Published by The College of Science & Information Technology (COSIT), TASUED, Vol. 18, No. 2, pp. 162-173.

COSIT, TASUED
Journal of Science and Information Technology (JOSIT)

An Ensemble Learning-Based Predictive Model for the Classification of Churn

Adedeji, O.B.

Department of Computer and Information science, Tai Solarin University of Education, Ijagun, Ogun state, Nigeria.
Corresponding Author: adedejiob@tasued.edu.ng

Abstract

Advances in ensemble learning and data analytics have transformed how businesses handle the problem of customer churn. Companies used customer surveys to get feedback and improve their services before predictive models were developed. Recognizing customer churn as a critical issue implies understanding that when customers leave a business at a higher rate than desired, it indicates deeper problems or challenges within the organization. The features that are associated with the classification of churn among customers were identified from literature following which relevant data was collected from an online repository provided online by Kaggle. The predictive model for the classification of churn among customers was simulated using the holdout method based on three simulation runs for the two ensemble learning models LightGBM and Bagging with Decision Tree. The results of the study showed that among the ensemble methods adopted, the LightGBM classifier proved to have the best overall performance among all the ensemble models considered having 100% accuracy through the three simulation. However, it was observed that as the proportion of the training datasets increased, the performance of the machine learning classifiers improved. The features that are most important to the classification of churn among customers include: type of customer contract, tenure of using service, satisfaction level of customer support, online security service, device protection service, payment methods, and streaming services. The study concluded that each feature had a relative importance to one another regarding their usefulness in the classification of churn.

Keywords: Classification algorithm, Data mining, Decision tree, Naïve Bayes, K-nearest neighbour

INTRODUCTION

Advances in ensemble learning and data analytics have changed how businesses approach their issue of the customer churn (Rahaman, Rani, Islam and Bhuiyan; Adrin, 2023 & Fadaralika, 2020). Prior to predictive models, companies relied on customer surveys to gather feedback and enhance their services while establishing customer success teams to facilitate smoother on-boarding and provide support (Nhu, Ly, & Son, 2022).

Ensemble learning approaches offer several advantages over predictive models based on a single ensemble learning algorithm (Karalar, Kapucu, & Gürüler, 2021; Gore et al. (2023)). By incorporating multiple models together, ensemble learning strategies can improve predicted accuracy and robustness by utilizing each model's advantages while minimizing its limitations.

Moreover, ensemble learning methods can handle diverse types of data and modelling techniques, enabling a more comprehensive analysis of factors influencing prediction. Ensemble Modelling is widely used in various ensemble learning tasks, including classification, regression, and clustering. It has been shown to improve predictive performance, reduce variance, and increase model robustness compared to single-model approaches. However, ensemble modelling requires careful tuning of hyperparameters, selection of diverse base models, and consideration of

Cite as:

Adedeji, O. B. (2024). An Ensemble Learning-Based Predictive Model for the Classification of Churn. *Journal of Science and Information Technology (JOSIT),* Vol. 18 No. 2, pp. 162-173.

computational resources to achieve optimal results (Joolfoo, et al, 2020).

As mentioned earlier, analysing customer behaviour serves as the basis for predicting customers who might churn, which is important for many reasons. One reason is that for companies who rely on subscription-based income, it can make a big difference on whether they can keep a steady income level or if they need to make changes to their services to keep customers (Lalwani, Mishra, Chadha, & Sethi, 2022). Another reason is that, compared to retaining customers, attracting new ones is costlier and firms can save money by retaining their existing customer base (de Lima Lemos, Silva, & Tabak, 2022).

In the world of commercial enterprises, achieving financial success is closely tied to customer management. This means not only acquiring new customers but also minimizing customer churn, as these factors together contribute to a business's overall performance (Imani, 2023). Many studies have considered the subject matter of the development of predictive models required for the classification of customer churn from various dimensions but many have not considered the importance of features on the performance of predictive models. Also, it has been observed that many studies were limited to the use of a single simulation run for development whereas multiple simulation runs of varying proportions of training and testing datasets can provide better insights into the performance of predictive models thereby providing more optimal solutions. There is a need for the identification of the features that are likely to improve the performance of the classification of customer churn thereby mitigating the onset of churn among customers using ensemble learning algorithms, hence this study.

RELATED WORKS

Malik, Runwal, Shah, Raut, and Hire (2023), worked on the application of ensemble learning algorithms to the prediction of churn. The study collected data containing information about 20 features from 7043 anonymous clients records from a telecommunication company's database. The dataset was extracted using SQL and stored

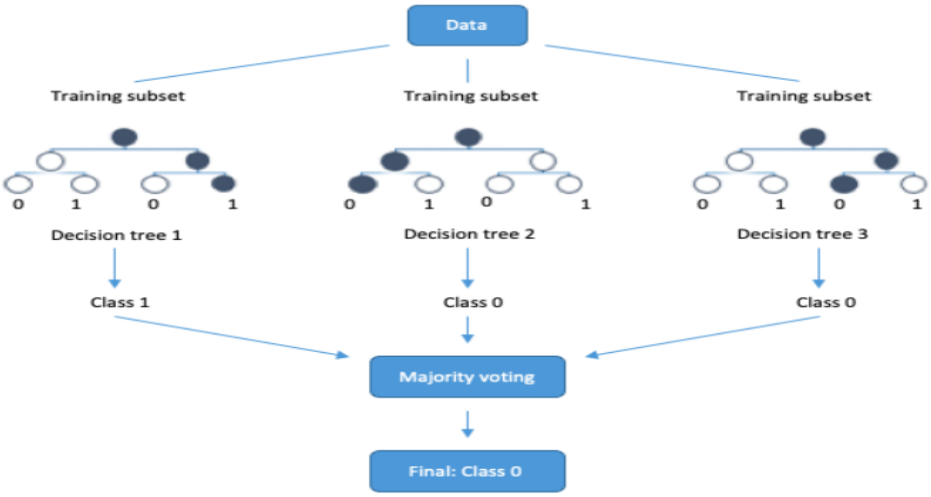

Figure 1. A visualization of the random forest approach (Source: Malik, Runwal, Shah, Raut, and Hire (2023).

as a spreadsheet file following data cleaning which invoved the removal of missing values, duplicate records and outliers from the data. The study simulated the predictive model using XGB Classifier, LightGBM Classifier, Random Forest Classifier, and Decision Tree Classifier including a stack of all classifiers based on 5-fold cross validation technique. The result of the study revealed that the best performance was achieved using LightGBM with an accuracy of 90.3% however the stack of classifier provided an accuracy of 90.9%. The study concluded that ensemble model show better performance at churn prediction compared to ensemble learning algorithms.

Sindikubwabo and Ndengo (2023), applied an ensemble learning to the prediction of churn among customers in the banking sector. The

study collected data containing information about churn among customers from an online repository provided by Kaggle. The data contained information about 14 features from 10000 customer records. The study performed a single simulation run consisting of 80% for training and 20% for testing the models. The results revealed that kNN showed the best performance with an accuracy of 86.5% among the ensemble learning algorithms however the overall best performance was achieved ranfom forest algorithm with an accuracy of 87.4%. The study failed to consider the impact of feature selection and varying proportion of training and testing datasets selected during simulation on the performance of the model.

de Lima Lemos, Silva, and Tabak (2022), worked on the application of ensemble learning alkgorithms to the determination of customers who were likely to churn. The study collected data containing information about 35 attributes from 500,000 customers which was subjected to a number of data preprocessing tasks with equal distribution of churners and non-churners. The study adopted the use of decision trees (DT), K-nearest neighbours (kNN), logistic regression (LR), support vector machines (SVM) and random forest for the classification of customer churn based on the collected dataset using 10-fold cross validation approach. The results revealed that the best performance was achieved using random forest algorithm with an accuracy of 82.8%. The study did not consider how the size of training dataset can affect the performance of the predictive models since it was limited to a single simulation run.

Dhangar and Anad (2021), assessed the performance of the application of ensemble learning to the prediction of customer churn. The study performed an exhaustive review of literature covering the various algorithms and toold adopted for the devepment of predictive models required for assessing customer churn.

The study collected data from 7045 client records containing information about 21 features. The study adopted a single simulation run for the assessment of logistic regression, Gaussian naïve Bayes, support vector nachines and random forest. The results revealed that the best performance was achieved using random forest algorithm with an accuracy of 87%. The study was limited to a single simuation run and did not consider the impact of the selection of relevant features on the performance of ensemble learning algorithms.

METHODOLOGY

Identification and Collection of Data

This section presents the methods that were adopted for the identification and collection of the dataset adopted for this study. The dataset used in this study was collected from the publicly accessible repository that is provided by Kaggle online via the Internet. It is a public dataset that has been made available for educational and research purposes and is available via https://www.kaggle.com/blastchar/telco-customerchurn The dataset consists of information about a set of features that was collected from 1870 non-churner records and 5173 churner records making a total of 7043 records.

The dataset was downloaded from the repository and stored as a spreadsheet file presented in .csv format. The dataset consists of 20 input features and the target feature. The target feature was represented by a binary value with class values: *churn* and *no churn.* The input features were divided into three groups, namely: demographic information with four (4) features, service-based information with nine (9) features and billing information with seven (7) features.

Table 1. Demographic information of customers.

Feature Name	Description	Data Label
Gender	Customer's gender	Male, Female
Senior Citizen	Is customer an elder?	Yes, No
Partner	Has a partner (or spouse)	Yes, No
Dependents	Has dependents (e.g., child)	Yes, No.

Table 1 shows a description of the demographic information of the customers which is composed of four (4) features. The features include: gender, been a senior citizen, having a partner (or spouse) and having dependents (e.g., child). The gender of the

customer was identified as a binary value with values Male and Female however the other three features were identified as binary value with values Yes and No.

Table 2. Service-based information of customers.

Feature Name	Description	Data Label
Phone	Use phone?	Yes, No
Multiple Lines	Have multiple lines?	Yes, No, NPS
Internet	Use Internet?	DSL, Fiber Optic, No
Online Security	Use online security?	Yes, No, FPS
Online Backup	Use online backup?	Yes, No, FPS
Device Protection	Use device protection?	Yes, No, FPS
Technical Support	Provided technical support?	Yes, No, FPS
Streaming TV	Streams TV?	Yes, No, FPS
Streaming Movies	Streams Movies?	Yes, No, FPS

Table 2 shows the description of the service-based information of the customers which was composed of nine (9) features. The features include knowledge of phone use, having multiple phone lines, using Internet, using online security, using online backup, using device protection, receiving technical support, TV and movie streaming. All the features were represented as categorical data such that the use of phone is binary while the rest are ternary. The use of Internet was represented as either: direct service line (DSL), fiber optic and no while the remaining features were represented as: yes, no and no phone service.

Table 3 shows a description of the account-based information of the customers which is composed of nine (9) features. Some of the features are numeric while others are categorical. The features with numeric data include: customer ID, monthly charges, total amount charged, and time spent using service. The categorical features include: contract represented as monthly, biennial, and yearly; using paper billing was represented as Yes and No while method of payment was presented as: electronic check, mailed check, bank transfer and credit card.

Table 3. Account information of customers.

Feature Name	Description	Data Label
Phone	Use phone?	Yes, No
Multiple Lines	Have multiple lines?	Yes, No, NPS
Internet	Use Internet?	DSL, Fiber Optic, No
Online Security	Use online security?	Yes, No, FPS
Online Backup	Use online backup?	Yes, No, FPS
Device Protection	Use device protection?	Yes, No, FPS
Technical Support	Provided technical support?	Yes, No, FPS
Streaming TV	Streams TV?	Yes, No, FPS
Streaming Movies	Streams Movies?	Yes, No, FPS

Model Simulation

To develop the predictive model that was required for the classification of churn among customers the dataset collected for this study was fed to a few ensemble learning and ensemble learning algorithms using the holdout method. The dataset was split into two parts, namely: training and testing dataset such that the training dataset was used to build the model while the testing dataset was used to validate the performance of the predictive model. The simulation was performed using three (3) simulation runs of varying proportions of the training and testing datasets following which the models were compared based on several performance evaluation metrics. The results of the simulation and evaluation of the predictive models was represented using a diagram called confusion matrix. The evaluation of the performance of the predictive models was done

based on the testing dataset and not the training dataset.

RESULTS AND DISCUSSION

Univariate distribution of features

Table 4 shows the distribution of the categorical features in the dataset which is represented using a frequency distribution table that presents the frequency of the feature labels alongside their respective percentage proportions. Regarding the socio-demographic information, distribution of the gender revealed that majority of the records were male owing to a proportion of 50.48%, majority of the records were non senior citizens owing to a proportion of 83.79%, majority of the records had partners owing to a proportion of 51.70%, and majority of the records showed had no dependents owing to a proportion of 70.04%.

Regarding service-based information, majority of the records used phone service owing to a proportion of 90.32%, majority of the records used multiple lines owing to a proportion of 48.13%, majority of the records used fibre optic for Internet service owing to a proportion of 43.96%, majority of the records were not subscribed to online security owing to a proportion of 49.67%, majority of the records were not subscribed to online backup service owing to a proportion of 43.84%, majority of the records were not subscribed to device protection service owing to a proportion of 43.94%, majority of the records were not subscribed to technical support owing to a proportion of 49.31%, majority of the records were not subscribed to streaming TV service owing to a proportion of 39.90%, and majority of the records were not subscribed to streaming movies service owing to a proportion of 39.54%.

Regarding the customer's accounting information, the distribution of the information about contract revealed that majority of the records were on monthly subscription owing to a proportion of 55.02%, the distribution of the information about paperless billing revealed that majority of the records were not subscribed to paperless billing owing to a proportion of 59.22%, and the distribution of the information about the payment methods revealed that majority of the records used electronic checks owing to a proportion of 33.58.

Table 4. Distribution of the categorical features in the churn dataset.

Variable Name	Label (Nominal)	Frequency	Proportion (%)
Gender	Male	3555	50.48
	Female	3488	49.52
Senior Citizen	No	5901	83.79
	Yes	1142	16.21
Partner	No	3641	51.70
	Yes	3402	48.30
Dependents	No	4933	70.04
	Yes	2110	29.96
Phone Service	Yes	6361	90.32
	No	682	9.68
Multiple Lines	No	3390	48.13
	Yes	2971	42.18
	No Phone Service	682	9.68
Internet Service	Fibre Optic	3096	43.96
	DSL	2421	34.37
	No	1526	21.67
Online Security	No	3498	49.67
	Yes	2019	28.67
	No Internet Service	1526	21.67
Online Backup	No	3088	43.84
	Yes	2429	34.49
	No Internet Service	1526	21.67

Device Protection	No	3095	43.94
	Yes	2422	34.39
	No Internet Service	1526	21.67
Technical Support	No	3473	49.31
	Yes	2044	29.02
	No Internet Service	1526	21.67
Streaming TV	No	2810	39.90
	Yes	2707	38.44
	No Internet Service	1526	21.67
Streaming Movies	No	2785	39.54
	Yes	2732	38.79
	No Internet Service	1526	21.67
Contract	Month-to-Month	3875	55.02
	Two Year	1695	24.07
	One Year	1473	20.91
Paperless Billing	Yes	4171	59.22
	No	2872	40.78
Payment Method	Electronic Check	2365	33.58
	Mailed Check	1612	22.89
	Bank Transfer	1544	21.92
	Credit Card	1522	21.61

Results of the Assessment of Feature Importance

This section presents he results of the assessment of feature importance among the features using two metrics, namely: Pearson's correlation and mutual information. Figure 2 shows the visualization of the feature-feature inter-correlation based on the Pearson's correlation coefficient function. It displays the value of the correlation of the features with respect to one another such that darker colours reflect higher correlation while light colours reflect lower correlations. Also, red colours signified negative correlation and blue colour signified positive correlation. However, since the focus of the study is on the association between the features and the classification of churn among customers, the values displayed in the last row at the bottom were considered. The values in the cell of the last row shows the correlation of the features with respect to the classification of churn.

According to the Pearson's correlation coefficient, it was revealed that the most important feature is *contract* with a high negative linear relationship, this was followed by *tenure, online security, technical support, device protection, payment method, streaming TV, streaming movies* and *paperless billing* all with a negative correlation, followed by *monthly charges, partner senior citizens, Internet service, multiple lines, phone service* and *gender* all with positive correlation.

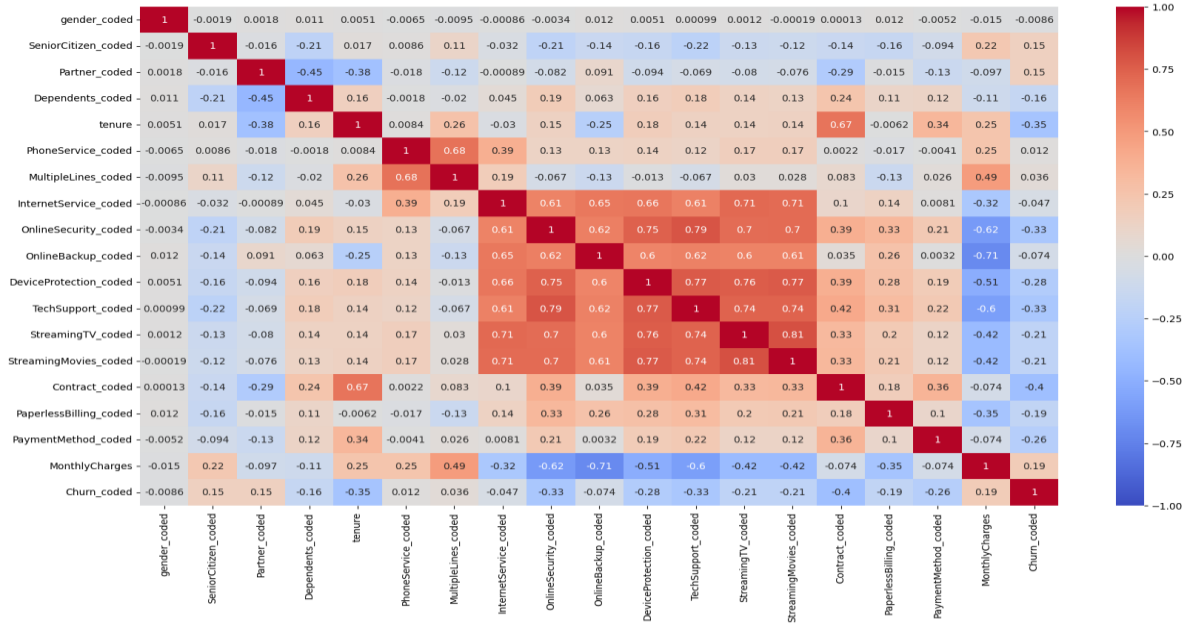

Figure 2. Visualization of heatmap of feature-feature intercorrelation.

Table 5 gives a summary of the ranking of features based on the value of their Pearson's correlation and mutual information.

Table 5. Identification of Feature Importance.

S/N	Pearson's Correlation		Mutual Information	
	Feature Name	Value	Feature Name	Value
1	Contract	-0.40000	Contract	0.101538
2	Tenure	-0.35000	Tenure	0.070407
3	Technical support	-0.33000	Technical support	0.063567
4	Online security	-0.33000	Online security	0.054984
5	Device protection	-0.28000	Payment method	0.052613
6	Payment method	-0.26000	Internet service	0.051970
7	Streaming TV	-0.21000	Online backup	0.050452
8	Streaming movies	-0.21000	Monthly charges	0.047425
9	Monthly charges	0.19000	Device protection	0.039607
10	Paperless billing	-0.19000	Streaming TV	0.032245
11	Dependents	-0.16000	Streaming movies	0.027995
12	Partner	0.15000	Paperless billing	0.019559
13	Senior citizen	0.15000	Partner	0.014404
14	Online backup	-0.07400	Dependents	0.013318
15	Internet service	-0.04700	Gender	0.012830
16	Multiple lines	0.03600	Phone service	0.009405
17	Phone service	0.01200	Senior citizen	0.009272
18	Gender	-0.00860	Multiple lines	0.005035

The results of the use of Pearson's correlation revealed the existence of a linear relationship between the features and classification of churn among customers. In most cases, most biological features hardly possess a linear relation but instead they possess a non-linear relationship between one another as revealed using the mutual information.

Results of the Simulation of Predictive Model

This section presents the results of the application of the three ensemble learning algorithms, namely: LightGBM and Bagging with Decision Trees classifier. The model simulation was conducted by splitting the dataset into two parts, train and test dataset using five simulations such that 60/40, 70/30, and 80/20 percent of the dataset was adopted for training/testing the predictive model. Table 6 shows the number of records that were adopted for each simulation that were considered in this study. As stated earlier, the train datasets were used to build the predictive model while the test data were used to evaluate the performance of the predictive models based on a number of performance evaluation metrics.

In simulation 1, 60% of the dataset was adopted for training and 40% was adopted for testing such that the train data consisted of 3753 no churn records and 472 churn record while the test data consisted of 1421 no churn and 1397 churn records. In simulation 2, 70% of the dataset was adopted for training and 30% was adopted for testing such that the train data consisted of 4103 no churn records and 827 churn record while the test data consisted of 1071 diabetes present and 1042 churn records. In simulation 3, 80% of the dataset was adopted for training and 20% was adopted for testing such that the train data consisted of 4473 no churn records and 1161 churn record while the test data consisted of 701 diabetes present and 708 churn records.

Table 6. Description of the number of records adopted for training and testing predictive models across three simulations.

Simulation#	Train Data			Test Data		
	No churn	**Churn**	**Total**	**No churn**	**Churn**	**Total**
Simulation 1 (60/40)	3753	472	**4225**	1421	1397	**2818**
Simulation 2 (70/30)	4103	827	**4930**	1071	1042	**2113**
Simulation 3 (80/20)	4473	1161	**5634**	701	708	**1409**

Results of the Evaluation of the Predictive Model

This section presents the results of the evaluation of the predictive models that were generated across the three simulations based on the ensemble learning techniques that were adopted in this study. The results are presented for each simulation following which the results of the performance of the algorithms were presented.

Evaluation of predictive models in simulation 1

As stated earlier, simulation 1 was validated using a dataset that was composed of 40% of the dataset and consisted of 1421 Non-churn and 1397 Churn records. Figure 3 shows the confusion matrices that were used to interpret the results of the evaluation of each predictive model developed in simulation 1 based on the test dataset. Using LightGBM classifier, it was observed that all 1421 Non-churn records were correctly classified while 1395 out of 1395 Churn records were correctly classified while 2 were incorrectly classified as no churn records owing to an accuracy of 99.93%. Using Bagging with DT classifier, it was observed that 1420 out of 1421 Non-churn records were correctly classified while 1 was not correctly classified and 1395 out of 1397 Churn records were correctly classified while 2 were incorrectly classified owing to an accuracy of 99.89%.

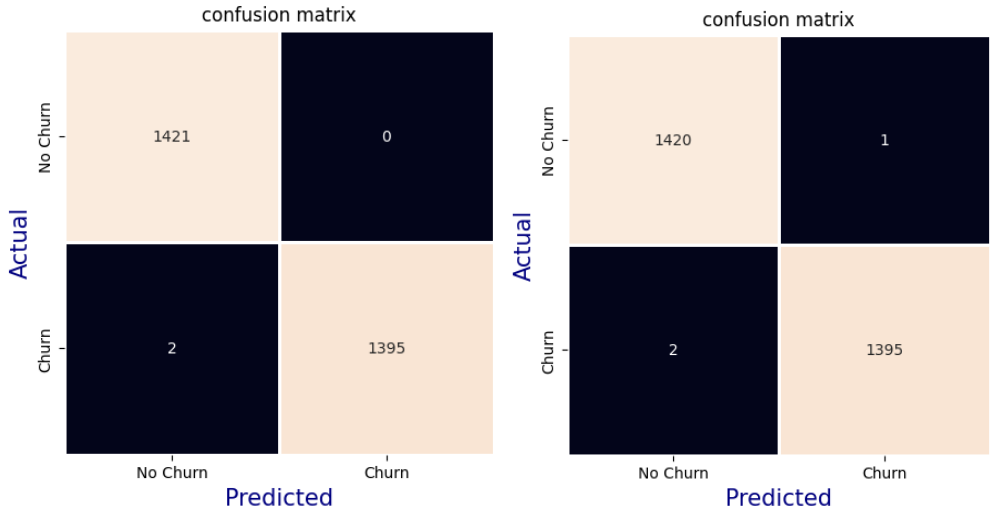

Figure 3. Confusion matrices for the evaluation of LightGBM (left) and Bagging with DT Classifier (right) for simulation 1.

Evaluation of predictive models in simulation 2

As stated earlier, simulation 2 was validated using a dataset that was composed of 30% of the dataset and consisted of 1071 Non-churn and 1042 Churn records. Figure 4 shows the confusion matrices that were used to interpret the results of the evaluation of each predictive model developed in simulation 2 based on the test dataset. Using LightGBM classifier, it was observed that 1071 Non-churn records were correctly classified and 1041 out of 1041 Churn records were correctly classified while 1 was classified as Non-churn records owing to an accuracy of 99.95%. Using Bagging with DT classifier, it was observed that all 1071 Non-churn records were correctly classified and all 1042 Churn records were correctly classified owing to an accuracy of 100%.

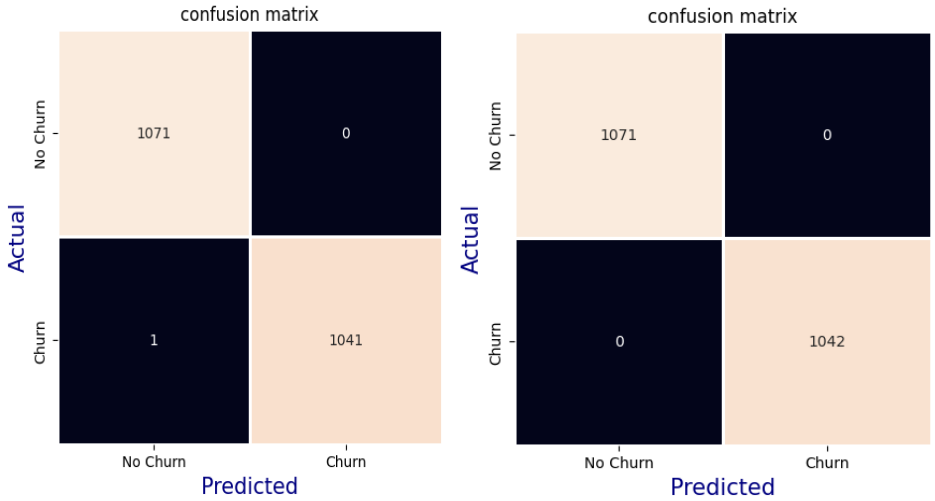

Figure 4. Confusion matrices for the evaluation of LightGBM (left) and Bagging with DT Classifier (right) for simulation 2.

Evaluation of predictive models in simulation 3

As stated earlier, simulation 3 was validated using a dataset that was composed of 20% of the dataset and consisted of 701 Non-churn and 708 Churn records. Figure 5 shows the confusion matrices that were used to interpret the results of the evaluation of each predictive model developed in simulation 3 based on the test dataset. Using LightGBM classifier, it was observed that all 701 Non-churn records were correctly classified and all 708 Churn records were correctly classified owing to an accuracy of 100%. Using DT classifier, it was observed that all 701 Non-churn records were correctly classified and all 708 Churn records were correctly classified owing to an accuracy of 100%.

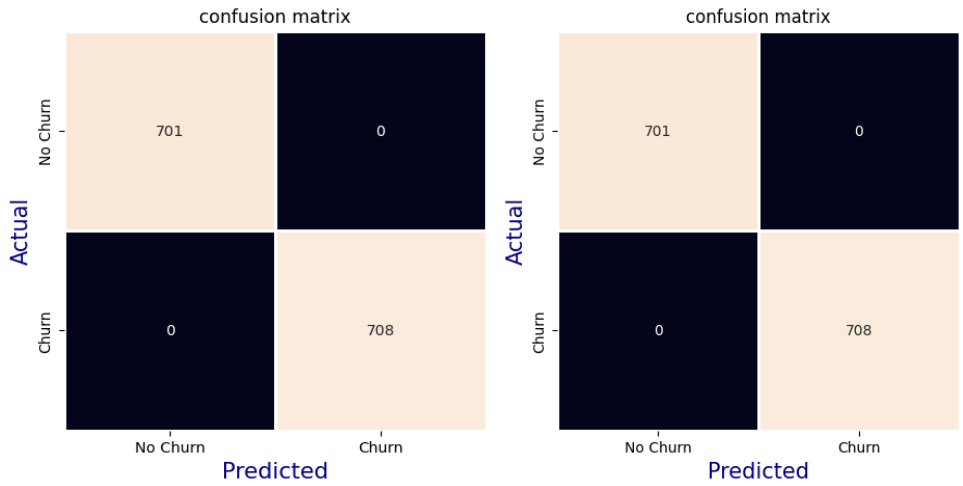

Figure 5. Confusion matrices for the evaluation of LightGBM (top-left) and Bagging with DT Classifier (bottom) for simulation 3.

DISCUSSION OF RESULTS

This section presents the discussion of the results of the various analyses that were performed in this study for the development of a predictive model required for the classification of churn among customers. The results of the study showed that among the ensemble methods adopted, the LightGBM classifier proved to have the best overall performance among all the ensemble models considered. However, it was observed that as the proportion of the training dataset increased, the performance of the machine learning classifiers improved.

It was revealed in this study that the ensemble learning algorithms selected in this study all showed very good performance in the classification of churn. The results of this study showed better performance than that of the study conducted by Malik, Runwal, Shah, Raut, and Hire (2023) which had an accuracy of 90.3% using Light GBM classifier and in the study by Sindikubwabo and Ndengo (2023) who showed an accuracy of 87.4% using random forest classifier and in the study by de Lima Lemos, Silva, and Tabak (2022), who showed an accuracy of 82.8%. However, the results of this study supported the view promoted in the study by Adrin and Fadaralika (2020) and Dhangar and Anad (2021) regarding their view concerning the effectiveness of ensemble models for development of predictive models. Table 7 presents the summary of the results of the evaluation of the performance of the ensemble learning approaches considered in this study.

Table 7. Results of the evaluation of the predictive models across five simulations based on performance metrics.

Simulation# (Test records)	Algorithm	Correct Records	Accuracy (%)	Precision		Recall		F1-score	
				No Churn	Churn	No Churn	Churn	No Churn	Churn
Simulation 1 (2818 records)	**LightGBM**	2816	99.93	1.000	1.000	1.000	1.000	1.000	1.000
	Bagging with DT	2815	99.89	1.000	1.000	1.000	1.000	1.000	1.000
Simulation 2 (2113 records)	**LightGBM**	2112	99.95	1.000	1.000	1.000	1.000	1.000	1.000
	Bagging with DT	2113	100.00	1.000	1.000	1.000	1.000	1.000	1.000
Simulation 3 (1409 records)	**LightGBM**	1409	100.00	1.000	1.000	1.000	1.000	1.000	1.000
	Bagging with DT	1409	100.00	1.000	1.000	1.000	1.000	1.000	1.000

CONCLUSION

The study identified that each feature had a relative importance to one another regarding their usefulness in the classification of churn. However, it was observed that the information provided by the mutual information metric seemed to be more reliable compared to the information provided by the Pearson's correlation since this is focused on assessing linear relationship which is hardly the case in medical data. The study concluded that ensemble learning models proved effective in the development of predictive models for the classification of churn among customers. The study revealed that the most important feature was the service-based information. Also, it was observed that the LightGBM classifier showed the nest performance among the ensemble models considered in this study.

REFERENCE

Adrin, M., & Fadaralika, D. (2020). A comparative model for predicting customer churn using supervised ensemble learning. *International Journal of Science and Research, 11*(2), 133-137.

Dhangar, K., & Anad, P. (2021). A review on customer churn prediction using ensemble learning approach. *International Journal of Innovations in Engineering Research and Technology, 8*(5), 193-201.

Gore, S., Chibber, Y., Bhasin, M., Mehta, S., & Suchitra, S. (2023). *Customer Churn Prediction using Neural Networks and SMOTE-ENN for Data Sampling.* Institute of Electrical and Electronics Engineers Incorporated.

Imani, M. (2020). Customer Chrun Prediction in Telecommunication using Ensemble learning: A Comparison Study. *AUT Journal of Modeling and Simulation, 52*(2), 229-250.

Joolfoo, M., Jugurnauth, R., & Joolfoo, K. (2020). Customer churn prediction in telecom using ensemble learning in big data platform. *Journal of Critical Reviews*, 1-13.

Kumar, A., & Jain, M. (2020). *Ensemble Learning for AI Developers: Learn Bagging, Stacking, and Boosting Methods with Use Cases.* Apress.

Lalwani, P., Mishra, M., Chadha, J., & Sethi, P. (2022). Customer churn prediction system: a ensemble learning approach. *Computing, 104*(2), 271-294.

Malik, S., Runwal, S., Shah, Y., Raut, V., & Hire, S. (2023). A study on customer churn prediction. *International Research Journal of Modernization in Engineering Technology and Science, 5*(4), 3557-3575.

Nhu, N., Ly, T., & Son, D. (2022). Churn prediction in telecommunication industry using kernel cupport vector machines. *PLoS ONE, 17*(5), 1-13.

Rahaman, M. M., Rani, S., Islam, M. R., & Bhuiyan, M. M. R. (2023). Machine Learning in Business Analytics: Advancing Statistical Methods for Data-Driven Innovation. *Journal of Computer Science and Technology Studies*, *5*(3), 104-111. https://doi.org/10.32996/

Sindikubwabo, E., & Ndengo, M. (2023). Analysis and prediction of customer churn using ensemble learning - a case study in the banking sector. *IOSR Journal of Computer Engineering, 25*(4), 27-33.

Yağcı, M. (2022). Educational Data Mining: Prediction of Students' Academic Performance using Machine Learning Algorithms. *Smart Learning Environments, 9*(11), 1-19.