Malaysian Journal of Computing, 8 (2): 1602-1619, 2023

eISSN: 2600-8238

PREDICTION OF CUSTOMER CHURN FOR ABC MULTISTATE BANK USING MACHINE LEARNING ALGORITHMS

Hui Shan Hon[1], Khai Wah Khaw[2*], XinYing Chew[3], Wai Peng Wong[4]

[1,2*] *School of Management, Universiti Sains Malaysia, 11800 Minden, Pulau Pinang, Malaysia*

[3] *School of Computer Sciences, Universiti Sains Malaysia, 11800 Minden, Pulau Pinang, Malaysia*

[4] *School of Information Technology, Monash University, Malaysia Campus, Selangor, Malaysia*

[1]honhuishan@gmail.com, [2*]khaiwah@usm.my, [3]xinying@usm.my, [4]waipeng.wong@monash.edu

ABSTRACT

Customer churn is defined as the tendency of customers to cease doing business with a company in a given period. ABC Multistate Bank faces the challenges to hold clients. The purpose of this study is to apply machine learning algorithms to develop the most effective model for predicting bank customer churn. In this study, six supervised machine learning methods, K-Nearest Neighbors, Support Vector Machine, Naïve Bayes, Decision Tree, Random Forest, and Extreme Gradient Boosting (XGBoost), are applied to the churn prediction model using Bank Customer Data of ABC Multistate Bank obtained from Kaggle. The results showed that XGBoost outperformed the other six classifiers, with an accuracy rate of 84.76%, an F1 score of 56.95%, and a ROC curve graph of 71.64%. The bank may use XGBoost model to accurately identify customers who are at risk of leaving, concentrate their efforts on them, and possibly make a profit. Future research should focus on various machine learning approaches for determining the most accurate models for bank customer churn datasets.

Keywords: *Bank Customer Churn, Machine Learning, Supervised Machine Learning.*

Received for review: 19-06-2023; Accepted: 24-08-2023; Published: 10-10-2023
DOI: 10.24191/mjoc.v8i2.21393

1. Introduction

Nowadays, due to the diversification of sales strategies brought about by technological advancements and the ensuing severe market competition, businesses are now required to move their attention from their products to their customers, as the customer is considered the real ruler of the market (Khodabandehlou & Rahman, 2017). Customer churn (customer attrition), defined as the tendency of customers to cease doing business with a company in a given period (Chandar et al., 2006), has become a significant issue for ABC Multistate Bank. Churn in the banking sector is not only due to fierce competition but also due to increased choices available to customers, causing banks to feel pressure to maintain customer relationships in the long term. Predicting the customer churn rate in a business is important as acquiring new customers is more costly, sometimes five times the cost of retaining existing customers (Hung et al., 2006). Predicting the likelihood of customer churn will serve as a guide for the bank in developing different marketing strategies for different client bases while satisfying their needs. Banks might develop proactive marketing campaigns to keep their existing customers from leaving. As a result of predicting customer churn, the bank can build a blueprint for its future revenue and assist its businesses in identifying and improving areas where customer service is lacking.

Data mining is the process of examining large datasets to find hidden patterns, relationships, or even anomalies in the data (Pisal, 2022). Machine learning is a subfield of artificial intelligence (AI) and computer science that uses data and algorithms to replicate how humans learn, gradually improving its accuracy. It can be used in a wide range of fields, including pavement performance prediction (Marcelino et al., 2021), spectroscopy (Meza Ramirez et al., 2021), medical field (Kaur & Kumari, 2022), machining industries (Cica et al.,2020), employee promotion prediction (Shafie et al., 2023) and in finance (Malhotra et al., 2020; Malik et al., 2022). Machine learning is mostly used in customer churn prediction to assist firms in better predicting customer attrition and providing targeted and effective marketing strategies for possible churners. However, customer churn prediction might be challenging for machine learning for various reasons:

- Imbalance data sets have the issue that they often misclassify minority instances more often than majority instances (Zhu et al., 2018).
- Data heterogeneity (Tékouabou et al., 2022).

The main contribution of this paper is the development of machine learning classification models for the Bank Customer Data acquired from Kaggle in order to predict customer churn for ABC Multistate Bank. The six classification methods of supervised machine learning used in this paper include K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Naïve Bayes (NB), Decision Tree (DT), Random Forest (RF), and Extreme Gradient Boosting (XGBoost). The accuracy and performance of these methods will be compared in order to identify the most appropriate methods for future researchers and business owners to accurately predict customer behaviour. The scope of the study is limited to supervised machine learning classification in bank customer churn prediction as the predicted target will fall into one of two specified categories: churned and non-churned. To evaluate and compare the prediction performance of the machine learning algorithms, a real-world dataset was utilised along with several evaluation metrics.

The remainder of this paper is organized as follows. Section 2 presents a literature review and Section 3 discusses in detail the methods and materials including the data collection and preparation. Section 4 discusses the model conducted in this study and the obtained results. The discussion is outlined in Section 5. Finally, the recommendations and conclusion are presented in Section 6.

2. Literature Review

This section reviews some work related to supervised machine learning classification algorithms applied for the prediction of customer churn.

Among the extensively used supervised algorithms in the context of churn prediction are K-Nearest Neighbors (KNN), Naïve Bayes (NB), Linear Regression, Logistic Regression, Linear Discriminant Analysis (LDA), Decision Tree (DT) (Keramati et al., 2016) and Support Vector Machine (SVM) (Zhao & Dang, 2008; He et al., 2014). Various studies propose using ensembles to solve the churn prediction issue, such as Random Forest (RF) (de Lima Lemos et al., 2022), AdaBoost, Gradient Boosting, or XGBoost (Guliyev & Tatoğlu, 2021).

Ha Thanh and Vy (2022) utilised three supervised models, which are KNN, RF, and Gradient Boosting, to predict the churn rate of customers using mobile banking services in Vietnam and to determine the most effective measurement for forecasting churners. The dataset was obtained from the Vietnam International Commercial Joint Stock Bank, which holds 66,736 records of client activity and the status of their financial activities. Gradient Boosting offers the highest accuracy (79.71%) and the best performance in the three above classifiers, as well as an 86.23% ROC (Receiver Operating Characteristic) curve graph. Furthermore, the decision tree algorithm produces eligible criteria for churner and non-churner classification, which may be useful to managers.

Rahman and Kumar (2020) worked with KNN, SVM, DT, and RF classifiers to validate the best model for predicting client churn in a startup bank. By analysing customer behaviour, the study encourages the investigation of the likelihood of churn. The dataset used in this study was obtained through Kaggle, and it contained 13 features and a small number of records (10,000 samples), and it was highly imbalanced. The author used feature selection methods such as Minimum Redundancy Maximum Relevance (mRMR) and the Relief algorithm to find more relevant features and simplify the models. The author has also used oversampling and undersampling to configure the class distribution of the given data. The results show that the RF classifier, after oversampling, performed better in terms of accuracy (95.74%) than the other models. Furthermore, because the bank dataset is imbalanced, SVM is unable to handle the data adequately, as oversampling lowers the score.

Guliyev and Tatoğlu (2021) used real banking data to estimate the explainable machine learning model and test data to evaluate a variety of machine learning models. The AUC values were used by the authors in this study to evaluate the performance of each classifier. High AUC values indicate a machine learning predictive model is effective. Since the study's dataset is unbalanced, the authors used tree-based models like DT, RF, and XGBoost to address these issues. The results show that, with an AUC of 96.97%, the XGBoost model outperformed other machine learning algorithms in classifying churn customers, with the RF model ranking second.

Kaur and Kaur (2020) investigate the use of Logistic Regression (LR), DT, KNN, and RF techniques in predicting the likelihood of churning customers. The authors examined the effectiveness of each model using a dataset from Kaggle that included 28,382 records and 21 features. Exploratory Data Analysis (EDA) is used in the pre-processing stage to discover missing values, normalise and scale data, and determine the relationship between two variables so that redundant features can be removed. The results of the models are compared using recall value, precision, ROC AUC, and accuracy. In order to improve the model's performance with less accuracy, ensembling techniques like averaging and max voting are used. Finally, Random Forest outperforms all of the models tested.

3. Methodology

This paper aims to predict the customer churn for ABC Multistate Bank using machine learning algorithms.

3.1 Dataset Description

Kaggle provided the churn dataset used in this analysis. Figure 1 shows the dataset contains information on 10,000 bank customers from European countries, with the target parameter being a binary variable indicating whether the customer has left the bank or is still a customer. The data is significantly imbalanced, with 7963 positive class (non-churned) samples and 2037 negative class (churned) samples; the imbalance rate is roughly 25.6%. The dummy variable will be 1 when the client closes his/her bank account. When the client is retained, the dummy variable 0 is displayed. The dataset contains twelve feature vectors (predictors) derived from customer data and transactions that the customer processed. Table 1 provides more details on these features.

3.2 Data Preprocessing

Data preprocessing is a vital step in the data mining process because it directly affects the task success rate. Real-world data is messy because it is frequently generated, processed, and stored by numerous people, business processes, and software programs. An incomplete data set, manual input errors, duplicate data, or the use of multiple names to describe the same thing are all possible outcomes from real-world data. The descriptions of the predictors after preprocessing are presented in Table 2, and these are the features used in this study to evaluate churn prediction. In this stage, the data will be transformed into a format that can be handled by data mining more quickly and efficiently. As part of our preparation, we handled data cleaning, selected features, transformed categorised features and scaled features.

	customer_id	credit_score	country	gender	age	tenure	balance	products_number	credit_card	active_member	estimated_salary	churn
0	15634602	619	France	Female	42	2	0.00	1	1	1	101348.88	1
1	15647311	608	Spain	Female	41	1	83807.86	1	0	1	112542.58	0
2	15619304	502	France	Female	42	8	159660.80	3	1	0	113931.57	1
3	15701354	699	France	Female	39	1	0.00	2	0	0	93826.63	0
4	15737888	850	Spain	Female	43	2	125510.82	1	1	1	79084.10	0
...	...	...	...	...	...	...	...	...	...	...	...	...
9995	15606229	771	France	Male	39	5	0.00	2	1	0	96270.64	0
9996	15569892	516	France	Male	35	10	57369.61	1	1	1	101699.77	0
9997	15584532	709	France	Female	36	7	0.00	1	0	1	42085.58	1
9998	15682355	772	Germany	Male	42	3	75075.31	2	1	0	92888.52	1
9999	15628319	792	France	Female	28	4	130142.79	1	1	0	38190.78	0

Figure 1. The dataset consists of 10,000 customer data.

Table 1. Dataset description.

Feature Name	Feature Description
customer_id	Unique IDs for bank customer identification
credit_score	A credit score of the customer
country	The country from which the customer belongs.
gender	Male or Female.
age	Age of the customer.
tenure	Number of years for which the customer has been an ABC Bank account holder.
balance	Bank account balance of the customer
products_number	A number of bank products the customer is utilizing (savings account, mobile banking, internet banking etc.).
credit_card	Binary flag for whether the customer holds a credit card with the bank or not.
active_member	Binary flag for whether the customer is an active member of the bank or not.
estimated_salary	Estimated salary of the customer in Dollars.
churn	Binary flag 1 if the customer has left the bank during some period and 0 if the customer is retained.

Table 2. Preprocessed dataset description.

Feature Name	Min	Max	Mean	STD
credit_score	350	850	650.5288	96.653299
age	18	92	38.9218	10.487806
tenure	0	10	5.0128	2.892174
balance	0	250898.09	76485.889288	62397.405202
products_number	1	4	1.5302	0.581654
credit_card	0	1	0.7055	0.45584
active_member	0	1	0.5151	0.499797
estimated_salary	11.58	199992.48	100090.239881	57510.492818
churn	0	1	0.2037	0.402769

3.2.1 Data Cleaning

Data cleaning is the process of deleting duplicate data, organising raw data, and filling in missing variables. Data cleaning in data mining allows users to identify inaccurate or incomplete data before business analysis and insights. If bad data is not cleaned, the accuracy of the final analysis may be compromised, and misleading conclusions may be drawn. As shown in Figure 2, the dataset utilised in this paper is unaffected by missing data.

3.2.2 Feature Selection

Feature selection methods are widely used in data processing fields such as machine learning because they reduce computing complexity while classifying high-dimensional data (Patro et al., 2022). It only keeps the features that are valuable by deleting irrelevant and redundant features (Guan et al., 2014). Data or features that are conditionally independent or have no bearing on the outcome of the topic under discussion are considered irrelevant. Keeping such attributes may decrease and impact classifier accuracy and performance. The features named customer_id and country were manually removed when analysing the churn dataset because they had no direct impact on the prediction. A correlation-based filter was also utilised to evaluate the relationship between the numerical features. Figure 3 shows that age has the highest positive correlation (0.28) with churned customers, followed by balance (0.12) and estimated_salary (0.012). The lowest correlation coefficient (-0.3) was recorded between the balance and the product number. There is no moderate or strong correlation among bank churn predictors. Therefore, all the predictors will be used to estimate the bank customer churn for ABC Multistate Bank.

3.2.3 Encoding Categorical Features

The majority of machine learning methods require the transformation of categorical data into numerical data in order to operate properly. This means that the categorical attribute in the churn dataset, which is gender, must be transformed into numbers before developing a prediction model. There is no preference order for the labels in categorical attributes. Machine learning models will mistakenly assume that the data contains a hierarchy since the data was composed of string labels. One Hot Encoding technique is a method for assigning a numerical value to these labels. Each category is transformed into one binary variable, with each observation indicating whether the classification level is present (1) or absent (0). Figure 4 is an instance of how one-hot encoding functions. There will be two columns labelled gender_Female and gender_Male that distinguish the categorical attribute Gender with the categories Female and Male. The value in the gender_Male columns will be 1 in the Male column and 0 in the Female column, and vice versa. This technique can be implemented using the get_dummies function in the Pandas software library.

```
customer_id         0
credit_score        0
country             0
gender              0
age                 0
tenure              0
balance             0
products_number     0
credit_card         0
active_member       0
estimated_salary    0
churn               0
dtype: int64
```

Figure 2. Missing value for each column in the dataset.

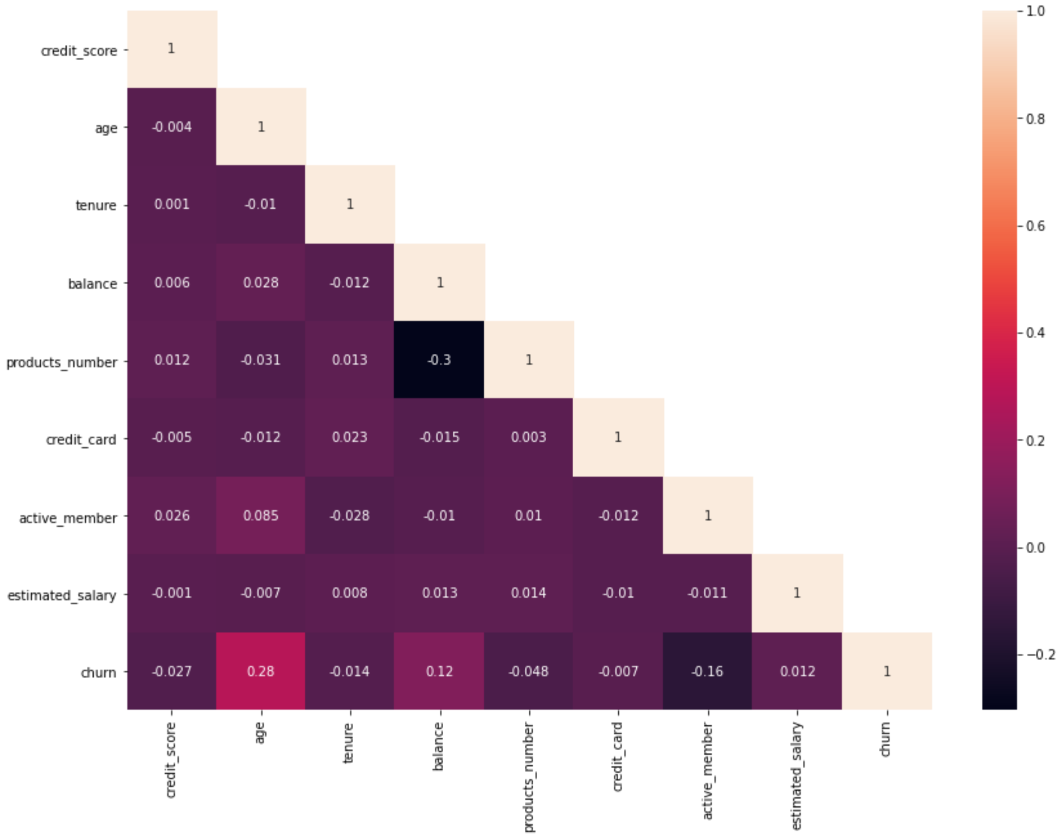

Figure 3. Correlation-based filter.

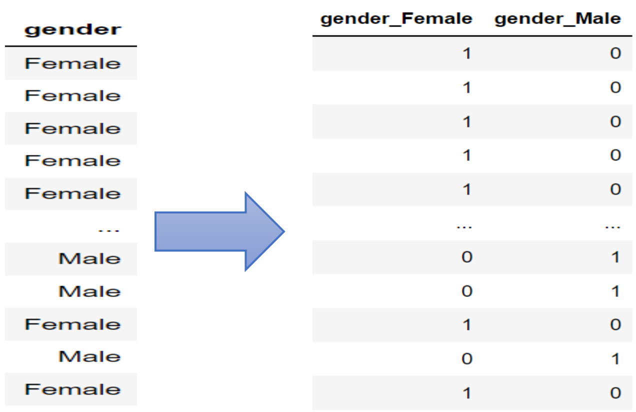

Figure 4. One-hot encoding.

3.2.4 Feature Scaling

Feature Scaling is a technique to normalise the independent features present in the data on the same scale. It is typically conducted during the data pre-processing step to manage highly variable magnitudes, values, or units. Without feature scaling, the machine learning model would give higher weights to higher values and lower weights to lower values. It also takes a long time to train the machine learning model. In this research, the StandardScaler function from the Sci-Kit Learn library is imported to neutralise the range of values for the feature variables. The Standardscaler assumes that data within each feature is normally distributed and rescales it so that it has a distribution with a mean value of 0 and a variance of 1.

3.3 Classification

The classification algorithms were applied to the preprocessed data. *K*-Nearest Neighbors (KNN), Support Vector Machine (SVM), Naïve Bayes (NB), Decision Tree (DT), Random Forest (RF), and Extreme Gradient Boosting (XGBoost) are used for comparison of results.

3.3.1 *K*-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) Classification is a technique that is frequently beneficial for taking more than one neighbour into account where k nearest neighbours are applied to determine the class (Cunningham & Delany, 2021). There are several methods for determining 'k' values, but we can simply run the classifier multiple times with different values to see which produces the best results (Taunk et al., 2019).

3.3.2 Support Vector Machine (SVM)

SVM is a binary classification algorithm based on statistical learning theory that is suitable for processing high-dimensional datasets. SVM is better suited for predicting datasets with small sample sizes (Yan et al., 2022). The SVM algorithm seeks to find a hyperplane in an N-dimensional space that clearly classifies input points. The dimensions of the hyperplane will vary depending on the number of features.

3.3.3 Naïve Bayes (NB)

The Bayes theorem, which states that many variables are not dependent on one another, was used to develop Naïve Bayes (NB) (Costache et al., 2022). Due to its conditional independence assumption, Naive Bayes avoids the problem of most attribute value combinations not being present in the training data or being present in insufficient quantities (Chen et al., 2020). In light of this strict independence requirement, Naive Bayes is an extremely capable classifier in a wide range of real-world applications.

3.3.4 Decision Tree (DT)

Continually separating data based on a particular parameter is a type of supervised machine learning known as a decision tree. The decision tree algorithm is one of the most fundamental and widely used classification algorithms. It can be used to solve regression and classification issues. By learning simple decision rules from training data, this algorithm aims to produce a training model that can predict the class or value of the target variable.

3.3.5 Random Forest (RF)

Random forests are one of the most widely used machine learning techniques for problem prediction (Zhang et al., 2020). A variety of classifiers are generated by the RF ensemble model, which aggregates the results to conclude (Breiman, 2001). The term "ensemble model" refers to a method of creating multiple models and combining the output of each model to produce conclusions (Oh et al., 2022). The model is less susceptible to overfitting problems because RF creates multiple classification trees using subsamples and predictors that are chosen at random (Berk, 2012).

3.3.6 Extreme Gradient Boosting (XGBoost)

XGBoost is a variant of the gradient boosting methods that are primarily used to solve complex regression and classification problems (Sahin, 2020). Asselman et al. (2021) describe the XGBoost ensemble method as a technique in which new models are constructed to correct the residuals or errors of previous models before being combined to produce the final prediction. In experiments, XGBoost outperforms many other ensemble classifiers (such as AdaBoost). It efficiently processes data and significantly reduces the risk of model overfitting (Gumus & Kiran, 2017).

3.4 Performance model evaluation

3.4.1 Accuracy

Accuracy is defined as the ratio of the number of accurate predictions to all other predictions. The confusion matrix is used to calculate the accuracy rate (Table 3). The definitions of True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) are defined in Section 4.2. A good model should be highly accurate. However, reaching high or 100% accuracy on a machine learning model does not imply that the model is well-built; rather, it is usually an indication of a problem, such as overfitting.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \quad (1)$$

Table 3. Confusion Matrix.

	Predicted Negative	Predicted Positive
Actual Negative	True Negative (TN)	False Positive (FP)
Actual Positive	False Negative (FN)	True Positive (TP)

3.4.2 Precision

Precision is also one of the methods used to measure the accuracy of outcomes. It can be defined as the number of correct outputs given by the model or the percentage of all positive classes accurately predicted by the model.

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

3.4.3 Recall

The recall is defined as the percentage of positive classes predicted correctly by the model out of all positive classes. The recall rate should be as high as possible.

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

3.4.4 F1 Score

The F1 Score is also known as the F Score or the F Measure. The F1 Score is used to evaluate the recall and precision simultaneously in cases where two models have low precision and high recall, or vice versa.

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

3.4.5 Area Under the Curve-Receiver Operating Characteristics (ROC AUC)

The Receiver Operating Characteristic (ROC) curve is a graph that displays a classifier's performance across all classification thresholds. The true positive rate (on the Y-axis) and the false positive rate (on the X-axis) are plotted on a graph. AUC (area under the ROC curve) is a well-known approach for comparing classifier performance in two-class tasks (Wang & Chen, 2020). The AUC was created along with the Success Rate and Prediction Rate (Costache et al., 2019). An AUC of 1 indicates a perfect classifier model (Li et al., 2018).

4. Result

After data preprocessing, the data will be in the operational form. Among the bank customer churn dataset that contained 10,000 samples, 25% of the data will be used for testing and the remaining 75% will be used for training. Six supervised machine learning algorithms, namely KNN, SVM, NB, DT, RF, and XGBoost will be analysed by using Python. The performance of the models will be presented in this section.

4.1 Data Visualization

Figure 5 shows the percentage of bank customer churn for ABC Multistate Bank. Among bank customers, 79.63% have not left the bank, while 20.37% have left at a certain point.

Figure 6 shows the churned customer per gender for ABC Multistate Bank. Males were more likely than females to continue using the bank's services among these customers. However female customers churn at a higher rate than male customers.

Figure 7 shows the churned customer per age for ABC Multistate Bank. The majority of the bank's customers are between the ages of 29 and 41. Customers between the ages of 40 and 52 have a higher churn rate.

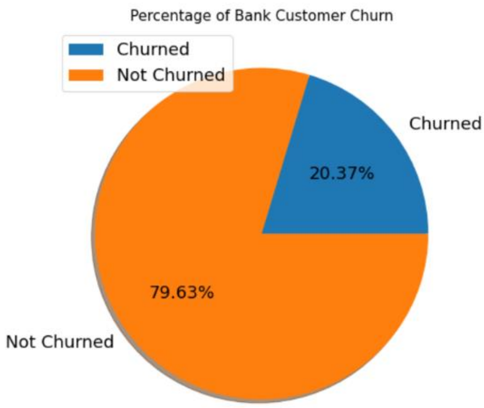

Figure 5. Percentage of bank customer churn.

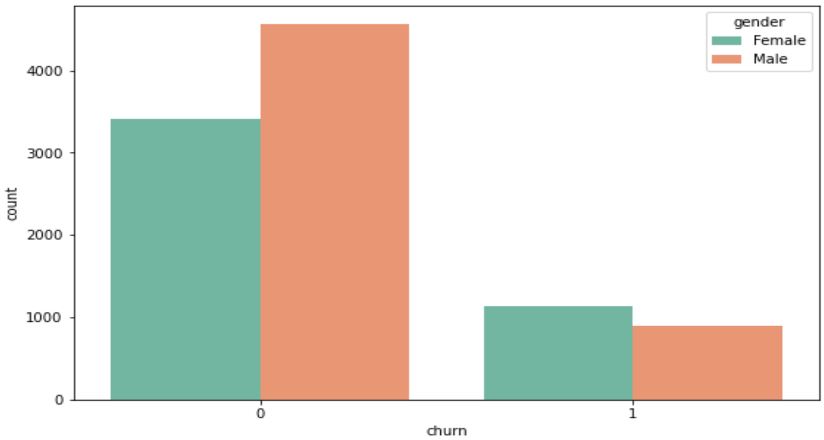

Figure 6. Churned customers per gender.

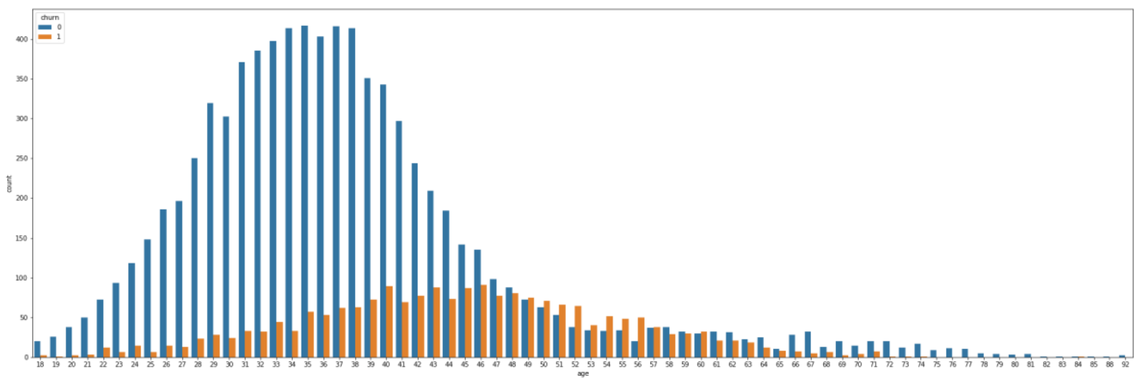

Figure 7. Churned customers per age.

4.2 Confusion Matrix

A confusion matrix is a method for summarising the classification algorithm's performance (Patgiri et al., 2018). The information contained in a confusion matrix is used to calculate accuracy, precision, and recall. The following are explanations for terms used in the applied metrics:

- TP denotes the number of customers correctly classified as churners.
- TN denotes the number of customers correctly classified as non-churners.
- FP denotes the number of customers who are actually non-churners classified as churners.
- FN denotes the number of customers who are actually churners classified as non-churners.

The confusion matrix demonstrates a confusion matrix with six models used by the authors: KNN, SVM, NB, DT, RF, and XGBoost. The authors implement performance metrics based on the above matrix. The confusion matrix is shown in Figure 8.

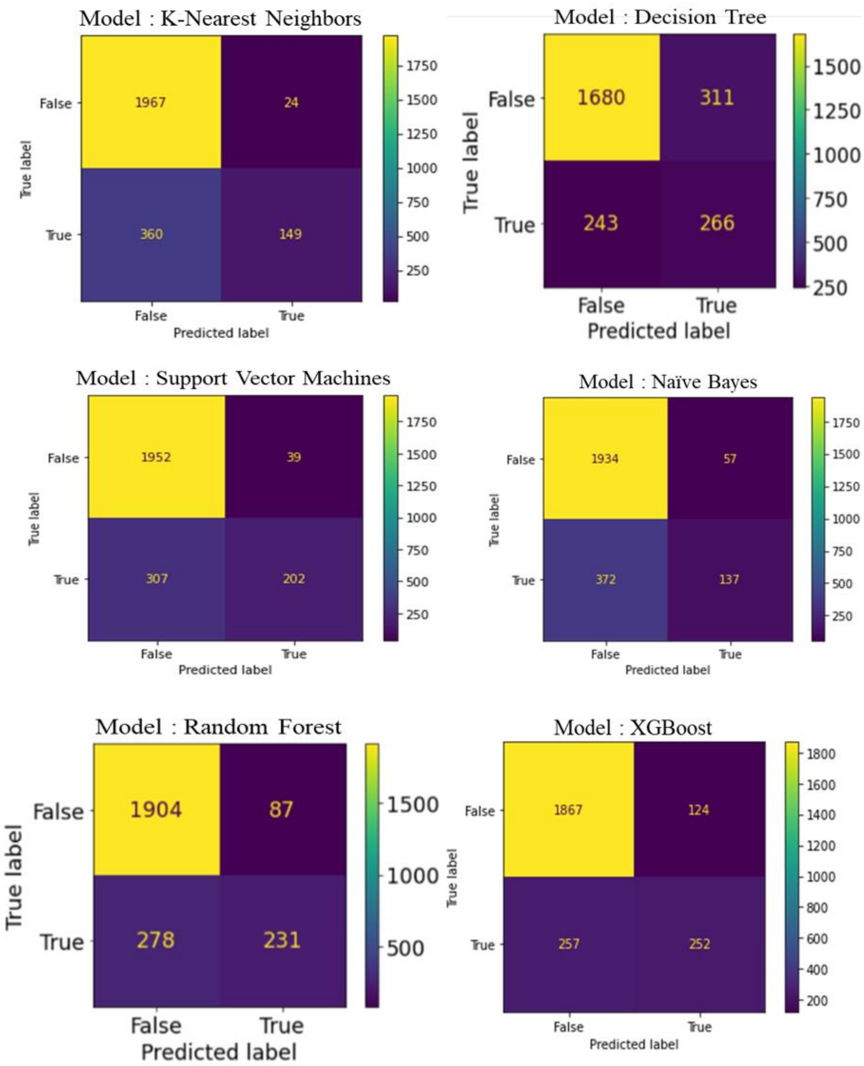

Figure 8. Confusion matrix.

4.3 Performance Results

Table 4 depicts the accuracy, ROC AUC, F1-score, Recall, and Precision performance results for the six models. In terms of accuracy, five of the models achieved an accuracy rate greater than 80%, with the SVM model achieving the highest (86.16%), followed by the RF (85.52%), XGBoost (84.76%), KNN (84.64%), and NB (82.84%). DT has the lowest accuracy of 77.92%.

Regarding ROC AUC, all models produced results that were relatively similar (62%-72%). XGBoost ranked in the top (71.64%), while the NB received the lowest with a ROC AUC score of 62.03%. The ROC curves of the models are depicted in Figure 9. XGBoost is closest to the top left of the graph on the ROC curve, indicating that this model works perfectly. Since the ROC curve of NB is so close to the random guessing line, it is not an effective classifier in this study. The ROC curve for RF exhibits the same trend as the ROC curve for XGBoost.

In terms of precision, which refers to the rate of correctly predicting leaving customers out of the total number of customers predicted to be leaving customers, DT reaches a low of 46.29%, XGBoost reaches 67.02%, NB reaches 70.62%, RF 72.48%, SVM reaches 83.82% and KNN reaches the highest at 86.13%.

Regarding Recall, which indicates the ratio of correctly predicting leaving customers out of the total number of customers actually leaving, the effectiveness of the KNN, SVM, and NB had significantly decreased to scores of 29.27%, 39.69%, and 26.92%, respectively. The remaining models predict 45% or higher (DT – 52.65%, RF – 46.56%, XGBoost - 49.51%).

Recall and Precision should be as high as possible, but if we adjust the model to make Recall too high, it could result in a decrease in Precision, and vice versa. F1-Score is used to balance the Precision and Recall ratio because it can harmonic average these two ratios. XGBoost has the highest rate of F1-Score (56.95%). It is possible to ask for the best model to predict customer churn in this data set when the authors apply the algorithm.

Table 4. Effective criteria of models.

Algorithm	**Accuracy**	**Roc_Auc**	**F1_Score**	**Recall**	**Precision**
K-Nearest Neighbors	84.64%	64.03%	43.70%	29.27%	86.13%
Decision Tree	77.92%	68.52%	49.26%	52.65%	46.29%
Support Vector Machines	86.16%	68.86%	53.87%	39.69%	83.82%
Naïve Bayes	82.84%	62.03%	38.98%	26.92%	70.62%
Random Forest	85.52%	71.02%	56.70%	46.56%	72.48%
XGBoost	84.76%	71.64%	56.95%	49.51%	67.02%

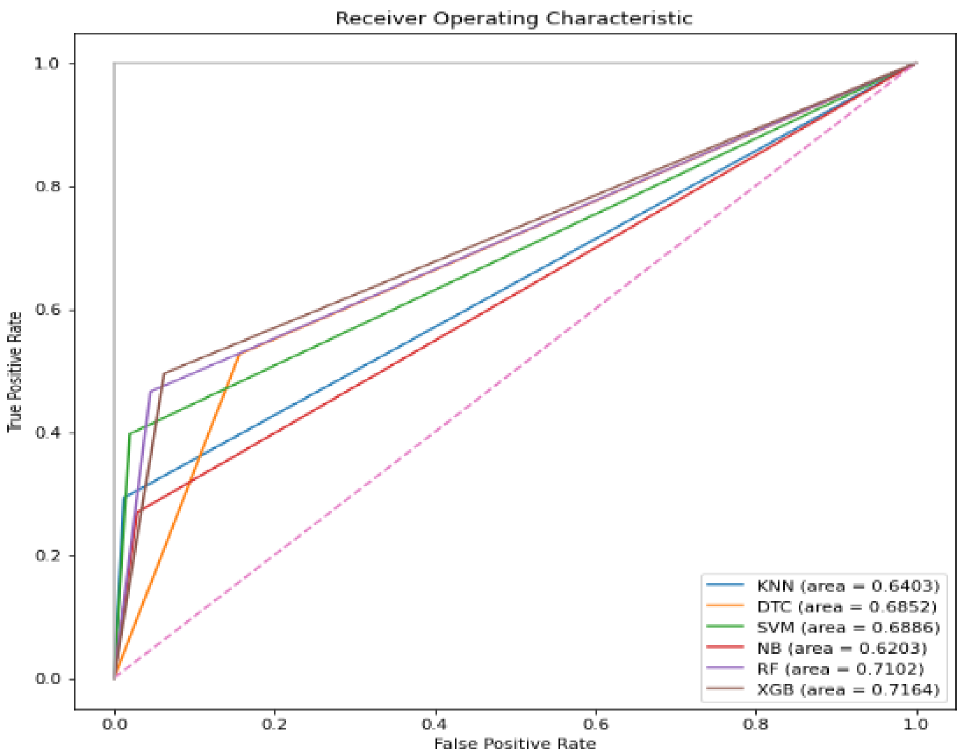

Figure 9. ROC curve.

5. Discussion

The data on bank customer churn for ABC Multistate Bank was examined in this study using six machine learning models. The purpose of this research is to develop the best model for predicting bank customer churn. All of the models are created in Python using the Sklearn library. In terms of Recall value, Precision, ROC AUC, and Accuracy, the performance of KNN, SVM, NB, DT, RF, and XGBoost were compared. XGBoost was found to be significantly more efficient than the other models. The XGBoost model outperformed all others in terms of accuracy in bank customer prediction (84.76%), F1 score (56.95%), recall and precision (49.51%, 67.02%), and an area under the receiver operating characteristic curve (ROC AUC) of 71.64%, respectively. These findings are consistent with the findings of Guliyev and Tatoğlu (2021), who discovered that XGBoost is the best model for predicting customer churn. When an imbalanced dataset occurs, XGBoost is one of the models that can be used. Based on the results, the author makes management recommendations for strengthening the decision-making consideration process at ABC Multistate Bank. The bank may use XGBoost models to accurately identify customers who are at risk of leaving, concentrate their efforts on them, and possibly make a profit.

6. Recommendations & Conclusion

6.1 Recommendations

It is recommended that the resampling technique be used to enhance the performances of the models used in this study. Resampling techniques can be used to make data adjustments to provide balanced training materials for machine learning algorithms because the data used in this study is highly unbalanced in the target variable and the size of the available data sample is small. SMOTE is a common oversampling technique used by researchers. By using the SMOTE, synthetic samples are generated by randomly sampling characteristics from occurrences in the minority class.

Besides, a one-hot encoding technique must be used to change the value of the country and gender columns in order for the machine learning models to be able to read and predict the dataset. It is necessary to convert the data types of both columns from string to integer because the XGBoost model cannot read string data types even though the value in the column is a number.

According to the chart visualizations, younger customers constituted the majority of the retained customer base, while older customers were more likely to leave. Due to the lower likelihood of churn among young customers, ABC Multistate Bank should enhance its marketing approach to attract them. They should also analyze to determine why older customers were more likely to leave, and then adjust their strategy to meet the satisfaction of the older customers. Customers of ABC Multistate Bank tend to be more male, so an investigation is required to determine why the majority of churned customers were women.

Lastly, the bank's churn rate was 20.37%, which is lower than the industry standard of 25% for financial institutions. Therefore, a recommendation is made for the bank to keep bringing in new strategies to lower the churn rate.

6.2 Conclusion

There are plenty of options available to customers today when deciding where to invest their money in the banking industry. To compete in this market, banks must identify customer churn possibilities as soon as possible. Numerous studies have been conducted previously to determine the methods that can be used to predict bank customer churn; however, the purpose of this study was to identify the most appropriate machine learning models to predict customer churn for ABC Multistate Bank.

As part of a study to predict bank customer churn, six supervised machine learning algorithms, which are KNN, SVM, NB, DT, RF and XGBoost, were developed and their effectiveness was investigated. All machine learning algorithms perform well in terms of accuracy score, but since the bank customer churn dataset used in this study is class-imbalanced, it might not be a good idea to use an accuracy score for evaluation. Therefore, the confusion matrix, Precision, Recall, and F1 score, as well as the ROC AUC, have been presented because they can provide greater insight into prediction than accuracy performance measures. After comparing the performance of the models, the author concludes that XGBoost is the best model for this dataset. Results from XGBoost show that the model has an accuracy rate of 84.76%, an ROC AUC of 71.64%, and an F1 score of 56.95%. Another finding is that NB is not capable of handling this dataset adequately because it received the lowest ROC AUC and F1 scores (62.03% and 38.98%, respectively).

However, the efficient churn prediction model only assists companies in identifying the customers who are poised to leave. Effective retention strategies are a crucial component of successful churn management. To meet the needs of customers, bank management must develop appealing retention programs to increase customer loyalty to the bank's services. The proper strategies to retain valuable customers can also be designed with the aid of integrating churn scores with customer segments and applying customer value.

Further research in this area may be required. Future research could concentrate on various machine learning approaches for determining the most accurate models for bank customer churn datasets. Deeper feature selection and extraction techniques should also be studied in order to predict bank customer churn and determine how accurate predictions are affected.

Acknowledgement

Special thanks were extended to the School of Management, Universiti Sains Malaysia.

Funding

This work is supported by the Ministry of Higher Education Malaysia, Fundamental and Research Grant Scheme under Grant No. FRGS/1/2019/STG06/USM/02/6 for the project entitled “A New Hybrid Model for Monitoring the Multivariate Coefficient of Variation in Healthcare Surveillance”.

Author Contribution

All authors have involved themselves equally in every section when writing this paper.

Conflict of Interest

The authors have no conflicts of interest to declare.

References

Asselman, A., Khaldi, M., & Aammou, S. (2021). Enhancing the prediction of student performance based on the machine learning XGBoost algorithm. *Interactive Learning Environments*, 1-20. DOI: https://doi.org/10.1080/10494820.2021.1928235

Berk, R. (2012). *Criminal justice forecasts of risk: A machine learning approach*. Springer Science & Business Media.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

Chandar, M., Laha, A., & Krishna, P. (2006, March). Modeling churn behavior of bank customers using predictive data mining techniques. In *National Conference on Soft computing techniques for engineering applications (SCT-2006)* (pp. 24-26).

Chen, S., Webb, G. I., Liu, L., & Ma, X. (2020). A novel selective naïve Bayes algorithm. *Knowledge-Based Systems*, *192*, 105361.

Cica, D., Sredanovic, B., Tesic, S., & Kramar, D. (2020). Predictive modeling of turning operations under different cooling/lubricating conditions for sustainable manufacturing with machine learning techniques. *Applied Computing and Informatics*. DOI: https://doi.org/10.1016/j.aci.2020.02.001

Costache, R., Arabameri, A., Moayedi, H., Pham, Q. B., Santosh, M., Nguyen, H., ... & Pham, B. T. (2022). Flash-flood potential index estimation using fuzzy logic combined with deep learning neural network, naïve Bayes, XGBoost and classification and regression tree. *Geocarto International, 37*(23), 6780-6807.

Costache, R., Hong, H., & Wang, Y. (2019). Identification of torrential valleys using GIS and a novel hybrid integration of artificial intelligence, machine learning and bivariate statistics. *Catena*, *183*, 104179.

Cunningham, P., & Delany, S. J. (2021). K-nearest neighbour classifiers-a tutorial. *ACM Computing Surveys (CSUR), 54*(6), 1-25.

de Lima Lemos, R. A., Silva, T. C., & Tabak, B. M. (2022). Propension to customer churn in a financial institution: A machine learning approach. *Neural Computing and Applications*, *34*, 11751-11768.

Guan, D., Yuan, W., Lee, Y. K., Najeebullah, K., & Rasel, M. K. (2014). A review of ensemble learning based feature selection. *IETE Technical Review, 31*(3), 190-198.

Guliyev, H., & Tatoğlu, F. Y. (2021). Customer churn analysis in banking sector: Evidence from explainable machine learning models. *Journal Of Applied Microeconometrics*, *1*(2), 85-99.

Gumus, M., & Kiran, M. S. (2017, October). Crude oil price forecasting using XGBoost. In *2017 International conference on computer science and engineering (UBMK)* (pp. 1100-1103). IEEE.

Ha Thanh, N., & Vy, N. (2022). Building a proper churn prediction model for Vietnam's mobile banking service. *International Journal of Advanced and Applied Sciences*, *9*(7), 139-149.

He, B., Shi, Y., Wan, Q., & Zhao, X. (2014). Prediction of customer attrition of commercial banks based on SVM model. *Procedia computer science*, *31*, 423-430.

Hung, S. Y., Yen, D. C., & Wang, H. Y. (2006). Applying data mining to telecom churn management. *Expert Systems with Applications*, *31*(3), 515-524.

Kaur, H., & Kumari, V. (2022). Predictive modelling and analytics for diabetes using a machine learning approach. *Applied Computing and Informatics, 18*, 90-100.

Kaur, I., & Kaur, J. (2020, November). Customer churn analysis and prediction in banking industry using machine learning. In *2020 Sixth International Conference on Parallel, Distributed and Grid Computing (PDGC)* (pp. 434-437). IEEE.

Keramati, A., Ghaneei, H., & Mirmohammadi, S. M. (2016). Developing a prediction model for customer churn from electronic banking services using data mining. *Financial Innovation*, *2*(1), 1-13.

Khodabandehlou, S., & Rahman, M. Z. (2017). Comparison of supervised machine learning techniques for customer churn prediction based on analysis of customer behavior. *Journal of Systems and Information Technology*, *19*, 65-93.

Li, C., Hou, L., Sharma, B. Y., Li, H., Chen, C., Li, Y., ... & Chen, H. (2018). Developing a new intelligent system for the diagnosis of tuberculous pleural effusion. *Computer Methods and Programs in Biomedicine*, *153*, 211-225.

Malhotra, D. K., Malhotra, K., & Malhotra, R. (2020). Evaluating consumer loans using machine learning techniques. In *Applications of Management Science*. Emerald Publishing Limited.

Malik, E. F., Khaw, K. W., & Chew, X. Y. (2022). A new hybrid data preprocessing technique for fraud detection prediction. *Computing and Informatics*, *41*, 981-1001.

Marcelino, P., de Lurdes Antunes, M., Fortunato, E., & Gomes, M. C. (2021). Machine learning approach for pavement performance prediction. *International Journal of Pavement Engineering*, *22*(3), 341-354.

Meza Ramirez, C. A., Greenop, M., Ashton, L., & Rehman, I. U. (2021). Applications of machine learning in spectroscopy. *Applied Spectroscopy Reviews*, *56*(8-10), 733-763.

Oh, G., Song, J., Park, H., & Na, C. (2022). Evaluation of random forest in crime prediction: Comparing three-layered random forest and logistic regression. *Deviant Behavior*, *43*(9), 1036-1049.

Patgiri, R., Varshney, U., Akutota, T., & Kunde, R. (2018, November). An investigation on intrusion detection system using machine learning. In *2018 IEEE Symposium Series on Computational Intelligence (SSCI)* (pp. 1684-1691). IEEE.

Patro, K. K., Jaya Prakash, A., Jayamanmadha Rao, M., & Rajesh Kumar, P. (2022). An efficient optimized feature selection with machine learning approach for ECG biometric recognition. *IETE Journal of Research*, *68*(4), 2743-2754.

Pisal, N. S., Abdul Rahman, S., Hanafiah, M., & Kamarudin, S. I. (2022). Prediction of life expectancy for Asian population using machine learning ALGORITHMS. *Malaysian Journal of Computing (MJoC)*, *7*(2), 1150-1161.

Rahman, M., & Kumar, V. (2020, November). Machine learning based customer churn prediction in banking. In *2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA)* (pp. 1196-1201). IEEE.

Sahin, E. K. (2020). Assessing the predictive capability of ensemble tree methods for landslide susceptibility mapping using XGBoost, gradient boosting machine, and random forest. *SN Applied Sciences*, *2*(7), 1-17.

Shafie, S., Ooi, S. P., & Khaw, K. W. (2023). Prediction of employee promotion using hybrid sampling method with machine learning architecture. *Malaysian Journal of Computing*, *8(1)*, 1264-1286.

Taunk, K., De, S., Verma, S., & Swetapadma, A. (2019, May). A brief review of nearest neighbor algorithm for learning and classification. In *2019 International Conference on Intelligent Computing and Control Systems (ICCS)* (pp. 1255-1260). IEEE.

Tékouabou, S. C., Gherghina, Ș. C., Toulni, H., Mata, P. N., & Martins, J. M. (2022). Towards Explainable Machine Learning for Bank Churn Prediction Using Data Balancing and Ensemble-Based Methods. *Mathematics*, *10*(14), 2379.

Wang, M., & Chen, H. (2020). Chaotic multi-swarm whale optimizer boosted support vector machine for medical diagnosis. *Applied Soft Computing*, *88*, 105946.

Yan, C., Liu, X., Boota, M. W., & Pan, Z. (2022). A Comparison of Machine Learning Methods Applied to the Automated Selection of River Networks. *The Cartographic Journal*, *59*(3), 187-202.

Zhang, H., Zimmerman, J., Nettleton, D., & Nordman, D. J. (2020). Random forest prediction intervals. *The American Statistician*, *74*(4), 392-406.

Zhao, J., & Dang, X. H. (2008, October). Bank customer churn prediction based on support vector machine: Taking a commercial bank's VIP customer churn as the example. In *2008 4th International Conference on Wireless Communications, Networking and Mobile Computing* (pp. 1-4). IEEE.

Zhu, B., Baesens, B., Backiel, A., & Vanden Broucke, S. K. (2018). Benchmarking sampling techniques for imbalance learning in churn prediction. *Journal of the Operational Research Society*, *69*(1), 49-65.