The Scientific Temper (2023) Vol. 14 (3): 618-624
Doi: 10.58414/SCIENTIFICTEMPER.2023.14.3.08
E-ISSN: 2231-6396, ISSN: 0976-8653
https://scientifictemper.com/

ORIGINAL RESEARCH

Customer churn prediction using machine-learning techniques in the case of commercial bank of Ethiopia

Temesgen Asfaw

Abstract

The number of service providers is increasing rapidly in every business. These days, there is plenty of options for customers in the banking sector when choosing where to put their money. As a result, customer churn and engagement have become one of the top issues for most of the banks. In this paper, a method to predict customer churn in a Bank using machine learning techniques, which is a branch of artificial intelligence, is proposed. The research promotes the exploration of the likelihood of churn by analyzing customer behavior. random forest (RF), logistic regression (LR), gradient boosting classifier (GBC), extreme gradient boosting classifier (EGBC), and light gradient boosting machine classifier (LGBCMC) are used in this study. Also, some feature selection methods have also been done to find the more relevant features and verify system performance. The experimentation was conducted on the churn modeling dataset from Kaggle. The results are compared to find an appropriate model with higher precision and predictability. As a result, using the Random Forest model after oversampling is better than other models in terms of accuracy. The experimental result shows that the Light Gradient Boosting Machine classifier outperformed with an accuracy of 98%, a precision of 97%, and a recall of 100%, with an AUC of 99% than other proposed supervised machine learning algorithms with balanced datasets across all evaluation metrics.

Keywords- Customer churn, Commercial bank of ethiopia, Gradient boosting classifier, Extreme gradient boosting classifier, and Light gradient boosting machine classifier.

Introduction

The market is very dynamic and highly competitive nowadays. It is because of the availability of a large number of service providers. The challenges of service providers are changing customer behavior and rising expectations. The rising aspirations of current generation consumers and their diverse demands for connectivity and innovative, personalized approaches are very distinct from previous generations of consumers. They are well-educated and better informed about emerging approaches. Such advanced knowledge has changed their purchasing behavior, resulting in a trend of 'analysis-paralysis' over-analyzing the selling and purchase scenario, which ultimately helps them to improve their purchase decisions. Therefore, this is a big challenge for the new generation of service providers to think innovatively to fulfill and add value to the customer's needs.

Commercial Bank of Ethiopia (CBE) was initiated as the community bank of Ethiopia in 1942. CBE had currently 22 million customers, and the total of Internet banking users is around 2.5 million. Furthermore, as the banking business grows and advances, the institutions want to track existing customers' status to check whether they churned or not.

According to Sashi (2012), there are two reasons customers churn from one institution: accidental and intentional. Intentional churn occurs when a customer is unsatisfied with the current institution and finds alternative institutions that provide comparable services. This may include interest rates, customer relationships, etc. Accidental churn occurs when a customer accidentally stops using one institution. This may include death, financial problems, etc. According to Nagadevara *et al.,* (2015), Kaya *et al.,* (2018), and Karvana *et al.,* (2019), customer churn prediction focuses on ascertaining customers who will churn so that basic mediation can be taken to keep current customers. Banking and financial organizations have anticipated monitoring their current customer relations to determine who will quit Deng *et al.,* (2021).

School of Computing and Informatics Mizan Tepi University, Ethiopia

***Corresponding Author:** Temesgen Asfaw, School of Computing and Informatics Mizan Tepi University, Ethiopia, E-Mail: temesgenabera@mtu.edu.et

How to cite this article: Asfaw T. (2023). Customer Churn Prediction Using Machine-Learning Techniques in the Case of Commercial Bank of Ethiopia. The Scientific Temper, **14**(3):618-624.

Doi: 10.58414/SCIENTIFICTEMPER.2023.14.3.08

Source of support: Nil

Conflict of interest: None.

Received: 10/07/2023 **Accepted:** 20/08/2023 **Published** : 25/09/2023

The study aimed to develop and compare supervised machine learning algorithms such as RF, LR, GBC, EGBC, and LGBMC for customer churn prediction using a commercial bank of Ethiopia dataset. Machine learning algorithms play an important role in the predictive analysis of business organizations such as banks and financial institutions Kumar *et al.,* (2021). The supervised machine learning algorithm is one machine learning algorithm used in case classification problems**.** Classification is the process of separating different things into discrete classes. Customer churn prediction is a classification problem that identifies whether the customer will quit or not.

Related Work

Many researchers in the past few decades have studied customer churn prediction by supervised machine learning. Some of the noteworthy related works are as follows:

Dang (2008) has proposed a novel approach using a machine-learning algorithm to predict customer churn. The author compared various machine learning algorithms, such as decision trees, artificial neural networks, naive Bayesian, and logistic regression classifiers, by performance evaluation metrics to suggest the best fitting model. According to the experimental results, support vector machines outperformed other machine learning algorithms and provided an actual application in predicting customer churn. Pan *et al.,* (2009) have proposed a new technique using a data mining framework cross-industry standard process (CRISP) and decision trees for customer churn and stockholder prediction. According to the decision tree and the CRISP model, the fund transferal pattern is the most important factor.

Furthermore, the models suggest low return rates have destructive consequences for silent stakeholders. The misclassification rate and social finance philosophy are the most important factors in both the decision tree and the CRISP models. Zhang *et al.,* (2009) have developed a new approach using a Fuzzy C-means clustering algorithm to predict customer churn and customer segmentation applied to the customer churn model to improve the prediction accuracy. Finally, the results of the proposed model by applying customer segmentation to predict customer churn have shown great promise compared to traditional techniques.

Niu *et al.,* (2009) have proposed a new approach using least squares support vector machine (LS-SVM) and rough set theory (RST) models for credit card churn predictions, that could precisely and accurately predict customer churn. The RST and LS-SVM models were compared to traditional machine learning techniques. The experimental result shows that the LS-SVM outperformed in all performance evaluation metrics compared to traditional machine learning algorithms and RST. Qin *et al.,* (2015) have developed a new approach using a data mining-based decision tree algorithm to assess e-commerce-shared customers and to find features of customer churn that aid organizations in maintaining good relations with current customers and avoiding churning customers. The experiment shows that data mining techniques have a great role in dealing with customer churn prediction. Bayazıt *et al.,* (2015) have developed a novel approach using a data mining model for customer churn prediction. Studies suggest that data mining greatly affects business analytics and churn prediction. Performance evaluation factors such as kappa statistics, sensitivity, accuracy, specificity, and ROC results were compared for customer churn prediction. The experimental result shows that data mining models greatly influence customer churn prediction.

Similarly, Gou *et al.,* (2015) have proposed a novel approach using a data mining model for customer churn prediction. The author analyzed customer objections to an information service using a data-mining model to predict customer churn. The experimental result shows that data mining models greatly influence customer churn prediction.

Mishra (2017) has proposed a new approach to predict customer churn in the telecommunication sector using ensemble-learning techniques. Ensemble learning techniques were compared with supervised machine learning algorithms for telecommunication churn predictions. The experimental result shows that random forest algorithms have a high accuracy of 91.6% compared to other supervised machine learning and ensemble learning techniques. Azzopardi *et al.,* (2018) have developed a novel approach using a digital set-top box and have overwhelming histories from cable networks' financial systems for customer churn prediction. According to the findings of the studies, customer payment practices, customer consumption, and customer viewing concertation all significantly impact customer churn prediction. According to the sightings, intentional user churn factors are well prepared and capable of as long as real advertising approaches for cable network businesses. Karvina *et al.,* (2019) proposed a new approach using machine-learning algorithms for customer churn prediction and compared five machine-learning models. The experimental results showed that random forest algorithms outperformed other machine learning models. Ullah *et al.,* (2019) have developed a new approach for predicting customer churn in banking centers using cluster-based local outlier factors, local outlier factors, and K-means. The model's effectiveness in predicting customer churn is measured by performance evaluation metrics such as accuracy, recall, precision, and F1 score.

Hu *et al.,* (2020) have proposed a new approach using an integrated model for customer churn prediction. The proposed integrated model shows high accuracy when compared to the single churn prediction model. The experimental result shows that the integrated model for customer churn prediction is better in all performance evaluation metrics than the single churn prediction model.

Deng *et al.,* (2021) have developed a new approach using ensemble-learning techniques such as Light GBM, chat boost, and RF to improve the prediction of customer churn in banking centers. Furthermore, the authors compared the models' performance measurement metrics such as recall, precision, accuracy, and f1-score. The experimental results show that Random Forest outperformed all performance measurement metrics light GBM and chat boost. Similarly, Kumar *et al.,* (2021) have compared different machine learning algorithms for customer churn prediction. The experimental result shows that random forest outperformed other machine learning algorithms.

Material and Methods

The proposed model for customer churn prediction in the case of the commercial bank of Ethiopia is illustrated and explained in Figure 1 below. After explaining the proposed model, we discussed the dataset and data preparation. Furthermore, different supervised machine-learning algorithms used for the problems mentioned above are explained theoretically and mathematically below. Finally, the phases of the proposed system are discussed below. Figure 1 above shows that the proposed diagram architecture for customer churn prediction and the whole process are discussed below.

Data collection phase

The first step in machine learning research is collecting the necessary datasets from concerned organizations or sites Zhang *et al.,* (2003). The dataset used for the study was collected from the CBE. It holds statistics of current customers of the commercial bank of Ethiopia from 2019 to 21. The total number of records in the dataset used for the study was 49707 records and 11 features.

Data preparation phase

According to Zhang *et al.,* (2003), data preparation is making raw data suitable for a machine learning model to analyze it. It is part of cleaning noisy data, filling or removing missing data, data normalization, feature selection, and target class balancing.

Missing data

Machine learning algorithms do not handle data with missing values, so it is compulsory to handle data with missing values before fitting them to the model Kotsiantis *et al.,* (2006). According to Kang *et al.,* (2013), variables with a missing value above 60% were removed from the dataset. For continuous variables, mean values are credited for missing values ranging from 2 to 30% of the total.

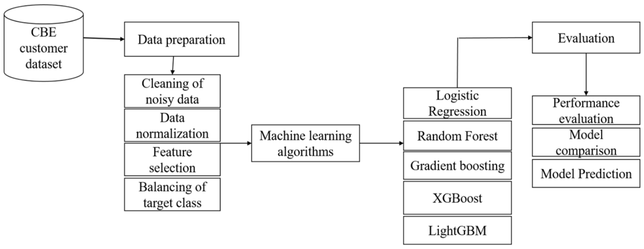

Figure 1: Proposed architecture for customer churn prediction.

Data normalization

Data normalization is the process of converting different formats of data scales to one standard measure. Data collected from the CBE was on different scales, so data normalization is required. For normalizing data, we used from sklearn library Min Max Scaler class.

Feature selection

According to Wieringa *et al.,* (2005), not all collected datasets are useful for predicting machine learning models. The collected dataset from commercial banks of Ethiopia was subjected to feature selection to select important features for customer churn prediction. According to Miao *et al.,* (2016), we need to select important features before giving the features to machine learning algorithms. Feature removed from the CBE dataset with a *p-value* greater than 0.05 or the standard cut-off value and a small chi-square.

Class imbalance

The Synthetic Minority Oversampling Technique (SMOTE) is a technique used to balance class imbalance problems in the training dataset. The dataset collected from the commercial bank of Ethiopia was target class imbalanced. According to Kotsiantis *et al.,* (2006), Burez *et al.,* (2009) and Maheshwari *et al.,* (2017), the collected dataset must be balanced before training to the proposed model. The synthetic minority oversampling technique (SMOTE), works by nominating related records from the smaller class and altering them one column at a time by a random amount to balance the data. The final dataset was divided into train and test datasets for training, and testing proposed supervised machine learning algorithms. According to previous studies, 80% of the data is used for training and 20% for testing Gou *et al.,* (2015), Fabris *et al.,* (2017), and Hu *et al.,* (2020).

Model fitting phase

For customer churn prediction in the case of commercial banks in Ethiopia, we developed and compared the following supervised machine learning models:

Logistic regression

Logistic regression is a classification technique that is used in the case of classification problems. Logistic regression uses a sigmoid function to convert values into discrete classes. Input values b are joined linearly using coefficient values to predict a target value (y). The main difference between logistic regression and linear regression is that the target value is a (0, 1) or binary value rather than a continuous value.

$$y = e^{\frac{a_0 + a_1 + b}{\left(1 + e^{a_0 + a_1 + b}\right)}} \quad 1$$

Where y represents a prediction, a_0 represents the intercept value, and a_1 represents a coefficient for b. The sigmoid function is a function that takes any value and maps it into a value [0, 1].

$$\text{Sigmoid} = \frac{1}{1+e^{y}} \qquad 2$$

Random Forest

RF is a predictive model based on decision trees. It predicts the target class by averaging the results of the decision tree. As the number of decision trees gets bigger, the accuracy of the result is also better. Random forest algorithms are implemented in the following manner:

Step I, suppose the training data contains O observations and F features. First, randomly selected samples are taken with replacements from the training set.

Step II, a random subset of F features, is nominated based on criteria that provide the best split, which is then used to divide nodes iteratively.

Step III, repeat Step I and Step II for N numbers of trees predicted based on aggregation.

Gradient boosting classifiers

Gradient boosting classifiers is a collection of machine learning algorithms based on combining different weak classifiers to generate strong classifiers, such as decision trees. In gradient boosting, each predictor keeps trying to outperform the one before it by reducing the number of errors. The fascinating concept behind gradient boosting is that instead of trying to fit a predictor to the data at each iteration, it conforms a new predictor to the error variance left by the previous predictor.

XG Boost

The acronym XG Boost stands for "EGB." It is quick, adaptable, and versatile due to its decentralized nature. The XG Boost classifiers are ensemble learning decision tree-based classifiers that use gradient boosting. Gradient boosting structure implementations are based on machine learning models. Gradient boosting classifiers are used nowadays for a wide range of data science issues.

Light GBM

Light Gradient Boosted Machine (Light GBM) is a gradient boosting structure based on a decision tree that divides the tree leaf-wise, whereas other boosting classifiers divide the tree level-wise. Most commonly used for hierarchies, classifications, and other machine-learning tasks. Light GBM is used for most classification problems because it gives better accuracy than other boosting algorithms because a leaf-wise approach minimizes much more loss than level-wise classifiers when getting bigger on the same leaf. It is quicker than all-existing boosting classifiers.

Experimental Evaluation and Results

The experiments were conducted using Python and Pycharm IDE 3.7. Two experiments were conducted on the collected dataset of commercial banks of Ethiopia. The first experiment was done using a class-imbalanced dataset, which was directly collected from the commercial bank of Ethiopia. The second experiment was conducted using a class-balanced dataset by the technique of SMOTE. Both experiments were compared based on the performance measurement metrics such as recall, precision, accuracy, f-1 score, and AUC.

Table 1 shows that the proposed supervised machine-learning algorithms achieved high accuracy with a class-imbalanced dataset. Because many datasets belong to one target class, the experiments were biased; this is why we need the second experiment.

Table 2 shows that the accuracy evaluation metrics decreased as the target class was balanced and other performance evaluation metrics increased. More essential performance measurement metrics in the case of bank customer churn prediction is recalled evaluation metrics because recall evaluation metrics show a truly categorized positive target class or predict the future of current customers, whether they churn or not

In Figure 2 we can see a comparison of proposed supervised machine learning techniques with a class-imbalanced dataset. As a result, the accuracy of all proposed supervised machine learning was above 92% because 93% of the dataset belonged to one class. Other performance evaluation metrics were very low, like recall, which is the best metric for churn prediction.

In Figure 3 we can see that the accuracy of all proposed supervised machine-learning algorithms with class balanced dataset was lower than the Figure 2 result because the dataset is now balanced. However, the recall for all supervised machine-learning algorithms using class balanced dataset was very high compared to a class imbalanced dataset.

Table 1: Results of class imbalanced dataset

Models	Accuracy	Precision	Recall	AUC
Logistic regression	0.922	0.000	0.000	0.500
Random forest	0.937	0.904	0.204	0.977
Gradient boosting	0.922	0.000	0.000	0.938
XG Boost	0.977	0.848	0.856	0.994
Light GBM	0.955	0.893	0.481	0.993

Table 2: Results of a class-balanced dataset

Models	Accuracy	Precision	Recall	AUC
Logistic regression	0.619	0.604	0.691	0.661
Random forest	0.916	0.856	1.000	0.979
Gradient boosting	0.832	0.749	1.000	0.930
XG Boost	0.989	0.979	1.000	0.994
Light GBM	0.989	0.978	1.000	0.994

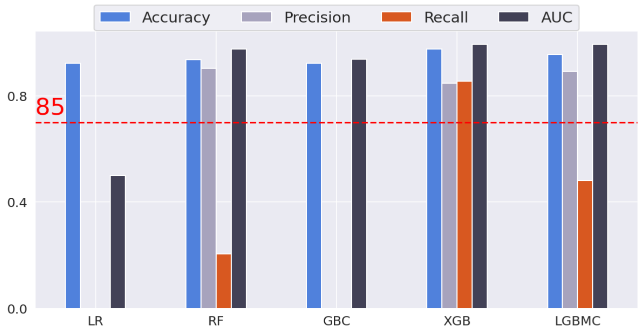

Figure 2: Comparison of proposed algorithms with class imbalanced dataset

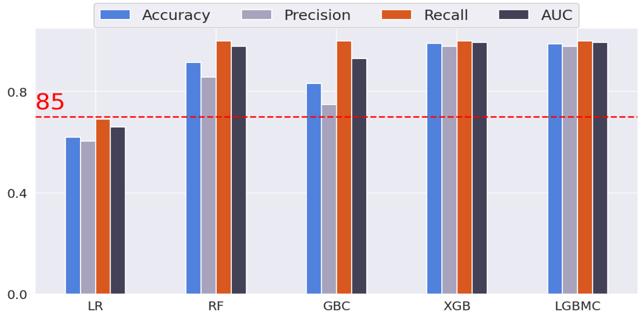

Figure 3: Comparison of proposed algorithms with a class-balanced dataset

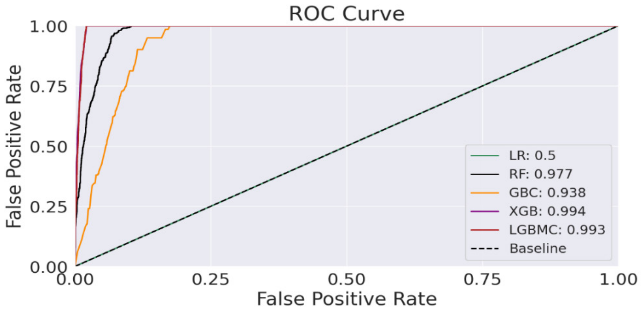

Figure 4: ROC curve of a class imbalanced dataset.

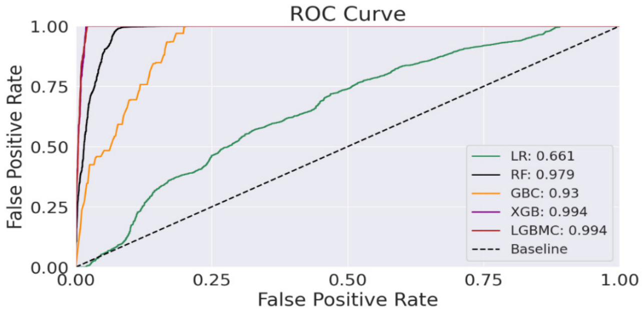

Figure 5: ROC curve of a class balanced dataset.

Figure 4 shows that the ROC curve was plotted to compare the proposed supervised machine with a class-imbalanced dataset. The higher the area under the curve, the best the prediction result. The XGB classifier obtains a higher area under the curve.

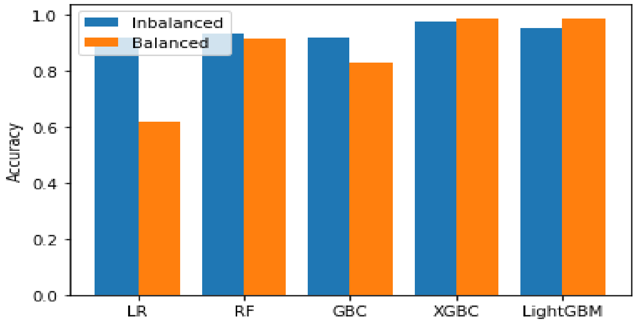

Figure 6: Comparison of class balanced and class imbalanced datasets using accuracy.

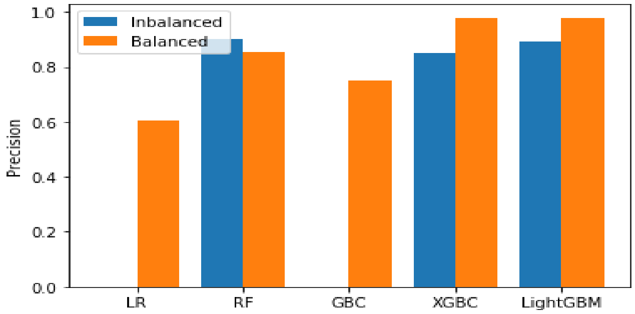

Figure 7: Comparison of class balanced dataset and class imbalanced dataset using precision

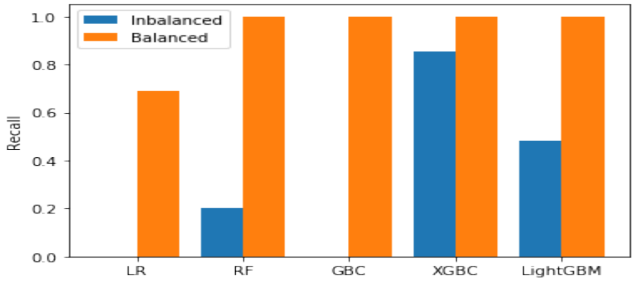

Figure 8: Comparison of the class-balanced and class-imbalanced datasets using recall

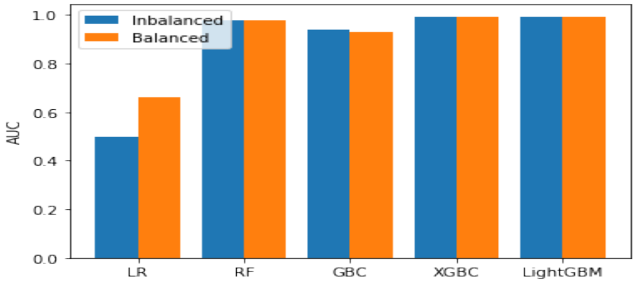

Figure 9: Comparison of class balanced and class imbalanced datasets using AUC.

In Figure 5 we can see that the ROC curve was plotted to compare a supervised machine with a class-balanced dataset. The higher the area under the curve, the best the prediction result. Both XGB and LGBM classifiers obtain a higher area under the curve.

In Figure 6 we can see a comparison of the class-imbalanced and class-balanced datasets using accuracy evaluation metrics for customer churn prediction. The result shows that the accuracy values obtained with class imbalanced dataset were high because a majority of the collected dataset belongs to one class.

In Figure 7 we can see the comparison of class imbalance and class balanced datasets using precision evaluation metrics for customer churn prediction. The result shows that precision evaluation metrics in the case of a class-imbalanced dataset are very low because the majority of the class belongs to one target class.

In Figure 8 we can see a comparison of a class-balanced and a class-imbalanced dataset using recall evaluation metrics for customer churn prediction. The result shows that the class-balanced dataset using SMOTE techniques shows high recall compared to the class-imbalanced dataset.

In Figure 9 we can see a comparison of class-balanced and class-imbalanced datasets using AUC evaluation metrics for customer churn prediction. The AUC assessment metrics show that a class-balanced dataset was influential in determining customer loss.

Conclusion

Customer churn is a key issue that is rising in all business organizations like banks and other institutions nowadays. Hence, customer churn prediction is the only way for the organization to take the necessary steps to protect customers who are willing to leave the institution. The goal of this research is to develop a supervised machine-learning algorithm for customer churn prediction in the case of the commercial bank of Ethiopia. The dataset collected was class imbalanced; only about 7.2% of records are related to customer churn. The SMOTE technique solved the class imbalance problems. Finally, two experiments were conducted on the collected dataset of the commercial banks of Ethiopia, one with a class imbalanced dataset and the other with class balanced dataset using SMOTE techniques. The first experiments show that the accuracy evaluation metrics for all proposed supervised machine learning are very high compared to other performance evaluation metrics because most target classes belong to just one target class. The second experiment was based on a class-balanced dataset, and the result of all performance evaluation metrics was nearly equal. The recall evaluation metrics were more critical for bank churn prediction, as recall shows the properly categorized element of a positive class or customer who would churn. The second experimental result showed that Light GBM outperformed in all performance evaluation metrics compared to other supervised machine learning algorithms.

Declaration of Competing Interest

The authors declare that there is no conflict of interest regarding the publication of this manuscript.

Data Availability

The data used to support the finding of this study are included in this research article. For simulation, we have used data from other research papers, which are properly cited

References

E. D. X. S. B. S. B. B. P. A. Kaya, " Behavioral Attributes and Financial Churn Prediction.," *EPJ Data Science,* pp. 7(1), 1-18 , 2018.

K. G. M. Y. S. S. A. &. M. P. Karvana, "Customer Churn Analysis and Prediction Using Data Mining Models in Banking Industry.," *International Workshop on Big Data and Information Security (IWBIS),* 2019 .

C. Sashi, "Customer engagement, buyer-seller relationships, and social media." *Management Decision,* pp. 50(2), 253–27, 2012.

Y. L. D. Y. L. T. J. &. Z. J. Deng, "Analysis and prediction of bank user churn based on ensemble learning algorithm.," in *IEEE International Conference on Power Electronics, Computer Applications (ICPECA).* , 2021.

V. Nagadevara, "CUSTOMER CHURN PREDICTION IN BANKING INDUSTRY.," *California Business Review,* pp. 3(1), 41–46., 2015.

F. M. J. P. D. &. F. A. A. Fabris, "A review of supervised machine learning applied to aging research." *Biogerontology,* pp. 18(2), 171–188., 2017.

Y. L. D. Y. L. T. J. &. Z. J. Deng, "Analysis and prediction of bank user churn based on ensemble learning algorithm.," in *IEEE International Conference on Power Electronics, Computer Applications (ICPECA).*, 2021.

S. L. Kumar, "Bank Customer Churn Prediction Using Machine Learning.," *International Journal for Research in Applied Science and Engineering Technology (IJRASET).,* 2021.

A. &. R. U. S. Mishra, " A comparative study of customer churn prediction in telecom industry using ensemble based classifiers.," in *International Conference on Inventive Computing and Informatics (ICICI).*, 2017.

X. Y. Y. C. L. &. Z. S. Hu, "Research on a Customer Churn Combination Prediction Model Based on Decision Tree and Neural Network.," in *IEEE5thInternational Conference on Cloud Computing and Big Data Analytics (ICCCBDA).*, 2020.

S. Y. A. S. a. P. M. K. G. M. Karvina, "Customer Churn Analysis and Prediction Using Data Mining Models in Banking Industry," *International Workshop on Big Data and Information Security (IWBIS),* 2019.

H. H. I. A. a. A. L. Ullah, "Churn Prediction in Banking System using K-Means, LOF, and CBLOF," in *International Conference on Electrical, Communication, and Computer Engineering (ICECCE)*, 2019.

M. S. A. G. Azzopardi, "Customer Churn Prediction for a Motor Insurance Company," in *Thirteenth International Conference on Digital Information Management (ICDIM)*, 2018.

K. Ş. a. N. G. Bayazıt, "Customer churn modelling in banking," in *23nd Signal Processing and Communications Applications Conference (SIU)* , 2015.

X. Z. a. J. Gou, "Warning model of customer churn based on emotions," in *International Conference on Logistics, Informatics and Service Sciences (LISS)*, 2015.

F. G. A. H. Qin, "The Analysis of Customer Churns in e-Commerce Based on Decision Tree," in *International Conference on Computer Science and Applications (CSA)*, 2015.

N. W. a. D.-x. Niu, "Credit card customer churn prediction based on the RST and LS-SVM," in *6th International Conference on Service Systems and Service Management*, 2009.

C. Y. a. X. Y. Pan Yan, "Predict the churn and silent customers: A case study of individual investors," in *IEEE International Conference on Intelligent Computing and Intelligent Systems*, 2009.

G. F. a. H. H. X. Zhang, "Customer-Churn Research Based on Customer Segmentation," in *International Conference on Electronic Commerce and Business Intelligence*, 2009.

J. Z. a. X. Dang, "Bank Customer Churn Prediction Based on Support Vector Machine: Taking a Commercial Bank's VIP Customer Churn as the Example," in *4th International Conference on Wireless Communications, Networking and Mobile Computing*, 2008.

S. Z. C. Y. Q. Zhang, "Data Preparation for Data Mining," *Applied Artificial Intelligence,* p. 375–381, 2003.

S. K. D. &. P. P. Kotsiantis, "Handling imbalanced datasets: A review.," *GESTS International Transactions on Computer Science and Engineering,* pp. 1-12, 2006.

H. Kang, "The prevention and handling of the missing data." *Korean Journal of Anesthesiology,* pp. 402-409. , 2013.

J. &. N. L. Miao, "A Survey on Feature Selection.," *Procedia Computer Science,* pp. 919-926, 2016.

R. M. N. M. N. &. R. C. Wieringa, "Requirements engineering paper classification and evaluation criteria: a proposal and a discussion.," *Requirements Engineering,* pp. 102-107, 2005.

S. K. D. &. P. P. Kotsiantis, "Handling imbalanced datasets: A review.," *GESTS International Transactions on Computer Science and Engineering,* pp.1-12, 2006.

S. J. R. &. J. R. Maheshwari, "a Review on Class Imbalance Problem: Analysis and Potential Solutions.," *International Journal of Computer Science,* pp. 43-51, 2017.

J. &. V. d. P. D. Burez, "Handling class imbalance in customer churn prediction.," *Expert Systems with Applications,* pp. 4626-4636, 2009.