International Journal of Technology 12(5) 1030-1037 (2021)
Received August 2021 / Revised October 2021 / Accepted November 2021

International Journal of Technology

http://ijtech.eng.ui.ac.id

Predicting Customer Churn using ensemble learning: Case Study of a Fixed Broadband Company

Arian Dhini[1*], Muhammad Fauzan[1]

[1]*Department of Industrial Engineering, Faculty of Engineering, Universitas Indonesia, Kampus UI Depok, Depok 16424, Indonesia*

Abstract. Technology advancement has developed a shift perception towards better service from internet providers, and the power to move easily to another provider to secure improved quality results in customer churn. Internet service providers must detect the risk of churn at the earliest opportunity if they want to retain their customers. This study aimed to predict churn using recent developments in machine learning approaches, and customer data from one of the biggest fixed broadband companies in Indonesia was selected as a case study. Ensemble learning is the collaboration of meta-algorithms to improve model performance, and two such approaches were performed in this study, namely random forest and extreme gradient boosting (XGBoost). The results show that the ensemble learning models outperform classical technique and XGBoost is the best algorithm for predicting customer churn. Customers are thereby clustered as being at high, medium, or low risk of churn, and the company can specify particular retention strategies towards each customer cluster.

Keywords: Customer churn prediction; Ensemble learning; Fixed broadband, Random Forest (RF); Extreme gradient boosting (XGBoost)

1. Introduction

The internet has become a necessity in the daily activities of modern life. The number of fixed broadband subscribers in Indonesia has proliferated in recent years, with an average annual increase of 2.8% (Statista, 2019), and the global volume of fixed broadband-based internet usage has itself risen by a yearly average of 6.8% (World Bank, 2020). The provision of high-speed internet connections will be an essential step towards achieving Industry 4.0 in terms of boosting productivity and economic performance (Sarachuk and Mißler-Behr, 2020). The emergence of these increasingly demanding internet needs has been leveraging competition among internet providers, and this high level of competition has encouraged companies to provide the best possible service (Huang et al., 2009; Khan et al., 2010; Qureshi et al., 2013; Do et al., 2017).

To sustain sales performance, an internet service provider must acquire and retain customers. Based on data in 2020, from one of Indonesia's largest fixed broadband companies, the costs incurred through acquisition exceed those of retention, and a monthly customer churn rate of 8.2% had affected the company's growth. Churn is often due to customer dissatisfaction, with higher prices, low service quality, incomplete features, and

*Corresponding author's email: arian@ie.ui.ac.id, Tel.: +62-21-78888805; Fax: +62-21-78885656
doi: 10.14716/ijtech.v12i5.5223

privacy concerns being some of the reasons that motivate customers to switch providers (Amin et al., 2019).

Different ways to reduce or manage churn rate have been reported, including investing in retention activities, targeted marketing, campaign management, and customer relationship management (CRM) (Chen and Popovich, 2003). Building and maintaining long-term relationships with customers through CRM can gain both empathy and loyalty (Jain et al., 2021), and loyal customers can become great ambassadors in the market and help attract new business (Amin et al., 2019).

Due to shifting telecommunications behavior and intensified competition from deregulation, there is an increasing need to secure core business by strengthening CRM and to improve the profitability of each customer (Wei and Chiu, 2002; Huang et al., 2009; Khan et al., 2010; Qureshi et al., 2013; Vafeiadis et al., 2015; Do et al., 2017). To do so, companies need to understand the churn characteristics of their customers by grouping them according to risk level. Some companies have faced difficulty in executing churn management because of an inability to predict who will leave using high volume of customer data (Ahmad et al., 2019).

Machine learning (ML) algorithms have been extensively applied for decision makers to predict the possibility of customer churn (Ahmad et al., 2019), and the current study focuses on ensemble learning in this regard. Ensemble learning is a progressive ML paradigm in which multiple models are applied in processes known as bagging, boosting, or stacking (Breiman, 2001; Chen and Guestrin, 2016; Jain et al., 2020). In decision tree, bagging generates various trees from which a synthesized model will be generalized through aggregation. Boosting applies a specific algorithm to minimize error in developing its trees (Chen and Guestrin, 2016; Do et al., 2017). Stacking techniques combine several algorithms and apply them to the meta-learner in order to result in prediction (Abbasimehr et al., 2014).

Previous studies have demonstrated that ensemble learning is best for prediction models, including both classification and regression problems and it has been proven that the approach can increase model performance through generalization and error adjustment (Jain et al., 2021). Stacking techniques are particularly complex and time-consuming, and so this study employs bagging in the form of random forest (RF) and extreme gradient boosting (XGBoost) trees for their simplicity and robustness in predicting customer churn (Do et al., 2017). These techniques can also predict missing values, control overfitting, and generalize output (Breiman, 2001; Chen and Guestrin, 2016).

Through the early detection of customer churn, the alignment of sales and marketing strategies can be improved and churn management activities made more efficient (Mattison, 2006). A relevant method here is to cluster the churning customers according into groups, an approach that has been used in various industries to profile customers and apply appropriate strategies (Larasati et al., 2019; Ullah et al., 2019). Using a clustering approach, this study develops profiling to group customers into several churn clusters so that the company can determine specific retention strategies for each group. A nonhierarchical clustering technique, K-means, is used to find the criteria for churn behavior, as in previous studies (Ullah et al., 2019).

The current study proposes a churn prediction model by comparing ensemble learning approaches. The RF and XGBoost are each performed with and without hyperparameter tuning and compared with the classic logistic regression (LR) approach. Data preprocessing to balance churn and non-churn data is conducted using the synthetic minority oversampling technique (SMOTE), and classification is performed by considering the

external factor of competitor availability as part mimicking the original reality competition in the market.

2. Methods

Based on previous telecommunications studies, a number of factors can be seen to affect churn, such as demographic variance, the rate of market competition, available products, and customer perception of the service provider (Wei and Chiu, 2002; Mattison, 2006; Huang et al., 2009; Huang et al., 2012; Qureshi et al., 2013; Springer et al., 2014; Do et al., 2017; Ullah et al., 2019; Ahmad et al., 2019;). These factors are likely to influence the behavioural patterns related to customer need, and companies need to apply prediction techniques to gain insight from their data (Do et al., 2017; Ahmad et al., 2019).

Various approaches to predicting churn have been used, including classic and hybrid ML (Vafeiadis et al., 2015; Zhu et al., 2017). The most common prediction method for customer churn is decision trees since tree-based algorithms perform well with complex nonlinear connections between attributes (Ullah et al., 2019). The data used in this study is nonlinear, and we therefore intended to identify rules and hidden patterns through decision trees.

Applying SMOTE for data balancing was expected to ensure that the model has robust output (Chawla et al., 2002; Sisodia and Verma, 2019). Relatedly, most algorithms in ensemble learning have several parameters that need to be tuned to obtain optimal accuracy, and so these adjustments were made using the heuristic RandomSearchCV during development in order to generate a model with the best performance (Bergstra and Bengio, 2012).

Figure 1 shows the steps involved in the proposed churn prediction model. After input, reprocessing is performed during the first step, and this includes data transformation, such as noise removal, feature labelling, and normalization; feature selection, using correlation and variance ranking to define informational attributes; and then data balancing. In the second step, classification algorithms are applied to categorize customers into the churn and non-churn classes using an ensemble learning approach. Finally, clustering is implemented to group customers at high risk of churn into several groups with K-means for marketing and sales strategy alignment. The programming was conducted using Python 3.9.

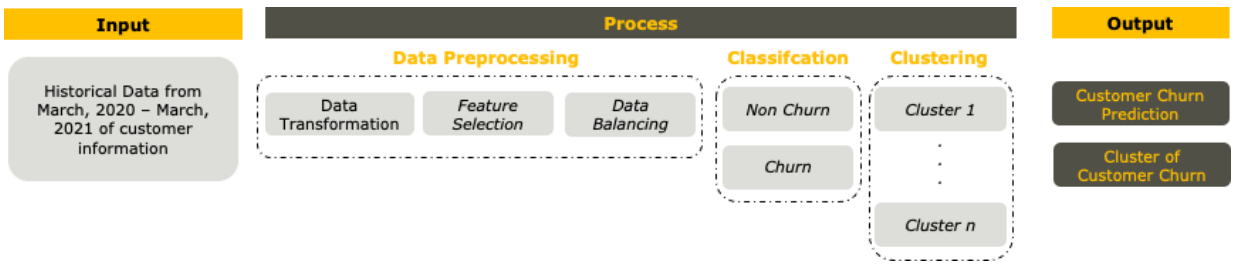

Figure 1 Steps of the proposed churn prediction model

2.1. Data Preprocessing

This study used customer data from a fixed broadband company in Indonesia for the year from March 2020 to March 2021. The data consists of 45 features for 1,638,076 subscribers, including demographic, internal, and external information. Through data transformation and feature selection, 16 variables that predict customer churn or non-churn behavior were identified (Table 1) with the six most important features as follows: the monthly billing; total number of assets; internet speed; onboard month; customer lifetime value; and limitation of internet usage.

Table 1 Features of predicting customer churn

No	Variable	No	Variable
1	Speed Rate	9	Total Number of Assets
2	Type of Domicile	10	Limitation of Internet Usage
3	Voice Usage (0/1)	11	Number of repairs required
4	Internet Usage (0/1)	12	Type of TV Channel Promotion
5	Usage of TV (0/1)	13	Type of Internet Speed
6	The monthly billing	14	Type of Sales Channel Promotion
7	Customer Lifetime Value	15	Type of Area Promotion
8	Onboard Month	16	Competitor in Area (0/1)

2.2. Model Development

The process of creating models for predicting customer churn is outlined below. Two ensemble learning approaches (i.e., RF and XGBoost) were employed and compared with LR across six scenarios, with and without heuristic parameter optimization. Each scenario was subsequently evaluated for accuracy and recall performance.

To tune the LR algorithm, the chosen parameters were "*C*" and "*Solver*;" for RF, they were "*n_estimator*," "*max_depth*," "*min_sample_leaf*," "*min_sample_split*," and "*max_features*;" and due to the establishment of the model decision in XGBoost, the chosen parameters were "*subsample,*" "*Colsample_bylevel,*" "*Colsample_bytree,*" and "*n_estimator.*" In addition, the gradient loss approach was used to reduce error rate as part of adjusting the model adjustment for overfitting. These parameters were chosen to set up the number of trees and proportion of each level. Table 2 outlines the hyperparameter tuning for the three algorithms.

Table 2 Hyperparameters for LR, RF, and XGBoost

Algorithm	Hyperparameter	Default value	Optimized value
LR	*C*	1	0.01
	Solver	*lbfgs*	*newton-cg*
RF	*n_estimator*	100	136
	Max_depth	0	28
	Min_sample_leaf	1	2
	Min_sample_split	2	38
	Max_features	*auto*	*sqrt*
XGBoost	*subsample*	1	0.8
	Colsample_bylevel	1	0.5
	Colsample_bytree	1	0.8
	N_estimator	100	340

2.3. Model Evaluation

Several evaluation methods were used to determine the proposed model's performance, including accuracy and recall assessments used in previous studies (Wei and Chiu, 2002; Fawcett, 2006; Huang et al., 2009; Khan et al., 2010; Do et al., 2017; Jain et al., 2021;). Problems regarding customer churn prioritize the effectiveness of predicting its true positive data (Ullah et al., 2019). Recall was therefore more suitable than other performance measures to evaluate this model's performance in addition to accuracy.

Accuracy is the percentage of the complete model predicting results correctly, calculated by True Positive (TP) + True Negative (TN) divided by all prediction targets (Equation 1). Recall, on the other hand, is the proportion of correct values predicted

correctly or looking at the model's performance in predicting the label's intended target, in this case customer churn (Equation 2).

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

$$Recall = \frac{TP}{TP + FN} \quad (2)$$

3. Results and Discussion

3.1. Comparison of Model Prediction Performance

The ensemble learning approach using RF and XGBoost exhibited better accuracy than LR. Where RF uses a divide-and-conquer approach, creating numerical decision trees and training each one by picking a random attribute from the whole predictive set, XGBoost uses gradient boosting which minimizes errors in the model.

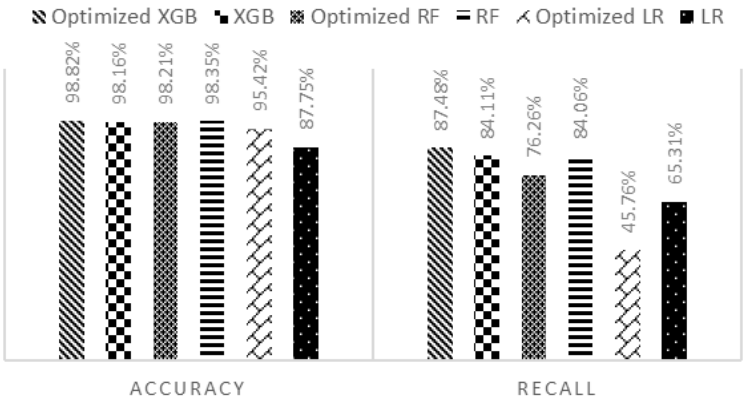

Figure 2 Model prediction performance

As shown in Figure 2, the algorithms demonstrate good accuracy with a minimum score of 87.75%. The XGBoost algorithm outperforms the other techniques and correctly classifies the data with 98.35% accuracy. This study supports earlier findings that an ensemble learning tree-based algorithm with boosting or XGBoost adjustment has the best performance for predicting customer churn (Do et al., 2017). Even though the large dataset had some missing values, XGBoost was able to process the data well.

This study also proves that heuristic optimization based on RandomSearchCV can positively and negatively affect ML models. The hyperparameter tuning aimed to improve the model's accuracy, but it resulted the separate problem of local optima. Figure 2 shows some optimized models seem to face local optima. RandomSearchCV focuses on increasing accuracy which can affect performance in recall. From the results of the optimized RF and optimized XGBoost, accuracy was only improved for the XGBoost approach, and so it seems that the RF developed local optima. Consequently, this study found that the best model for predicting customer churn is the optimized XGBoost.

3.2. Customer Profiling for Retention Strategies

K-means was performed to cluster the customers which generated four groups based on the six classification features identified earlier to build insight into retention strategy alignment (Figure 3). Each group was then subcategorized into three groups based on the risk of churn.

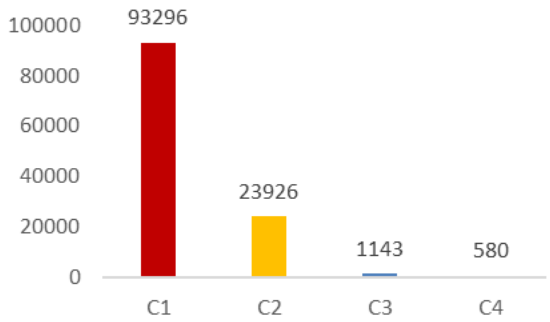

Figure 3 Composition of customer churn clusters

Clusters 1 and 2 dominate, and these are categorized as Newcomer and Baby (Soldani et al., 2011; Yang, 2013; Springer et al., 2014). Cluster 1 (C1) is characterized by a low ability to pay, infrequent internet use, and the tendency to use a service as a trial. Customers in this cluster tend to choose the lowest features and rarely use internet or tv subscription. Accordingly, C1 can be classified as high risk of churn due to low empathy and knowledge regarding the company and its services, thus the company must focus its retention activities in this cluster and work to attract customers to use the internet more frequently. Urgent strategies that could be used are bundling offerings together with consideration of the customer's long-term preferences (Yang, 2013). Another possible strategy is to provide a series of product differentiations (Soldani et al., 2011)

In contrast to C1, Cluster 2 (C2) can be categorized as medium risk. It is characterized by customers that have the capacity to use the internet although most only have one type of product offering, i.e. internet subscription, and their monthly billing is very low due to their online behavior. Here, the company should conduct in-depth analysis, i.e. channel distribution, type of services that underperform, for each customer to provide tailored offerings that would obtain greater empathy and loyalty leading to the automatic selection of that provider in their daily activities. Product bundling and loyalty programs can be provided to improve perceived quality and further increase empathy and experience for the customers (Springer et al., 2014). The overall strategy that needs to be applied for C2 is one that offers various long-term promotions accompanied by intense relationship management.

Meanwhile, clusters 3 and 4 (C3 and C4) can be categorized as low risk as they contain customers who tend to be entirely satisfied with the quality of service provided and those who tend to persist in using the same provider. Despite having distinct characteristics, both C3 and C4 have an expensive tier of services although customers in C3 tend to have lower monthly bills than those in C4. In these clusters, marketing adjustments could be made to ensure customers remain as subscribers, including loyalty programs to increase retention, moving assistance and internet installation in new addresses, and monthly or annual rewards so that every customer is happy with their product (Soldani et al., 2011; Yang, 2013).

4. Conclusions

Predicting churn using ensemble learning approaches and clustering customers according to risk has identified the most appropriate model for assisting decision makers in designing customer retention activities. As a result, the productivity of CRM in terms of retention could be improved, and both product and service quality could be enhanced which would increase customer commitment toward the company. The optimized XGBoost algorithm produced the best accuracy and recall in predicting customer churn at 98.82%

and 87.48%, respectively. This study also provides a framework for assessing churn based on low, medium, and high risk which could help the company focus and prioritize its retention activities by group to ultimately positively impact its profitability.

The main limitation of the study is that it considers competitor availability without categorizing them, for example as strong or weak. Attention to this particular feature is therefore of high importance to future studies. Two key adjustments to the study's methods have been identified: Firstly, competitor categorization, and secondly, more advanced methods should be applied to obtain an accurate & quicker outcome in predicting customer churn, for example stacking and deep learning (DL). Stacking algorithms would be expected to significantly improve model performance because they combine multiple classifiers, such as logistics regression and random forest combines for generating fewer errors (Abbasimehr et al., 2014). On the other hand, DL uses multiple layers to progressively extract higher-level features from the raw input by which the machine can learn customer behavioural patterns more thoroughly and effectively (Agrawal et al., 2018). However, this approach would cost much more in terms of computational time which must also be considered for any future studies.

References

Abbasimehr, H., Setak, M., Tarokh, M.J., 2014. A Comparative Assessment of the Performance of Ensemble Learning in Customer Churn Prediction. *International Arab Journal of Information Technology*, Volume 11(6), pp. 599–606

Agrawal, S., Das, A., Gaikwad, A., Dhage, S., 2018. Customer Churn Prediction Modelling based on Behavioural Patterns Analysis using Deep Learning. *In:* 2018 International Conference on Smart Computing and Electronic Enterprise (ICSCEE), pp. 1–6

Ahmad, A.K., Jafar, A., Aljoumaa, K., 2019. Customer Churn Prediction in Telecom using Machine Learning in Big Data Platform. *Journal of Big Data*, Volume 6(1), pp. 1–24

Amin, A., Al-Obeidat, F., Shah, B., Adnan, A., Loo, J., Anwar, S., 2019. Customer Churn Prediction in Telecommunication Industry using Data Certainty. *Journal of Business Research*, Volume 94, pp. 290–301

Bergstra, J., Bengio, Y., 2012. Random Search for Hyper-Parameter Optimization. *Journal of Machine Learning Research*, Volume 13(2), pp. 281–305

Breiman, L., 2001. Random Forests. *Kluwer Academic Publishers*, Volume 45, pp. 5–35

Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. SMOTE: Synthetic Minority Over-Sampling Technique. *Journal of Artificial Intelligence Research*, Volume 16, pp. 321–357

Chen, I.J., Popovich, K., 2003. Understanding Customer Relationship Management (CRM). *Business Process Management Journal,* Volume 9(5), pp. 672–688

Chen, T., Guestrin, C., 2016. XGBoost: A Scalable Tree Boosting System. *In:* Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794

Do, D., Huynh, P., Vo, P., Vu, T., 2017. Customer Churn Prediction in an Internet Service Provider. *In:* 2017 IEEE International Conference on Big Data (Big Data), pp. 3928–3933

Fawcett, T., 2006. An Introduction to ROC Analysis. *Pattern Recognition Letters*, Volume 27(8), pp. 861–874

Huang, B., Kechadi, M.T., Buckley, B., 2012. Customer Churn Prediction In Telecommunications. *Expert Systems with Applications*, Volume 39(1), pp. 1414–1425

Huang, B.Q., Kechadi, M.T., Buckley, B., 2009. Customer Churn Prediction for Broadband Internet Services. *In:* International Conference on Data Warehousing and Knowledge

Discovery, pp. 229–243
Jain, H., Khunteta, A., Srivastava, S., 2020. Churn Prediction in Telecommunication using Logistic Regression and Logit Boost. *Procedia Computer Science*, Volume 167, pp. 101–112
Jain, H., Yadav, G., Manoov, R., 2021. Churn Prediction and Retention in Banking, Telecom and IT Sectors using Machine Learning Techniques. *In:* Advances in Machine Learning and Computational Intelligence, pp. 137–156
Khan, A.A., Jamwal, S., Sepehri, M.M., 2010. Applying Data Mining to Customer Churn Prediction in an Internet Service Provider. *International Journal of Computer Applications*, Volume 9(7), pp. 8–14
Larasati, A., Hajji, A.M., Handayani, A.N., Azzahra, N., Farhan, M., Rahmawati, P., 2019. Profiling Academic Library Patrons using K-means and X-means Clustering. *International Journal of Technology*, Volume 10(8), pp. 1567–1575
Mattison, R., 2006. *The Telco Churn Management Handbook*. Lulu.com
Qureshi, S.A., Rehman, A.S., Qamar, A.M., Kamal, A., Rehman, A., 2013. Telecommunication Subscribers' Churn Prediction Model using Machine Learning. *In:* Eighth International Conference on Digital Information Management (ICDIM 2013), pp. 131–136
Sarachuk, K., Mißler-Behr, M., 2020. Is Ultra-Broadband Enough? The Relationship between High-Speed Internet and Entrepreneurship in Brandenburg. *International Journal of Technology*, Volume 11(6), pp. 1103–1114
Sisodia, D.S., Verma, U., 2019. Distinct Multiple Learner-Based Ensemble Smotebagging (ML-ESB) Method for Classification of Binary Class Imbalance Problems. *International Journal of Technology*, Volume 10(4), pp. 721–730
Soldani, D., Hou, X.J., Luck, B., 2011. Strategies for Mobile Broadband Growth: Traffic Segmentation for Better Customer Experience. *In:* 2011 IEEE 73rd Vehicular Technology Conference (VTC Spring), pp. 1–5
Springer, T., Kim, C., Debruyne, F., Azzarello, D., Melton, J., 2014. Breaking the Back of Customer Churn. *Bain & Company*, pp. 1–8
Statista, 2019. *Telkom Indonesia: Fixed Broadband Market Share 2019 | Statista*. Available Online at https://www.statista.com/statistics/1058240/telkom-indonesia-fixed-broadband-market-share
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*, Volume 7, pp. 60134–60149
Vafeiadis, T., Diamantaras, K.I., Sarigiannidis, G., Chatzisavvas, K.C., 2015. A Comparison of Machine Learning Techniques for Customer Churn Prediction. *Simulation Modelling Practice and Theory*, Volume 55, pp. 1–9
Wei, C.-P., Chiu, I.-T., 2002. Turning Telecommunications Call Details to Churn Prediction: A Data Mining Approach. *Expert Systems with Applications*, Volume 23(2), pp. 103–112
World Bank, 2020. *World Development Indicators | DataBank*. Available Online at https://databank.worldbank.org/reports.aspx?source=2&series=IT.NET.BBND&country=IDN#
Yang, M., 2013. Churn Management and Policy: Measuring the Effectiveness of Fixed-Mobile Bundling on Mobile Subscriber Retention. *Journal of Media Economics*, Volume 26(4), pp. 170–185
Zhu, B., Baesens, B., vanden Broucke, S.K.L.M., 2017. An Empirical Comparison of Techniques for the Class Imbalance Problem in Churn Prediction. *Information Sciences*, Volume 408, pp. 84–99