Bulletin of Electrical Engineering and Informatics
Vol. 14, No. 1, February 2025, pp. 398~408
ISSN: 2302-9285, DOI: 10.11591/eei.v14i1.8112

Enhancing costumer churn prediction with stacking ensemble and stratified k-fold

Rofik, Jumanto Unjung, Budi Prasetiyo
Department of Computer Science, Faculty of Mathematics and Natural Science, Universitas Negeri Semarang, Semarang, Indonesia

Article Info

Article history:

Received Dec 28, 2023
Revised Aug 9, 2024
Accepted Aug 25, 2024

Keywords:

Churn prediction
Stratified K-fold
Synthetic minority oversampling technique
Stacking classifier
XGBoost

ABSTRACT

In the era of rapid technological advancement, the telecommunications industry undergoes significant changes. Factors such as the speed of technological change, high customer expectations, and changing preferences are the main obstacles that affect the dynamics of telecommunications companies. One major issue faced is the high customer churn rate, adversely impacting company revenue and profitability. Previous studies indicate that customer churn prediction remains complex in the telecommunications industry, with opportunities to optimize algorithm selection and prediction model construction methods. This research aims to improve the accuracy of customer churn prediction by employing a complex model that utilizes stacking ensemble learning techniques. The proposed model combines 6 base algorithms: extreme gradient boosting (XGBoost), random forest, light gradient boosting machine (LightGBM), support vector machine (SVM), K-nearest neighbor (KNN), and neural network (NN), with XGBoost as the meta-learner model. The research process involves preprocessing, class data balance with synthetic minority oversampling technique (SMOTE), training using stratified k-fold, and model evaluation. The model is tested using the Telecom Churn dataset. The evaluation results show that the constructed stacking model achieves 98% accuracy, 98.74% recall, 98.03% precision, and 98.38% F1 score. This study demonstrates that optimizing the stacking ensemble model with SMOTE and stratified k-fold enhances customer churn prediction accuracy.

Corresponding Author:

Jumanto Unjung
Department of Computer Science, Faculty of Mathematics and Natural Science
Universitas Negeri Semarang
Kampus Sekaran Gunungpati, Semarang, Central Java, Indonesia
Email: jumanto@mail.unnes.ac.id

1. INTRODUCTION

In the modern era characterized by the rapid development of information technology, the telecommunications industry has undergone similar changes [1]. Telecommunications has gone beyond its role as a tool for people to communicate with each other to become the foundation of various services and innovations that affect the way people move and work. Telecommunications can become an auxiliary factor in various industrial sectors, both directly, and indirectly [2]. It can be said that telecommunications has become the backbone of an ever-evolving digital society.

Unfortunately, to fulfill its role as the backbone of the digital society, the telecommunications industry must face great challenges. This challenge is of course the existence of business competition that is getting tighter from time to time [3]. The speed of technological change, higher customer expectations, and changing customer preferences are the main factors that cause these challenges [4]. One of the main

problems often faced in this context of adaptation is how to maintain market share and increase profitability amid this fierce competition. Of course, to maintain market share and increase profitability in this industry, it is not enough to develop a more sophisticated infrastructure. But it is also necessary to understand customer preferences well.

When companies are unable to meet the preferences and needs of their customers, they will face the existence of customers they have turned disloyal or the term customer churn. Customer churn is a term that refers to customer disloyalty to a company's products or services, where customers leave the company not to resubscribe and switch to competitors [5], [6]. This is usually because customers are not satisfied with the company's services [7] and want to improve the quality of service and the price rates of other companies [8], [9]. Losing customers, which usually provide profit for the company, certainly reduces the company's income. Furthermore, customer churn also makes the company more burdened by having to try to attract new customers [10]. The costs and efforts incurred to attract new customers tend to be greater than maintaining old customers [11], [12].

Previous research stated that the annual customer churn rate in the telecommunications industry can reach 20-40%, and the cost of acquiring new customers can be 5-10 times higher than retaining existing customers [9], [13]. Thus, it can be said that customers are a useful resource for businesses. Inevitably to maintain their business, companies must be proactive in overcoming this problem [14], [15]. Customer churn detection efforts can be carried out by identifying customers who are at high risk of moving to competitors before the event occurs [16]–[18]. Therefore, companies can know which customers should receive more attention in ways such as offering special programs [19], providing additional services, or offering special offers to maintain their loyalty to the company [15].

The churn phenomenon is influenced by several complex factors that include service usage behavior, customer satisfaction, brand interaction, and economic factors. Manually identifying early signs of churn should be very difficult, coupled with the growth in the number of customers and the complexity of existing data [20] at a fantastic speed [21]. Therefore, strategies and methods are needed that can understand and predict customer churn effectively and efficiently. To meet this challenge, many researchers have developed techniques to predict customer churn using data-driven and machine learning approaches [22]. One of the popular methods for processing data into business-critical information is data mining. Data mining is a machine learning approach that focusses on finding information from existing patterns in a set of data. The classification method is one of the techniques in data mining to make predictions from a set of classes in a dataset. Previous studies that have applied classification methods to predict classes in several industrial sectors are [23]–[27]. The application of preprocessing techniques and solving data imbalance problems [28]–[31] has also become a common approach in data mining.

Several related studies have previously performed customer churn prediction in the telecommunications industry. Some of them are as done by Ullah *et al.* [32], who used a classification and clustering approach with the information gain feature selection technique and the correlation attribute ranking filter. This study shows that the random forest classification model produces good performance, which is 88.63 for correct classification. Research by Kanwal *et al.* [21] focusses on using particle swarm optimization (PSO) as a feature selection in classifying churn customers. This research applies decision tree, K-nearest neighbor (KNN), gradient boosting, and naive bayes algorithms. This research shows the success of the gradient boosted tree algorithm with the application of PSO feature selection, which shows the greatest accuracy compared to other model implementations, which is 93%. Amin *et al.* [33] integrated the distance factor in predicting customer churn with the highest accuracy reaching 89.01%.

Research was also carried out to classify customer churn by Amin *et al.* [11], who proposed an adaptive method with a naive bayes classifier and genetic algorithm. This research was tested on 3 datasets, namely the BigML Customer Churn dataset, IBM Customer Churn, and Cell2Cell, and obtained the highest accuracy of 98.5% on the Cell2Cell dataset. Research by Cenggoro *et al.* [34] uses embedding vectors in deep learning to identify churn customers with an accuracy of 89.82%. Ahmad *et al.* [35] focused on classifying churn customers using social network analysis (SNA) in feature selection, and with this, this study was able to increase the accuracy of the model from 84% to 93%. This research examines the potential performance improvement obtained by predicting customer churn, which has not been fully explored in previous studies. Meanwhile, researchers highlight that the stacking ensemble method tends to perform well in several cases [6], [36], [37]. The synthetic minority oversampling technique (SMOTE) method, which works well in balancing data [29], [38]–[41], and the stratified k-fold method which excels in training models [42]. With the excellence of these methods, this research proposes a stacking ensemble learning model with SMOTE and stratified k-fold to enhance the performance of customer churn prediction models in the telecommunications industry, which has not been conducted by previous studies.

2. METHOD

In this research, customer churn prediction is implemented through a series of steps, including data collection, pre-processing, oversampling, modelling, and model evaluation. Details of the flow of the steps carried out in the investigation can be seen in Figure 1. The framework shown in Figure 1 was created to provide a comprehensive guide for designing a research flow to predict customer churn with optimal accuracy. The main steps taken in this research are expected to produce an effective solution, especially to overcome challenges such as the imbalance of existing data to identify churn customers in the telecommunications industry. The following are details about each of the steps taken.

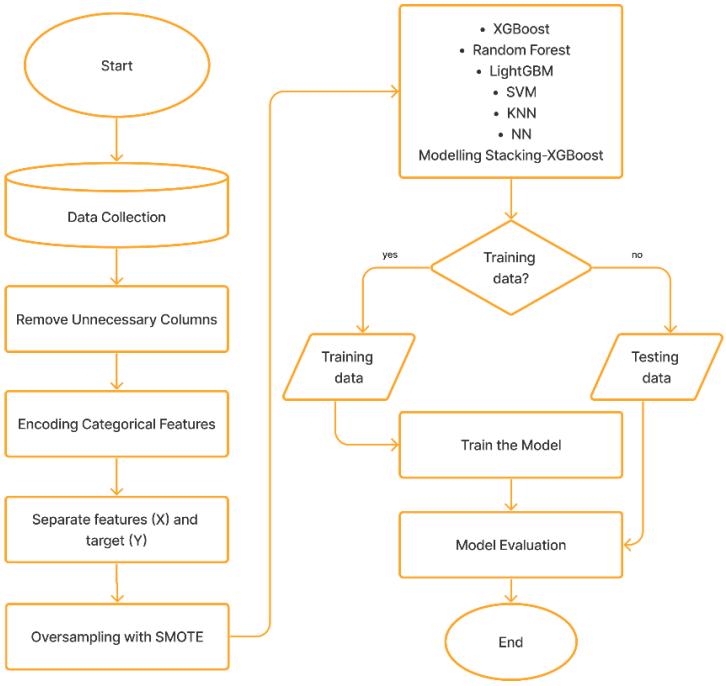

Figure 1. Workflow for customer churn prediction

2.1. Data collection

The dataset is obtained from a publicly accessible platform through Kaggle.com. This research takes the dataset, namely Telco Churn Data, which has also been used in several previous studies with a similar focus. The dataset used in this research can be accessed via the URL link: https://www.kaggle.com/code/mnassrib/customer-churn-prediction-telecom-churn-dataset/notebook. The dataset consists of 3,333 data records involving 21 features covering both churn true and churn false categories. This dataset presents very valuable information in the context of customer churn prediction research in the telecommunications industry. With a total of 2,850 customer data that do not churn and 483 customer data that churn, this dataset is expected to provide a comprehensive picture related to customer behavior and factors that influence customer decisions to churn. In Table 1 is presented a detailed explanation of the features in the dataset along with their data types.

2.2. Preprocessing

The next stage is preprocessing to prepare the data for further analysis. This research uses a Google Collaboratory platform with python programming language to process data. The preprocessing stage is performed by loading the data that have been collected previously. The data cleaning stage is carried out, first, by removing the 'phone number' feature which is considered not to provide a significant contribution to understanding customer churn. There are categorical features in the dataset, namely class, state, international plan, and voice email plane. A label encoder was applied to these features to convert them into numeric representations that can be used in modeling. This processing aims to improve the completeness and uniformity of the data, reduce the complexity, and improve the readability of the data for the model to be applied. Separation was also performed between the variable X (independent variable) and its class Y (dependent variable).

Table 1. Description of features in the dataset and their data types

No	Features	Description	Data type
1	State	Name of the state	String
2	Account_length	Number of days a customer has been a customer	Integer
3	Area code	Customer area code	Integer
4	Phone number	Customer's phone number	String
5	International plan	Customer status having or not having an international customer package (yes/no)	String
6	Voice mail plan	Customer status of having or not having a voicemail package (yes/no)	String
7	Number vmail messages	The number of voicemail messages received by the customer	Integer
8	Total day minutes	Total minutes of the customer during daylight hours	Float
9	Total day calls	Number of calls made by the customer during daylight hours	Integer
10	Total day charge	Total cost for calls during daylight hours	Float
11	Total eve minutes	Minutes of calls during the afternoon	Float
12	Total eve calls	Number of calls made by the customer during the afternoon	Integer
13	Total eve charge	Total cost for calls during the afternoon	Float
14	Total night minutes	Total minutes of calls during the night	Float
15	Total night calls	Number of calls made by the customer during the night	Integer
16	Total night charge	Total cost for calls during the night	Float
17	Total intl minutes	Total minutes of international calls	Float
18	Total intl calls	Number of international calls made by the customer	Integer
19	Total intl Charge	Total cost for international calls	Float
20	Customer service calls	Number of customer service calls made by the customer	Integer
21	*churn*	Customer churn status (true/false)	Boolean

2.3. Oversampling

Because the dataset has a class imbalance, where the number of customers who churn is less than that of customers who do not churn, this research applies an oversampling technique using SMOTE [38]. This method generates a synthesized sample of the minority class using an interpolation method to create data between the selected point and its closest data [43], thus creating a balance between the two classes. This step is done to avoid biases that may arise in modeling due to data imbalance. This method is also applied to create a balance so that the model can learn well from both classes. By (1) this is the data augmentation formula using SMOTE:

$$Xsyn = Xi + rand(0,1) \times |Xi - Xneighbour| \quad (1)$$

here, 'Xi' refers to an example from the minority class. The 'Xneighbour' is a randomly selected sample from the nearest neighbor. And 'rand(0,1)' is a random number that falls in the range between 0 and 1. After the implementation of this data balancing method using SMOTE, the data are now balanced between the positive churn and negative churn classes, which are 2,850 data each.

2.4. Split data

At the split data stage, this research divides the data into two important parts, namely training data and test data. Training data are used to train basic models and metalearner models, while test data are used to test the performance of the models that have been built. The division process is carried out with a proportion of 80% of the data as training data and the remaining 20% as test data. With the use of appropriate data division, this research can perform objective tests and produce reliable predictions. Therefore, this data-sharing stage is an important step to validate the performance of models in predicting customer churn in the telecommunications industry.

2.5. Modeling

At this stage, the basic models are built. This research develops a variety of classification algorithms that will act as important components in the stacking approach. Some of the basic models chosen include extreme gradient boosting (XGBoost), random forest, light gradient boosting machine (LightGBM), support vector machine (SVM), KNN, and neural network (NN). Each of these models has unique characteristics that contribute to the diversity of modeling. XGBoost is used for its ability to handle complex classification problems [44]. Random forest is a robust choice for tree-based classifiers [45]–[47]. LightGBM is used because it is known for its high performance and ability to process large data [48], [49]. SVM is used to handle non-linear classification problems [41], [50]. KNN is the choice for classification due to its ability to be based on neighborliness [51], [52]. A NN is used because it allows the model to handle deep learning [53].

Each base model is trained with preprocessed data, oversampled using SMOTE, and the results are used as contributions to a more robust ensemble model. By combining these various baseline models, this research seeks to create an ensemble model that can combine the advantages of each algorithm and produce more accurate predictions in the modeling [54], [55] customer churn prediction. The second level involves a

meta-learner model that takes the predictions of the first level as its input features [54]. The meta-learner model, in the context of customer churn prediction, is XGBoost. This meta-learner is then able to understand how to combine these predictions to produce a more accurate final prediction. XGBoost was chosen as the meta-learner because it is one of the most effective ensemble algorithms that can handle various classification problems with a high level of accuracy, and it also performed well on the stacking model created in previous research [29].

The stacking model training process is carried out using training data and is carried out with the k-fold stratified method. This is the concept of the cross-validation method, which divides the dataset into several subsets with fair class sizes [56]. Each of these sub-datasets acts as a validation dataset in turn, while the others are used as training datasets. In this process, K models will be built and trained with different training datasets and then tested with the validation dataset. Figure 2 shows how cross-validation works. The stratified K-fold is used to avoid and help mitigate the risk of bias, where the model may depend only on part of the data [57].

Figure 2. Cross-validation concept

2.6. Evaluation

Model evaluation is a critical step in this research to measure the performance of the developed ensemble model. Several evaluation metrics are used to represent the performance of the ensemble model. The evaluation metrics used are accuracy, precision, recall, and F1-score. Accuracy measures the degree to which the model can predict correctly. Precision measures the degree to which the positive predictions made by the model are correct. Meanwhile, recall measures the extent to which the model can detect customers who really churn. The model performance measure is obtained from the confusion matrix. The confusion matrix table can be seen in Table 2.

Table 2. Confusion matrix

Predictive values/actual values		Actual values 1	Actual values 0
Predictive Values	1	TP	FP
	0	FN	TN

The confusion matrix is a table that is used to visually analyze the extent to which the model can predict the classification of the data. This matrix consists of four main cells, namely true positive (TP), false positive (FP), true negative (TN), and false negative (FN). The TP reflects the amount of data that is correctly classified as positive by the model. FP represents the amount of data that was mistakenly classified as positive by the model, despite being negative. The TN represents the amount of data that was correctly classified as negative by the model. In contrast, FN records the amount of data that was incorrectly classified as negative by the model, when it should have been considered positive. With this, we also get parameters that show how the model has performed. Through the measurement of the parameters (2) to (5):

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)} \times 100\% \quad (2)$$

$$Precision = \frac{TP}{(TP+FP)} \times 100\% \quad (3)$$

$$Recall = \frac{TP}{(TP+FN)} \times 100\% \quad (4)$$

$$F1 - Score = \frac{2*(Presisi*Recall)}{(Presisi+Recall)} \times 100\% \quad (5)$$

with the confusion matrix and the calculation formula for each parameter, a value between 0% and 100% is obtained, which represents the value of how the model can correctly predict churn customers and those who do not churn. The higher the value obtained from each parameter, the better the performance of the classification model. Vice versa, if the value shown is close to 0%, it can be concluded that the model is still underperforming and there are likely to be deficiencies in the ability to perform proper classification.

3. RESULTS AND DISCUSSION

To further analyze the data, an exploratory data analysis (EDA) was performed. Figure 3 shows the data distribution of each feature of the dataset. From the data distribution image for each feature, it can be said that most features have a normal distribution, namely for the state distribution feature account length distribution, total day minutes distribution, total day calls distribution, total day charge distribution, total eve minutes distribution, total eve calls distribution, total eve charge distribution, total night minutes distribution, total night calls distribution, total night charge distribution, intl minutes distribution, and total intl charge distribution. Where the data spreads normally in that range. But some distributions are skewed to the right, where the data are spread over the minimum values. That is, in the total distribution 'intl calss' and customer service calls distribution features. Most people do not make much use of that feature. Most customers also do not use v-mail messages. Likewise, the utilization of international plans and voice mail plans, which customers rarely use. The distribution of area codes is mostly below the number 420, for the rest it is more than area code 500.

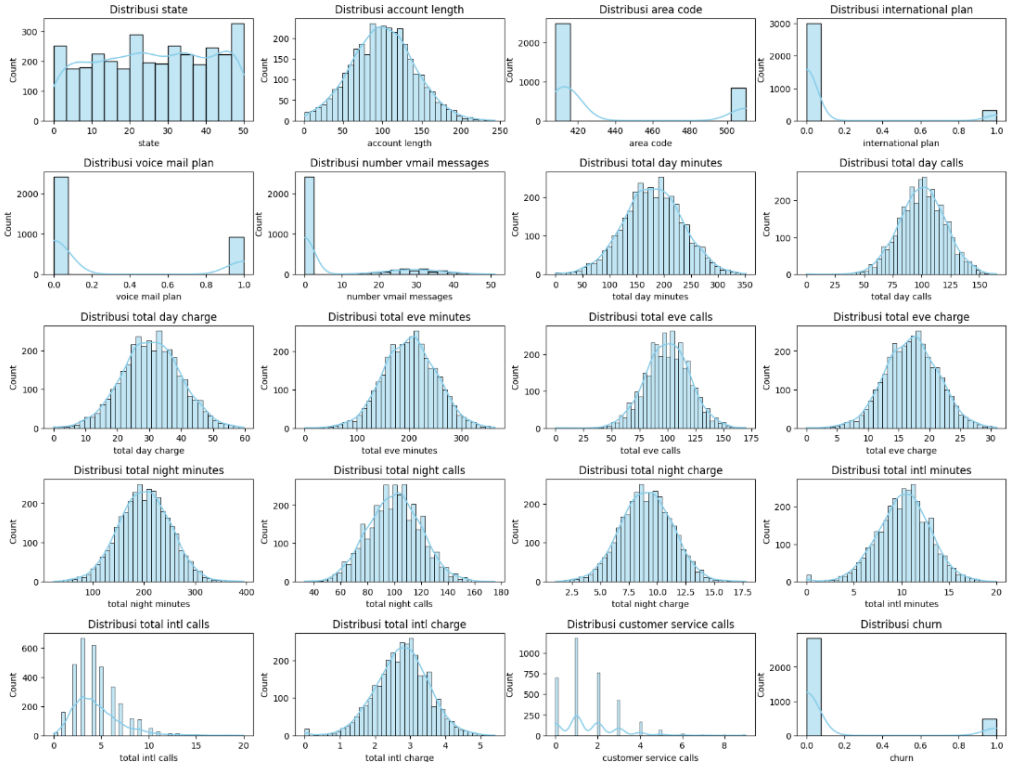

Figure 3. Thelco churn dataset EDA framework

When entering the preprocessing stage, it turns out that the data do not have NaN or duplicates, so no data reduction is done. Therefore, the original amount of data was used. The data consist of 3333 data records, of which 2,850 customers do not churn, while 483 are customers who churn in this telecommunications industry. Of course, if directly entered into modeling, it will make the model too inclined

to the class of data that does not churn. And this will result in misclassification. So, class balancing is done using SMOTE. This implementation can balance the data in each class. It works because the way it works is to synthesize the minority class. Figure 4 shows the data distribution between the churn class (true) and the no-churn class (false) before and after SMOTE.

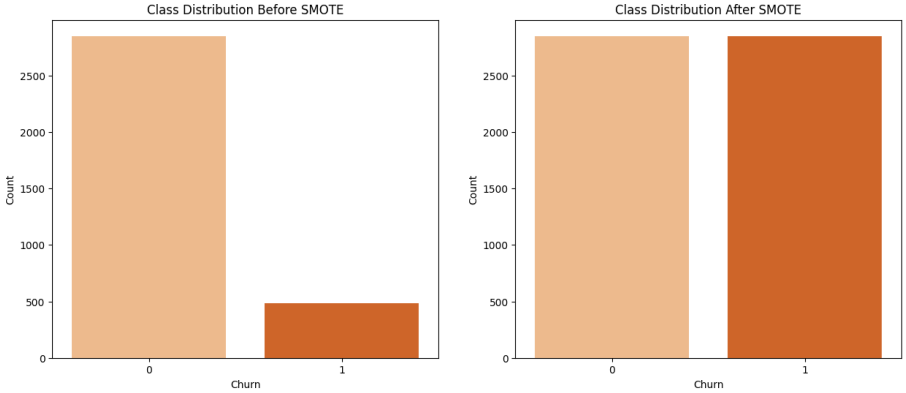

Figure 4. Data distribution before SMOTE and data distribution after SMOTE

According to Figure 4 illustrates the increase in the amount of data in the true (churn) class after SMOTE. Now the data distribution between the churn and non-churn classes is 2,850 data each. Therefore, this amount of data distribution will help the model to optimize its performance. To see the correlation between features and the correlation between features and their targets, a heat map is created. The heatmap showing the correlation between features can be seen in Figure 5.

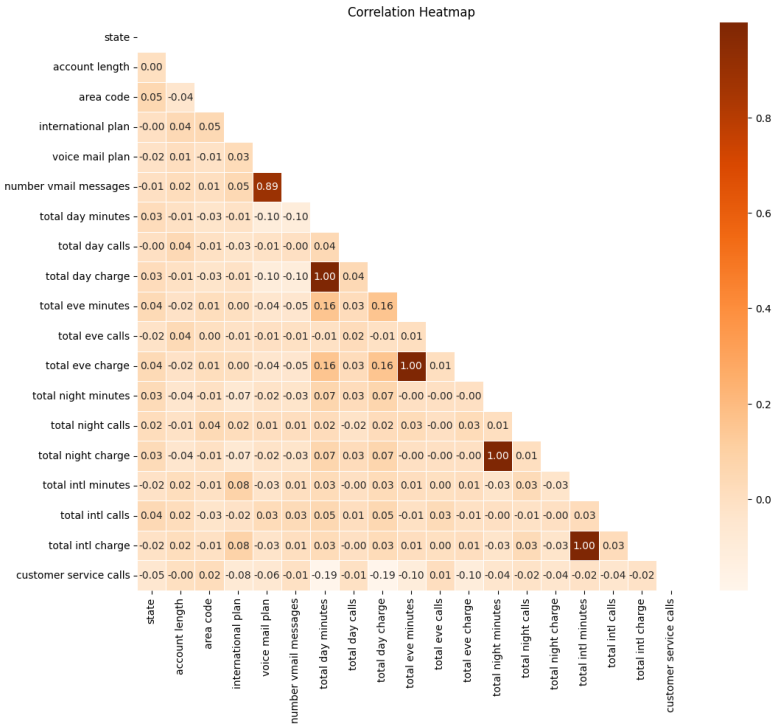

Figure 5. Correlation between features

The correlation approaching a value of 1 indicates that the feature has a strong correlation and the correlation is positive or direct. In the heatmap, this is depicted by colors that are approaching red. On the

other hand, negative correlation, which is an inverse correlation, is represented by colors approaching white, with values close to -1. From the heatmap, it can be concluded that features strongly correlated with the target are international plan, total day minutes, and total day charge. The larger these features, the higher the likelihood of customer churn. On the contrary, features with negative correlation (having an inverse relationship) are voice mail plan, number of vmail messages, and total int calls. The lower the values of these features, the higher the likelihood of customer churn.

The modeling is performed using the stacking ensemble learning method, which combines six basic algorithms: XGBoost, random forest, LightGBM, SVM, KNN, and NN. This is done to optimize the performance generated by combining the results of the decisions of these 6 algorithms, each with its own strengths and weaknesses. For the final estimator, the XGBoost algorithm is used due to its popularity in achieving optimal performance in various cases. Figure 6 illustrates the construction of the modeling conducted in this research to detect customers who churn and those who do not.

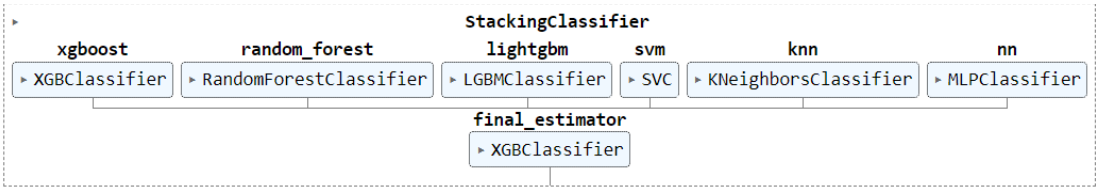

Figure 6. Depiction of how the stacking model works

The data was trained using the training data with the stratified k-fold method. In the cross-validation tuning, n was set to 5 folds, where the training data was divided into 5 parts of equal size, with the number of each class balanced in each part or fold. This was done to prevent the model from performing well only on a few specific data points. Subsequently, during model testing, the created model was tested to predict customers using testing data, which are data that the model has not seen or known before. The model was evaluated using a confusion matrix which can show the number of instances where the model correctly or incorrectly predicted the results. Figure 7 illustrates the number of correctly and incorrectly predicted data points, including instances where the actual outcome is true and predicted correctly, as well as cases where the actual outcome is false but predicted correctly and vice versa. From the confusion matrix, performance metrics of the model can be derived, such as accuracy, precision, recall, and F1-score.

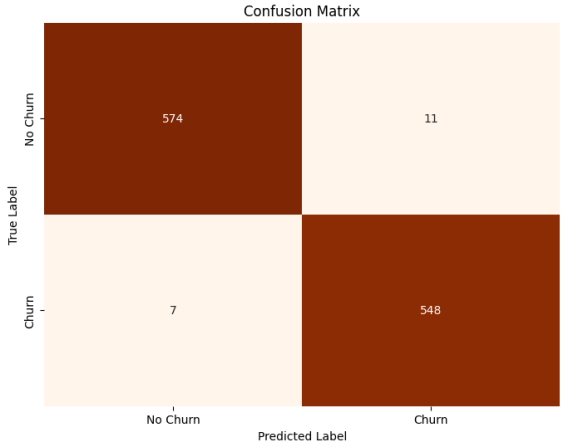

Figure 7. Confusion metric of stacking ensemble model

The constructed ensemble model demonstrates optimal performance, with a precision metric value of 98%, a recall of 98.74%, precision of 98.03%, and an F1-score of 98.38%. Our findings indicate that the stacking ensemble model with the combination of algorithms used along with the implementation of SMOTE and the stratified k-fold method works effectively with the data utilized. The proposed methods in this study tend to exhibit significantly better metric proportions compared to previous research. The model developed in this study successfully optimizes performance in predicting customer churn and non-churn. In terms of

accuracy, precision, recall, and F1-score, the model also effectively distinguishes between churn and non-churn classes.

The implemented SMOTE method also aids in the data synthesis without negatively impacting the classification of classes in the synthesized data. The training process using stratified k-fold does not significantly alter the resulting performance. This study constructs a complex stacking ensemble model and tests it extensively. However, further in-depth research may be necessary to confirm that the model can also perform well on new data. It is important to note that optimal results may be influenced by the relatively small size of the dataset, especially considering the original data has minimal examples for churn class. Subsequent research is recommended to confirm the performance of the model created in this study using larger and more varied datasets.

4. CONCLUSION

In this study, classification was performed to identify potential churn or retained subscribers in the telecommunications service using a stacking ensemble learning approach. The model was built with six basic algorithms: XGBoost, random forest, LightGBM, SVM, KNN, and NN. To address sample imbalance between churn and non-churn classes, the SMOTE oversampling technique was applied. Training utilized cross-validation, specifically the stratified K-fold technique, to prevent the model from excelling in classifying churn and non-churn customers on only certain portions of the data. Evaluation results showed that the stacking ensemble learning model successfully enhanced performance in distinguishing between churn and non-churn customers. The model achieved an accuracy of 98% accuracy, 98.74% recall, 98.03% precision, and 98.38% F1-score. This finding supports the effectiveness of the ensemble learning approach with multiple methods implemented to address classification problems, particularly in predicting potential churn customers in the telecommunications service industry. For future research, it is recommended to explore larger datasets and ensure model optimization for more diverse data.

ACKNOWLEDGESMENTS

Authors would like to acknowledge the support of the Checklist for budgetary execution (DPA) Faculty of Mathematics and Natural Sciences (FMIPA), Universitas Negeri Semarang, Indonesia for this research through a grant (grant number. DPA 107.28.3/UN37/PPK.04/2024).

REFERENCES

[1] S. Maldonado, J. López, and C. Vairetti, "Profit-based churn prediction based on minimax probability machines," *European Journal of Operational Research*, vol. 284, no. 1, pp. 273–284, Jul. 2020, doi: 10.1016/j.ejor.2019.12.007.

[2] P. Maneejuk and W. Yamaka, "An analysis of the impacts of telecommunications technology and innovation on economic growth," *Telecommunications Policy*, vol. 44, no. 10, pp. 1–19, Nov. 2020, doi: 10.1016/j.telpol.2020.102038.

[3] K. Coussement, S. Lessmann, and G. Verstraeten, "A comparative analysis of data preparation algorithms for customer churn prediction: A case study in the telecommunication industry," *Decision Support Systems*, vol. 95, pp. 27–36, Mar. 2017, doi: 10.1016/j.dss.2016.11.007.

[4] H. Jain, A. Khunteta, and S. Srivastava, "Churn prediction in telecommunication using logistic regression and logit boost," *Procedia Computer Science*, vol. 167, pp. 101–112, 2020, doi: 10.1016/j.procs.2020.03.187.

[5] E. Zdravevski, P. Lameski, C. Apanowicz, and D. Ślęzak, "From big data to business analytics: the case study of churn prediction," *Applied Soft Computing Journal*, vol. 90, pp. 1–15, May 2020, doi: 10.1016/j.asoc.2020.106164.

[6] T. Xu, Y. Ma, and K. Kim, "Telecom churn prediction system based on ensemble learning using feature grouping," *Applied Sciences*, vol. 11, no. 11, pp. 1–12, 2021, 10.3390/app11114742.

[7] K. Ljubičić, A. Merćep, and Z. Kostanjčar, "Churn prediction methods based on mutual customer interdependence," *Journal of Computational Science*, vol. 67, p. 101940, Mar. 2023, doi: 10.1016/j.jocs.2022.101940.

[8] U. J, N. Metawa, K. Shankar, and S. K. Lakshmanaprabu, "Financial crisis prediction model using ant colony optimization," *International Journal of Information Management*, vol. 50, pp. 538–556, Feb. 2020, doi: 10.1016/j.ijinfomgt.2018.12.001.

[9] N. N. Y. Vo, S. Liu, X. Li, and G. Xu, "Leveraging unstructured call log data for customer churn prediction," *Knowledge-Based Systems*, vol. 212, p. 106586, Jan. 2021, doi: 10.1016/j.knosys.2020.106586.

[10] E. Sivasankar and J. Vijaya, "Hybrid PPFCM-ANN model: an efficient system for customer churn prediction through probabilistic possibilistic fuzzy clustering and artificial neural network," *Neural Computing and Applications*, vol. 31, no. 11, pp. 7181–7200, Nov. 2019, doi: 10.1007/s00521-018-3548-4.

[11] A. Amin, A. Adnan, and S. Anwar, "An adaptive learning approach for customer churn prediction in the telecommunication industry using evolutionary computation and Naïve Bayes," *Applied Soft Computing*, vol. 137, p. 110103, Apr. 2023, doi: 10.1016/j.asoc.2023.110103.

[12] A. Alamsyah *et al.*, "Customer segmentation using the integration of the recency frequency monetary model and the k-means cluster algorithm," *Scientific Journal of Informatics*, vol. 9, no. 2, pp. 189–196, Nov. 2022, doi: 10.15294/sji.v9i2.39437.

[13] A. Amin *et al.*, "Cross-company customer churn prediction in telecommunication: a comparison of data transformation methods," *International Journal of Information Management*, vol. 46, pp. 304–319, Jun. 2019, doi: 10.1016/j.ijinfomgt.2018.08.015.

[14] S. A. Panimalar and A. Krishnakumar, "Customer churn prediction model in cloud environment using DFE-WUNB: ANN deep feature extraction with weight updated tuned Naïve Bayes classification with Block-Jacobi SVD dimensionality reduction," *Engineering Applications of Artificial Intelligence*, vol. 126, p. 107015, Nov. 2023, doi: 10.1016/j.engappai.2023.107015.

[15] B. Prabadevi, R. Shalini, and B. R. Kavitha, "Customer churning analysis using machine learning algorithms," *International Journal of Intelligent Networks*, vol. 4, pp. 145–154, 2023, doi: 10.1016/j.ijin.2023.05.005.
[16] K. W. De Bock and A. De Caigny, "Spline-rule ensemble classifiers with structured sparsity regularization for interpretable customer churn modeling," *Decision Support Systems*, vol. 150, pp. 1–14, Nov. 2021, doi: 10.1016/j.dss.2021.113523.
[17] P. Kate, V. Ravi, and A. Gangwar, "FinGAN: chaotic generative adversarial network for analytical customer relationship management in banking and insurance," *Neural Computing and Applications*, vol. 35, no. 8, pp. 6015–6028, Mar. 2023, doi: 10.1007/s00521-022-07968-x.
[18] L. McMillan and L. Varga, "A review of the use of artificial intelligence methods in infrastructure systems," *Engineering Applications of Artificial Intelligence*, vol. 116, pp. 1-21, Nov. 2022, doi: 10.1016/j.engappai.2022.105472.
[19] S. Baghla and G. Gupta, "Performance evaluation of various classification techniques for customer churn prediction in e-commerce," *Microprocessors and Microsystems*, vol. 94, p. 104680, Oct. 2022, doi: 10.1016/j.micpro.2022.104680.
[20] Z. P. Agusta and Adiwijaya, "Modified balanced random forest for improving imbalanced data prediction," *International Journal of Advances in Intelligent Informatics*, vol. 5, no. 1, pp. 58–65, Dec. 2019, doi: 10.26555/ijain.v5i1.255.
[21] S. Kanwal *et al.*, "An attribute weight estimation using particle swarm optimization and machine learning approaches for customer churn prediction," in *4th International Conference on Innovative Computing, ICIC 2021*, IEEE, Nov. 2021, pp. 1–6. doi: 10.1109/ICIC53490.2021.9693040.
[22] M. T. Vo, A. H. Vo, T. Nguyen, R. Sharma, and T. Le, "Dealing with the class imbalance problem in the detection of fake job descriptions," *Computers, Materials and Continua*, vol. 68, no. 1, pp. 521–535, 2021, doi: 10.32604/cmc.2021.015645.
[23] M. Z. Abedin, P. Hajek, T. Sharif, M. S. Satu, and M. I. Khan, "Modelling bank customer behaviour using feature engineering and classification techniques," *Research in International Business and Finance*, vol. 65, pp. 1-16, Apr. 2023, doi: 10.1016/j.ribaf.2023.101913.
[24] C. A. R. Pinheiro, M. Galati, N. Summerville, and M. Lambrecht, "Using network analysis and machine learning to identify virus spread trends in COVID-19," *Big Data Research*, vol. 25, pp. 1–9, Jul. 2021, doi: 10.1016/j.bdr.2021.100242.
[25] D. K. Sharma, S. Lohana, S. Arora, A. Dixit, M. Tiwari, and T. Tiwari, "E-commerce product comparison portal for classification of customer data based on data mining," *Materials Today: Proceedings*, vol. 51, pp. 166–171, 2021, doi: 10.1016/j.matpr.2021.05.068.
[26] D. Wu, Q. Wang, and D. L. Olson, "Industry classification based on supply chain network information using graph neural networks," *Applied Soft Computing*, vol. 132, p. 109849, Jan. 2023, doi: 10.1016/j.asoc.2022.109849.
[27] A. A. Nurdin, G. N. Salmi, K. Sentosa, A. R. Wijayanti, and A. Prasetya, "Utilization of business intelligence in sales information systems," *Journal of Information System Exploration and Research*, vol. 1, no. 1, pp. 39–48, Dec. 2022, doi: 10.52465/joiser.v1i1.101.
[28] A. Arafa, N. El-Fishawy, M. Badawy, and M. Radad, "RN-SMOTE: reduced noise SMOTE based on DBSCAN for enhancing imbalanced data classification," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 8, pp. 5059–5074, Sep. 2022, doi: 10.1016/j.jksuci.2022.06.005.
[29] M. A. Muslim *et al.*, "New model combination meta-learner to improve accuracy prediction P2P lending with stacking ensemble learning," *Intelligent Systems with Applications*, vol. 18, pp. 1-8, May 2023, doi: 10.1016/j.iswa.2023.200204.
[30] C. O. Sakar, S. O. Polat, M. Katircioglu, and Y. Kastro, "Real-time prediction of online shoppers' purchasing intention using multilayer perceptron and LSTM recurrent neural networks," *Neural Computing and Applications*, vol. 31, no. 10, pp. 6893–6908, Oct. 2019, doi: 10.1007/s00521-018-3523-0.
[31] A. U. Dullah, F. N. Apsari, and J. Jumanto, "Ensemble learning technique to improve breast cancer classification model," *Journal of Soft Computing Exploration*, vol. 4, no. 2, Jun. 2023, doi: 10.52465/joscex.v4i2.166.
[32] I. Ullah, B. Raza, A. K. Malik, M. Imran, S. U. Islam, and S. W. Kim, "A churn prediction model using random forest: analysis of machine learning techniques for churn prediction and factor identification in telecom sector," *IEEE Access*, vol. 7, pp. 60134–60149, 2019, doi: 10.1109/ACCESS.2019.2914999.
[33] A. Amin, F. Al-Obeidat, B. Shah, A. Adnan, J. Loo, and S. Anwar, "Customer churn prediction in telecommunication industry using data certainty," *Journal of Business Research*, vol. 94, pp. 290–301, Jan. 2019, doi: 10.1016/j.jbusres.2018.03.003.
[34] T. W. Cenggoro, R. A. Wirastari, E. Rudianto, M. I. Mohadi, D. Ratj, and B. Pardamean, "Deep learning as a vector embedding model for customer churn," *Procedia Computer Science*, vol. 179, pp. 624–631, 2021, doi: 10.1016/j.procs.2021.01.048.
[35] A. K. Ahmad, A. Jafar, and K. Aljoumaa, "Customer churn prediction in telecom using machine learning in big data platform," *Journal of Big Data*, vol. 6, no. 1, pp. 1–24, Dec. 2019, doi: 10.1186/s40537-019-0191-6.
[36] Y. Liu, Y. Li, X. Tan, P. Wang, and Y. Zhang, "Local discriminant preservation projection embedded ensemble learning based dimensionality reduction of speech data of Parkinson's disease," *Biomedical Signal Processing and Control*, vol. 63, pp. 1–13, Jan. 2021, doi: 10.1016/j.bspc.2020.102165.
[37] M. A. Muslim *et al.*, "An ensemble stacking algorithm to improve model accuracy in bankruptcy prediction," *Journal of Data Science and Intelligent Systems*, vol. 2, no. 2, pp. 79–86, Mar. 2023, doi: 10.47852/bonviewjdsis3202655.
[38] A. R. Safitri and M. A. Muslim, "Improved accuracy of Naive Bayes classifier for determination of customer churn uses SMOTE and genetic algorithms," *Journal of Soft Computing Exploration*, vol. 1, no. 1, pp. 70–75, Sep. 2020, doi: 10.52465/joscex.v1i1.5.
[39] A. Fernández, S. García, F. Herrera, and N. V. Chawla, "SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary," *Journal of Artificial Intelligence Research*, vol. 61, pp. 863–905, Apr. 2018, doi: 10.1613/jair.1.11192.
[40] A. Imakura, M. Kihira, Y. Okada, and T. Sakurai, "Another use of SMOTE for interpretable data collaboration analysis," *Expert Systems with Applications*, vol. 228, p. 120385, Oct. 2023, doi: 10.1016/j.eswa.2023.120385.
[41] Jumanto *et al.*, "Optimizing support vector machine performance for parkinson's disease diagnosis using GridSearchCV and PCA-based feature extraction," *Journal of Information Systems Engineering and Business Intelligence*, vol. 10, no. 1, pp. 38–50, 2024, doi: 10.20473/jisebi.10.1.38-50.
[42] T. R. Mahesh, K. V. Vinoth, K. V. Dhilip, O. Geman, M. Margala, and M. Guduri, "The stratified k-folds cross-validation and class-balancing methods with high-performance ensemble classifiers for breast cancer classification," *Healthcare Analytics*, vol. 4, pp. 1-10, Dec. 2023, doi: 10.1016/j.health.2023.100247.
[43] G. Kovács, "An empirical comparison and evaluation of minority oversampling techniques on a large number of imbalanced datasets," *Applied Soft Computing Journal*, vol. 83, pp. 1–13, Oct. 2019, doi: 10.1016/j.asoc.2019.105662.
[44] S. K. Kiangala and Z. Wang, "An effective adaptive customization framework for small manufacturing plants using extreme gradient boosting-XGBoost and random forest ensemble learning algorithms in an Industry 4.0 environment," *Machine Learning with Applications*, vol. 4, pp. 1–15, Jun. 2021, doi: 10.1016/j.mlwa.2021.100024.
[45] S. F. Sabbeh, "Machine-learning techniques for customer retention: a comparative study," *International Journal of Advanced Computer Science and Applications*, vol. 9, no. 2, pp. 273–281, 2018, doi: 10.14569/IJACSA.2018.090238.
[46] S. Stehani, N. Karunya, D. R. J. B. Ranjan, S. Sumathipala, and T. C. Sandanayake, "Customer churn reasoning in telecommunication domain," in *Proceedings of International Conference on Image Processing and Robotics, ICIPRoB 2020*,

IEEE, Mar. 2020, pp. 1–5. doi: 10.1109/ICIP48927.2020.9367342.
[47] A. F. Mulyana, W. Puspita, and J. Jumanto, "Increased accuracy in predicting student academic performance using random forest classifier," *Journal of Student Research Exploration*, vol. 1, no. 2, pp. 94–103, Jul. 2023, doi: 10.52465/josre.v1i2.169.
[48] W. Liu, H. Fan, M. Xia, and M. Xia, "A focal-aware cost-sensitive boosted tree for imbalanced credit scoring," *Expert Systems with Applications*, vol. 208, p. 118158, Dec. 2022, doi: 10.1016/j.eswa.2022.118158.
[49] Y. Dasril, M. A. Muslim, M. F. Al Hakim, Jumanto, and B. Prasetiyo, "Credit risk assessment in P2P lending using LightGBM and particle swarm optimization," *Register: Jurnal Ilmiah Teknologi Sistem Informasi*, vol. 9, no. 1, pp. 18–28, Feb. 2023, doi: 10.26594/register.v9i1.3060.
[50] R. Rofik, R. A. Hakim, J. Unjung, B. Prasetiyo, and M. A. Muslim, "Optimization of SVM and gradient boosting models using GridSearchCV in detecting fake job postings," *MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 23, no. 2, pp. 419–430, Mar. 2024, doi: 10.30812/matrik.v23i2.3566.
[51] A. Huang, R. Xu, Y. Chen, and M. Guo, "Research on multi-label user classification of social media based on ML-KNN algorithm," *Technological Forecasting and Social Change*, vol. 188, pp. 1-10, Mar. 2023, doi: 10.1016/j.techfore.2022.122271.
[52] L. Wang, "Imbalanced credit risk prediction based on SMOTE and multi-kernel FCM improved by particle swarm optimization," *Applied Soft Computing*, vol. 114, p. 108153, Jan. 2022, doi: 10.1016/j.asoc.2021.108153.
[53] R. K. Verma, L. Kaur, and N. Kaur, "Review on application areas of deep learning," *Advances in Mathematics: Scientific Journal*, vol. 10, no. 12, pp. 3725–3731, Dec. 2021, doi: 10.37418/amsj.10.12.12.
[54] Y. He, J. Xiao, X. An, C. Cao, and J. Xiao, "Short-term power load probability density forecasting based on GLRQ-Stacking ensemble learning method," *International Journal of Electrical Power and Energy Systems*, vol. 142, p. 108243, Nov. 2022, doi: 10.1016/j.ijepes.2022.108243.
[55] E. K. Sahin and S. Demir, "Greedy-AutoML: A novel greedy-based stacking ensemble learning framework for assessing soil liquefaction potential," *Engineering Applications of Artificial Intelligence*, vol. 119, p. 105732, Mar. 2023, doi: 10.1016/j.engappai.2022.105732.
[56] T. Yan, S. L. Shen, A. Zhou, and X. Chen, "Prediction of geological characteristics from shield operational parameters by integrating grid search and K-fold cross validation into stacking classification algorithm," *Journal of Rock Mechanics and Geotechnical Engineering*, vol. 14, no. 4, pp. 1292–1303, Aug. 2022, doi: 10.1016/j.jrmge.2022.03.002.
[57] L. Dora, S. Agrawal, R. Panda, and A. Abraham, "Nested cross-validation based adaptive sparse representation algorithm and its application to pathological brain classification," *Expert Systems with Applications*, vol. 114, pp. 313–321, Dec. 2018, doi: 10.1016/j.eswa.2018.07.039.

BIOGRAPHIES OF AUTHORS

Rofik currently a student in Engineering Informatics Study Program, Department of Computer Science, Faculty of Mathematics and Natural Sciences, Universitas Negeri Semarang, Indonesia. Her research interests are in computational mathematics, machine learning, and artificial intelligence. She can be contacted at email: rofikn4291@students.unnes.ac.id.

Jumanto Unjung received his master degree in computer science and electronics from Universitas Gadjah Mada, Indonesia. He currently is an Assistant Professors in Department of Computer Science, Faculty of Mathematics and Natural Sciences, Universitas Negeri Semarang, Indonesia. He has a research interest in pattern recognition, image processing, machine learning, and artificial intelligence applications. He is a fellow member Indonesian Association for Pattern Recognition (INAPR). He can be contacted at email: jumanto@mail.unnes.ac.id.

Budi Prasetiyo is currently a lecturer in the Department of Computer Science at Universitas Negeri Semarang (UNNES), Indonesia. He received his bachelor's degree in mathematics from UNNES. Then, he obtained his master's degree in computer science from Universitas Diponegoro, Indonesia. He is currently pursuing a Ph.D. degree in information system technology at Universitas Diponegoro. He has 8 years of teaching and research experience. He published 25 research papers in reputable journals and conferences. He is a fellow member IEEE Indonesia Section. His research interests are in data mining, computational mathematics, machine learning, and artificial intelligence. He can be contacted at email: bprasetiyo@mail.unnes.ac.id.