.1 Result of identifying highly correlated features

The focus of the following discourse is on the outcomes of a comparison of multiple algorithms' efficiency in forecasting non-churners and churners in the telecommunications sector. Several algorithms were leveraged and evaluated against specific metrics during the prediction process. Further exploration of this endeavor is provided in the following section. The correlation between datasets is demonstrated in Figure 2.

An extensive analysis was conducted during this procedure to comprehend the connection between various attributes and the target attribute. Pearson correlation analysis was utilized for this purpose. Figure 2 presents the numerical values and total scores obtained from comparing and contrasting the relationship between each variable and the target variable, 'churn.' The outcomes helped identify the highly correlated features with an absolute score of 0.5 or more.

Further analysis was then performed to recognize only three out of ten attributes with a specific correlation level with the target variable. Therefore, the remaining highly correlated features were successfully eliminated from the

a perfectly positive relationship, matter of -1 telling a perfectly negative relationship, and 0 indicating no relationship. In the context of customer churn analysis on telco datasets, the correlation between variables can help to identify the most influential variables in predicting customer churn behavior. It can be used to select the most essential features in the prediction model. In addition, correlation analysis can also help reduce the dataset's dimensions by eliminating highly correlated or redundant variables, thereby increasing the performance and efficiency of the prediction model.

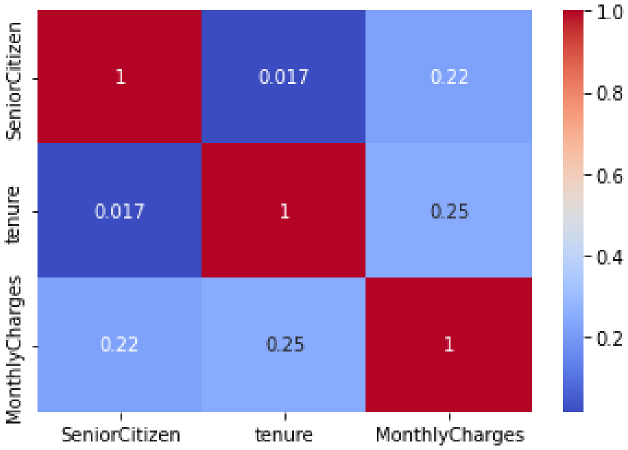

Figure 2: Correlation between variables of the data set.

After identifying the critical attributes, the dataset was divided into two non-equivalent sections: training and testing. These sections were then fed into different machine learning techniques, such as Adaboost, Gradient Boost, XGBoost, CatBoost, and LightGBM, to determine the best one for creating the final forecasting model.

The performance results during the forecasting process were assessed using several methods, including Accuracy, Confusion Matrix, Precision, Recall, and F1-Score, with K-fold cross-validation as the primary technique. [24]. Based on the results of the tests carried out, the values of each evaluation model can be seen in Table 2.

The results of machine learning algorithms used to build different models are presented in Table 2, which shows the accuracy, recall, precision, and F1 scores. Based on the outcomes, most ensemble algorithms effectively predicted customer churn, with an accuracy rate of over 80%, using 1409 datasets. Among the five algorithms tested, XGBoost demonstrated the highest accuracy, with an average accuracy value of 81.2%, recall of 91%, precision of 84%, and F1-Score of 88%. This confirms that XGBoost outperformed the other techniques and is the most suitable algorithm for the final predictive model. Moreover, to achieve even higher accuracy, a Grid Search CV hyper-parameter setting was performed with XGBoost, resulting in an accuracy of 81.2% in forecasting churning behavior.

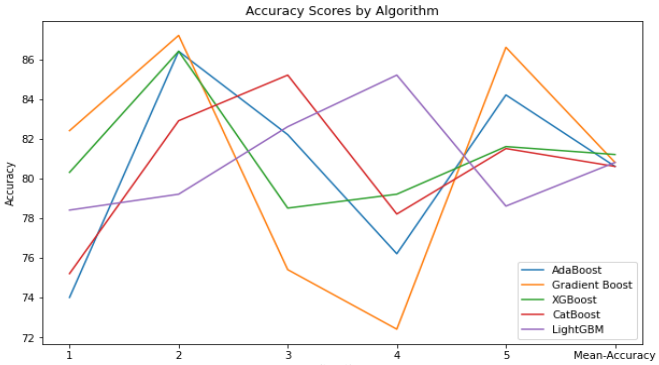

Figure3: Iteration process accuracies of 5 algorithms.

3.2 Comparison of performance

In addition to the results obtained from the tables as mentioned earlier, the K-fold cross-validation technique was performed with five folds to find the best approach to deal with any proportion of testing and training data can be seen in the following Figure 3. The graph presented illustrates the accuracy of various algorithms for each fold and the mean accuracy.

Overall, the accuracy decreases after the 2nd fold and continues declining until the 3rd fold, except for ensemble techniques demonstrating a more stable accuracy level. Although there is a slight improvement in the accuracy of LightGBM in the 5th fold, the highest mean accuracy is achieved only by XGBoost. XGBoost is proven to be the best algorithm in the class of ensemble algorithms because it has a better level of accuracy [25].

Based on Swetha et al., who conducted a customer churn analysis in the telecommunication industry using XGBoost, the accuracy was 99.6%. The XGBoost algorithm has a good improvisation with an increase in accuracy from 27.4% to 52%, which increases gradually

Fahd Idrissi Khamlichi et al. [26] conducted a study where various standalone machine learning techniques, including Random Forest, XGBoost, SVM, Decision Tree, Logistic Regression, and KNN, were applied to a publicly available dataset containing 5000 samples. According to their findings, XGBoost was the most successful technique, achieving an accuracy of 95% and an F-measure of 80%.

Table 3: Confusion matrix of the classifier

Models	Confussion Matrix			
	TP	FP	FN	TN
Adaboost	930	106	206	167
Gradient Boost	944	92	195	167
XGBoost	**947**	**89**	**197**	**176**
CatBoost	933	103	202	171
LightGBM	929	107	210	163

Sagar et al. used the XGBoost algorithm on a dataset with 52,332 customer records, of which 46,204 were non-churned and 6,128 were churned customers. They achieved an accuracy of 97% and an f1-score of 88% on this dataset. In addition, the researchers used a publicly available dataset with 3,333 subscribers to compare their results with previous research.

On this dataset, they obtained an improved accuracy and f1-score of 96.25% and 86.34%, respectively. [27]. Table 4 provides a comparison of previous research with the experimental results of this study. By examining this table, researchers can gain insights into the similarities and differences between previous research and the results obtained in this study.

This comparison can help researchers identify areas where further research is needed, or new experimental approaches may be necessary to improve the accuracy of customer churn analysis. By analyzing the strengths and weaknesses of previous research, researchers can refine their experimental approaches and better understand the most effective algorithms and techniques for analyzing customer churn in telco data. Table 4 compares the results of four previous studies that carried out comparisons using the best scheme.

The best scheme involves maximizing data preprocessing and optimizing the algorithm's performance by focusing on several parameters to achieve the highest accuracy possible. As a result, previous studies have found that XGBoost is the algorithm that has the best accuracy for analyzing customer churn. However, this study used the worst scheme instead of the best scheme. The worst scheme used a smaller number of datasets compared to previous studies. It did not perform any special treatment on the dataset other than processing it without changing several data types from each dataset. Despite using the worst scheme, the XGBoost algorithm still outperformed the different algorithms in terms of accuracy. The experiments carried out in this study provide new insights into how XGBoost can achieve high accuracy even without using the best scheme. However, it should be noted that in a real case study, it is not feasible to implement the worst scheme. The worst scheme only tests the algorithm's strength in detecting and analyzing data.

This study further strengthens previous research findings by using a different approach in the ensemble algorithm comparison. While previous studies tended to use the best schema to improve the algorithm's accuracy, this study utilized a different schema to identify the best algorithm from the existing ensemble algorithm. Despite using a different schema, the results and conclusions of this study are consistent with previous studies, which have consistently found that the XGBoost algorithm is the most effective algorithm for analyzing customer churn with telco data. The findings suggest that the XGBoost algorithm is robust and can achieve high accuracy even when a less optimal schema is utilized. These results can be valuable for businesses looking to improve customer retention and reduce customer churn. By using the XGBoost algorithm, companies can effectively predict customer churn and implement strategies to retain their customers.

4 Conclusion

The experiment costumer churn prediction with 5 ensemble algorithms shows the result the most most of the ensemble algorithms accurately predicted customer churn, achieving an accuracy rate of over 80% using 1409 datasets.

XGBoost performed the best among the five algorithms tested, achieving the highest accuracy rate. The dataset is then divided into training and testing sections and fed into various machine learning techniques, such as Adaboost, Gradient Boost, XGBoost, CatBoost, and LightGBM, to find the best algorithm for creating the final forecasting model. The performance of the models is evaluated using different metrics, including Accuracy, Confusion Matrix, Precision, Recall, and F1-Score, with K-fold cross-validation being the primary method. the results of different machine learning algorithms used to construct various models. These results show that the majority of ensemble algorithms were successful in predicting customer churn, with an accuracy rate of over 80% using 1409 datasets. Among the five algorithms evaluated, XGBoost showed the best accuracy, with an average of 81.2%, recall of 91%, precision of 84%, and an F1-Score of 88%.

Table 4: The comparison of recent research

Recent Work	Best Algorithm	Best Accuracy	Dataset	Schema Type
Ahmad et al (2019)	XGboost	93.3%	7 terabyte data transaction	Best Schema
Fahd Idrissi Khamlichi et al (2019)	XGBoost	95%	5000 sample	Best Schema
Swetha et al (2021)	XGBoost	99.6 %	3333 numerical forms	Best Schema
Sagar et al (2022)	XGboost	97%	52,332 costumer record	Best Schema
The Result of experiment	XGboost	81.2 %	1409 datasets	Worst Schema

This indicates that XGBoost outperformed the other algorithms and is the most appropriate algorithm to use in creating the final predictive model. Additionally, a Grid Search CV hyper-parameter setting was conducted with XGBoost to further enhance accuracy, resulting in an 81.2% accuracy rate in forecasting churning behavior. The final attempt resulted in the successful creation of a highly practical predictive model. This effort provided the advantage of accurately predicting the probability of customer churn. The model can be expanded further by creating a combined churn prediction model for the telecommunications industry. It can be regarded as the most suitable prediction model and can be used freely in various other companies. It was also observed that executing hyper-parameter tuning before conducting cross-validation can improve the accuracy of ensemble techniques.

References

[1] Chen, H., Chiang, R.H., Storey, V.C. Business intelligence and analytics: From big data to big impact. MIS quarterly.2012;1165–1188. https://doi.org/10.2307/41703503

[2] Ullah I, Raza B, Malik AK, Imran M, Islam SU, Kim SW. A churn prediction model using random forest: analysis of machine learning techniques for churn prediction and factor identification in telecom sector. IEEE Access 2019;(7). https://doi.org/10.1109/access.2019.2914999

[3] Labhsetwar, S. R, Predictive analysis of customer churn in telecom industry using supervised learning. *ICTACT Journal on Soft Computing*,2020;*10*(2), 2054-2060. https://doi.org/10.21917/ijsc.2020.0291

[4] Rajamohamed R, Manokaran J. Improved credit card churn prediction based on rough clustering and supervised learning techniques. Cluster Computing 21. 2018 ;(1):65–77. https://doi.org/10.1007/s10586-017-0933-1

[5] Lalwani, P., Mishra, M. K., Chadha, J. S., & Sethi, P. *Customer churn prediction system: a machine learning approach. Computing.* 2021. https://doi.org/10.1007/s00607-021-00908-y

[6] Amin, A., Al-Obeidat, F., Shah, B., Adnan, A., Loo, J., & Anwar, S. *Customer churn prediction in telecommunication industry using data certainty. Journal of Business Research.* 2018. https://doi.org/10.1016/j.jbusres.2018.03.003

[7] Vijaya J, Sivasankar E. Improved churn prediction based on supervised and unsupervised hybrid data mining system. In: Information and Communication Technology for Sustainable Development. Singapore: Springer, 2018; 485–499. https://doi.org/10.1007/978-981-10-3932-4_51

[8] Ali M, Rehman AU, Hafeez S, Ashraf MU. Prediction of churning behavior of customers in telecom sector using supervised learning techniques. In: International Conference on Computer, Control, Electrical, and Electronics Engineering (ICCCEEE). Piscataway: IEEE, 2018; 1–6. https://doi.org/10.1109/iccceee.2018.8515857

[9] Amin A, Al-Obeidat F, Shah B, Adnan A, Loo J, Anwar S. Customer churn prediction in telecommunication industry using data certainty. Journal of Business Research 94.2019;(8):290–301. https://doi.org/10.1016/j.jbusres.2018.03.003

[10] Matloob, F., Ghazal, T. M., Taleb, N., Aftab, S., Ahmad, M., Khan, M. A., … Soomro, T. R. *Software Defect Prediction Using Ensemble Learning: A Systematic Literature Review. IEEE Access, 9,* 2021; *98754–98771.* https://doi.org/10.1109/access.2021.3095559

[11] Bilal, S. F., Almazroi, A. A., Bashir, S., Khan, F. H., & Almazroi, A. A. An ensemble-based approach using a combination of clustering and classification algorithms to enhance customer churn prediction in telecom industry. *PeerJ Computer Science*, 2022; *8*, e854. https://doi.org/10.7717/peerj-cs.854

[12] A. N. R. Moparthi and B. D. N. Geethanjali.Design and implementation of hybrid phase-based ensemble technique for defect discovery using SDLC software metrics. in Proc. 2nd Int. Conf. Adv. Electr., Electron., Inf., Commun. Bio-Inform. (AEEICB), Feb. 2016, pp. 268–274. https://doi.org/10.1109/aeeicb.2016.7538287

[13] Ahmed M, Afzal H, Siddiqi I, Amjad MF, Khurshid K. Exploring nested ensemble learners using overproduction and choose approach for churn prediction in telecom industry. Neural Computing and Applications. 2020;32(8):3237–325. https://doi.org/10.1007/s00521-018-3678-8

[14] Brownlow J, Chu C, Fu B, Xu G, Culbert B, Meng Q. Cost-sensitive churn prediction in fund management services. In: International Conference on Database Systems for Advanced Applications. Cham: Springer, 2018;776–788. https://doi.org/10.1007/978-3-319-91458-9_49

[15] S. Jhaveri, I. Khedkar, Y. Kantharia and S. Jaswal, "Success Prediction using Random Forest, CatBoost, XGBoost and AdaBoost for Kickstarter Campaigns," 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2019; 1170-3. https://doi.org/10.1109/iccmc.2019.8819828

[16] Dong, X., Yu, Z., Cao, W., Shi, Y., & Ma, Q. *A survey on ensemble learning. Frontiers of Computer Science,* 2019; *14(2):241–258.* https://doi.org/10.1007/s11704-019-8208-z

[17] V. Umayaparvathi and K. Iyakutti, "Attribute selection and Customer Churn Prediction in the telecom industry," in International Conference on Data Mining and Advanced Computing (SAPIENCE), Ernakulam, 2016; 84-90. https://doi.org/10.1109/sapience.2016.7684171

[18] O. Celik and U. O. Osmanoglu.Comparing to Techniques Used in Customer Churn Analysis.J. Multidiscip. Dev. 2019;4(1):30–38. https://doi.org/10.1109/sapience.2016.7684171

[19] Deng, Y., Li, D., Yang, L., Tang, J., & Zhao, J. *Analysis and prediction of bank user churn based*

on ensemble learning algorithm. 2021 IEEE International Conference on Power Electronics, Computer Applications (ICPECA). 2021. https://doi.org/10.1109/icpeca51329.2021.9362520

[20] Thakkar, H. K., Desai, A., Ghosh, S., Singh, P., & Sharma, G. Clairvoyant: AdaBoost with cost-enabled cost-sensitive classifier for customer churn prediction. *Computational Intelligence and Neuroscience*. 2022. https://doi.org/10.1155/2022/9028580

[21] Ahmad, A. K., Jafar, A., & Aljoumaa, K. *Customer churn prediction in telecom using machine learning in big data platform. Journal of Big Data,* 2019 *6(1).* https://doi.org/10.1186/s40537-019-0191-6

[22] S. Raschka. Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning. 2018. http://arxiv.org/abs/1811.12808. https://doi.org/10.1186/s40537-019-0191-6

[23] M. A. Alonso, D. Vilares, C. Gómez-Rodríguez & J. Vilares.Sentiment analysis for fake news detection. Electronics, 2021;10(11). https://doi.org/10.3390/electronics10111348.

[24] Senthan, P., Rathnayaka, R., Kuhaneswaran, B., & Kumara, B. *Development of Churn Prediction Model using XGBoost - Telecommunication Industry in Sri Lanka. 2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS).* 2021. https://doi.org/10.1109/iemtronics52119.2021.9422657

[25] RB, D. Customer churn prediction in telecommunication industry through machine learning based Fine-tuned XGBoost algorithm.2021. https://doi.org/10.2139/ssrn.3835039

[26] Khamlichi, F.I., Zaim, D., Khalifa, K. A new model based on global hybridization of machine learning techniques for "customer churn prediction", in: 2019 Third International Conference on Intelligent Computing in Data Sciences (ICDS), IEEE. 2019:1–4. https://doi.org/10.1109/icds47004.2019.8942240

[27] Shrestha, S. M., & Shakya, A. A Customer Churn Prediction Model using XGBoost for the Telecommunication Industry in Nepal. *Procedia Computer Science*, *215*, 652-661. 2022. https://doi.org/10.1016/j.procs.2022.12.067