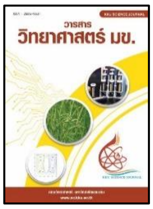
KKU SCIENCE JOURNAL

Journal Home Page : https://ph01.tci-thaijo.org/index.php/KKUSciJ

Published by the Faculty of Science, Khon Kaen University, Thailand



การประเมินวิธีการเรียนรู้ของเครื่องสำหรับแก้ปัญหาความไม่สมดุลของข้อมูลลูกค้าในธนาคาร: การศึกษาเชิงเปรียบเทียบ

Evaluating Machine Learning Methods for Solving Class Imbalance in Banking Customer Data: A Comparative Study

จิรกฤต บุญหมื่นไวย[1] ธีรภัทร จันทรักษา[1] และ เบญจวรรณ โรจนดิษฐ์[1*]

Jirakit Boonmunewai[1], Teerapat Chantaraksa[1] and Benjawan Rodjanadid[1]

[1]สาขาวิชาคณิตศาสตร์และภูมิสารสนเทศ สำนักวิชาวิทยาศาสตร์ มหาวิทยาลัยเทคโนโลยีสุรนารี จังหวัดนครราชสีมา 30000

[1]School of Mathematical Sciences and Geoinformatics, Institute of Science, Suranaree University of Technology, Nakhon Ratchasima, 30000, Thailand

บทคัดย่อ

วัตถุประสงค์ของงานวิจัยนี้ คือ การแก้ปัญหาข้อมูลไม่สมดุลที่ส่งผลต่อการทำนายการยกเลิกการใช้บริการของลูกค้าธนาคาร ในการศึกษานี้ใช้เทคนิคการสุ่มตัวอย่างสองแบบ ได้แก่ เทคนิคการชักตัวอย่างแบบวิธีสังเคราะห์ข้อมูลใหม่ และเทคนิคการชักตัวอย่างลดอย่างสุ่มร่วมกับเทคนิคการเรียนรู้ของเครื่อง 3 เทคนิค ในการสร้างตัวแบบในการทำนาย ได้แก่ เทคนิคต้นไม้ตัดสินใจ เทคนิคจำแนกประเภทแบบนาอีฟเบย์ และเทคนิคซัพพอร์ตเวกเตอร์แมชชีน จากผลการศึกษาพบว่าการใช้เทคนิคการชักตัวอย่างแบบวิธีสังเคราะห์ข้อมูลใหม่ร่วมกับการสร้างตัวแบบด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีนให้ประสิทธิภาพในการทำนายได้ดีที่สุดโดยจะมีประสิทธิภาพเหนือกว่าแบบจำลองอื่น ๆ โดยมีค่าความแม่นยำร้อยละ 92.99 ค่าวัดประสิทธิภาพร้อยละ 91.37 ค่า AUC ร้อยละ 96.4 และอัตราลบเท็จร้อยละ 7.01

ABSTRACT

The goal of this research was to address an imbalance problem that affects churn prediction for bank customers. In this study, we examined two sampling techniques Synthetic Minority Over-sampling Technique (SMOTE) and random under-sampling along with three predictive models: the decision tree classifier, Naïve Bayes classifier, and support vector machine classifier. The results indicated that the support vector machine classifier, when combined with SMOTE, was the most effective, achieving a recall of 92.99%, an F-score of 91.37%, an area under the curve (AUC) of 96.4%, and a false negative rate of 7.01%.

คำสำคัญ: การพยากรณ์การขอยกเลิกใช้บริการ เทคนิคการชักตัวอย่างแบบวิธีสังเคราะห์ข้อมูลใหม่ ต้นไม้การตัดสินใจ ตัวแบบนาอีฟเบย์ ซัพพอร์ตเวกเตอร์แมชชีน

Keywords: Churn Prediction, Synthetic Minority Over-sampling Technique, Decision Tree Classifier, Naïve Bayes Classifier, Support Vector Machine

*Corresponding Author, E-mail: benjawan@g.sut.ac.th

Received: date: 1 October 2024 | Revised date: 20 October 2024 | Accepted date: 12 November 2024

INTRODUCTION

Many businesses, including telephone service providers, credit card companies, insurers, and retail groups, face the challenge of retaining their existing customers. Accurately predicting which customers are likely to discontinue or cancel their services, known as churn prediction, is crucial. Churn prediction enables businesses to anticipate changes in customer behavior, allowing them to focus on customers who are at risk of leaving. This proactive approach helps companies create targeted campaigns to retain customers, thereby reducing the risk of service termination.

Customer data plays a pivotal role in churn prediction. It is essential to analyze how existing customer data can be utilized effectively in a prediction model and to assess the performance of the resulting model. Today, artificial intelligence (AI) techniques are increasingly popular and effective for analyzing and developing churn prediction models. Techniques such as decision trees, Naïve Bayes classification, and support vector machines are AI branches commonly used in building predictive models. However, the amount of data used to create a churn prediction model significantly influences the model's performance. Typically, a nearly equal amount of data from each group is required for effective AI training.

Burez and Van den Poel (2009) conducted an empirical study focusing on sampling techniques (random and advanced under-sampling) and more suitable evaluation metrics such as AUC and lift. They explored the performance improvements of sampling and two specific modeling techniques gradient boosting and weighted random forest compared to standard techniques like logistic regression and random forests. Their findings revealed that under-sampling can enhance prediction accuracy, particularly when evaluated using AUC, contrary to the results of Ling and Li (1998). However, the advanced sampling technique CUBE did not improve predictive performance, suggesting that other advanced sampling methods, such as SMOTE for over-sampling, might yield better results.

Amin *et al*. (2016) examined six well-known sampling techniques and compared their performance: mega trend diffusion function (MTDF), synthetic minority oversampling technique (SMOTE), adaptive synthetic sampling approach, majority weighted minority oversampling technique, immune centroids oversampling technique, and couples top n reverse k-nearest neighbor on four publicly available datasets from the telecommunications sector. The empirical results indicated that MTDF and rule-generation based on genetic algorithms outperformed the other oversampling methods and rule generation algorithms in terms of overall predictive performance.

Xie *et al*. (2019) proposed an improved oversampling algorithm based on a sample selection strategy for imbalanced data classification. Building on the Random-SMOTE algorithm, they extracted support vectors and used them as parent samples to generate new examples for the minority class, thereby balancing the data. The imbalanced datasets were then classified using an SVM algorithm, with F-measure, G-mean, ROC curve, and AUC values selected as performance evaluation metrics. Experimental results demonstrated that this improved algorithm provides strong classification performance on imbalanced datasets.

Wadikar (2020) compared several supervised machine learning techniques for creating a churn prediction model, including logistic regression, random forest, support vector machine (SVM), and neural networks. Initially, these models were applied to an imbalanced dataset, and the results were evaluated. Wadikar then used the SMOTE technique to balance the data and re applied the same models, comparing the results. In Wadikar's study, the Random Forest model emerged as the best overall classifier.

Kimura (2022) developed a customer churn prediction model to enhance customer retention, which is more cost-effective than acquiring new customers. Given the complexity of predicting churn due to diverse predictors and their unclear effect sizes, the research utilized advancements in data storage and analytics to apply machine learning techniques. The researchers focused on supervised machine learning algorithms, treating churn prediction as a binary classification problem. Traditional algorithms like logistic regression, K-Nearest Neighbor, and Decision Tree were compared with advanced ensemble learning models such as XGBoost, LightGBM, and CatBoost, which have shown high prediction performance but have been rarely used in churn prediction. To address the issue of imbalanced datasets, where churn cases are fewer than non-churn cases, the study employed hybrid resampling methods like SMOTE-ENN and SMOTE Tomek-Links, which are novel yet underutilized in this context. By integrating ensemble learning algorithms with these hybrid resampling methods, the study aimed to develop a more accurate prediction model. The model's performance was then compared with traditional methods and previous studies, contributing significantly to the research on customer churn prediction.

Srinivasan and Subalalitha (2023) addressed the challenge of class imbalance, a significant issue in sentiment analysis, particularly for code-mixed data where text alternates between two or more languages. The authors highlighted that most existing research focused on monolingual data, thus neglecting the complexities of code-mixed data. They introduced a novel solution that combined sampling techniques with Levenshtein distance metrics for preprocessing to handle both class imbalance and code-mixing issues. They evaluated the performance of various machine learning algorithms, including Random Forest Classifier, Logistic Regression, XGBoost Classifier, Support Vector Machine, and Naïve Bayes Classifier, using the F1-score as a metric for comparison. This comprehensive approach aimed to improve sentiment analysis accuracy in imbalanced code-mixed datasets, addressing a gap in current research.

Haddadi *et al*. (2024) provided a comprehensive analysis of Customer Churn Prediction by examining three public datasets characterized by significant class imbalance. The datasets spanned various business sectors, including telecommunications, online retail, and banking. The researchers conducted a comparative analysis of fourteen distinct classification methods, incorporating popular resampling strategies such as the Synthetic Minority Over-sampling Technique (SMOTE) and Adaptive Synthetic Sampling (ADASYN). Specifically, they investigated a novel configuration that integrated a two-phase resampling method, based on clustering and ensemble techniques, with Long Short-Term Memory (LSTM) networks. Their findings demonstrated competitive effectiveness, highlighting the method's potential for improving prediction accuracy through effective imbalance correction. The results suggested that this integrated approach generally outperformed standalone methods across different scenarios in the three datasets, particularly in terms of the Area Under the Curve (AUC). This research made a significant contribution to the field of churn prediction by addressing the challenges of class imbalance.

In this study, the data used for training the churn prediction model was classified into two groups: retained and closed accounts. A balanced representation of both groups enhances the efficiency of model creation. However, data is often unbalanced. To address this imbalance, several techniques can be employed. In this research, we considered the Synthetic Minority Over-Sampling Technique (SMOTE) and random under-sampling. These techniques help transform unbalanced data into balanced data.

This research focuses on the problem of imbalanced data classification. Various techniques for addressing this issue will be reviewed, with particular emphasis on the class imbalance problem in churn prediction models. An example related to this problem will be presented and compared. RapidMiner was the primary software used in this work for data normalization, re-sampling, balancing, and modeling the churn prediction.

LITERATURE REVIEW

1. Machine Learning

Machine learning is a branch of AI that enables systems to learn and improve autonomously from experience without explicit programming. It involves creating computer programs that can access and learn from data. This learning process starts with observations or data to identify patterns, allowing the system to make better decisions in the future based on the examples provided. The primary goal is to enable computers to learn independently, without human intervention or assistance, and to adjust their actions accordingly (Expert System Team, 2017).

1.1 Decision tree

This method is a popular machine learning algorithm because it is easy to understand and interpret. It is a flowchart-like structure used for classification and regression (Han *et al.*, 2011). The decision structure consists of the topmost node, which is the root node. Each internal node of the tree represents attribute tests, branches denote the outcomes of these tests, and leaf nodes signify the classes.

The algorithm for constructing a decision tree employs information theory principles. The information content to data is measured using entropy, which indicates uncertainty. Entropy $H(x)$ can be calculated as follows:

$$H(x) = -\sum_{i=1}^{n} P(x_i) log_2 P(x_i), \tag{1}$$

where $P(x_i)$ is the probability of even x_i, and there are n possible even. This entropy value is used to calculate the information gain (IG), which helps select the best independent variable X_a from the set of all independent variables X $(X_a \in X)$ in the set of all example S for each node. The IG for X_a is defined as follows:

$$IG(S, X_a) = H(S) - H(S|X_a), \tag{2}$$

where $H(S|X_a) = \sum_{v \in values(X_a)} \left|\frac{S_v}{S}\right| H(S_v)$

Here, v represents the values of the independent variable X_a , $S_v = \{s|s \in S, Value(s, X_a) = v\}$, and $|S|$ is the number of members in the set S. The gain ratio (GR) is then calculated using the following equation:

$$GR(S, X_a) = \frac{IG(S,X_a)}{-\sum_{v \in Values(X_a)} \left|\frac{S_v}{S}\right| log_2 \left|\frac{S_v}{S}\right|}. \tag{3}$$

The independent variable with the highest GR is chosen as the root or decision node.

1.2 Bayes classification methods

Naïve Bayes classifiers are classifiers that can predict the probability of class membership and are based on Bayes' theorem, this theorem provides a way to calculate conditional probability, which is the likelihood of an event occurring given that another event has already happened. The conditional probability of an event B, given that event A has occurred, is denoted as $P(B|A)$. This probability can be calculated using Bayes' theorem as follows:

$$P(B|A) = \frac{P(B \cap A)}{P(A)} = \frac{P(B)P(A|B)}{P(A)} \tag{4}$$

Given the vector $X = x_1, x_2, \dots, x_n$ representing n features assigns probabilities to this instance for each of the K possible classes C_k, $P(C_k|X)$ can be calculated as follows:

$$P(C_k|X) = \frac{P(C_k)P(X|C_k)}{P(X)} \tag{5}$$

For Naïve Bayes classifier, it assumes that each class is independent, leading to the following:

$$P(X|C_i) = \prod_{k=1}^{n} P(x_k|C_i) \tag{6}$$

1.3 Support vector machine (SVM)

SVM is a widely used supervised machine learning algorithm, it is a discriminative classifier characterized by the creation of a separating hyperplane, making it suitable for classifying both linear and nonlinear data. The algorithm relies on a set of mathematical functions known as kernels, which transform input data into the required form. Various SVM algorithms utilize different types of kernel functions, which can vary depending on the specific needs of the data. For instance, some kernels perform a nonlinear mapping to project the original training data into a higher-dimensional space, enabling more complex classifications.

Let $T = \{(x_1, y_1), (x_2, y_2), \dots, (x_l, y_l)\}$ be the training data where $x_i \in R^n$, $y_i \in R$, and $i = 1,2, \dots, l$. The equation of a SVM is

$$f(x) = (w \cdot x) + b, \tag{7}$$

where $f(x)$ is a predicted function, w is a weight vector and b is a bias value and $b \in R$, $(\cdot)$ is an inner product and x is an input data. Solving the regression problem uses an optimized weight parameter as in the following equation

$$\min_{w,b,\xi^*} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}\left(\xi_i + \xi_i^*\right), \tag{8}$$

with constraints

1. $y_i - (w \cdot x_i) - b \le \varepsilon + \xi_i,$

2. $(w \cdot x_i) + b - y_i \le \varepsilon + \xi_i^*,$

where $\xi_i, \xi_i^*, C \ge 0, \forall i = 1, 2, \ldots, l$, ξ_i, ξ_i^* are slack variables, $\varepsilon-$ insensitive function, and C is a penalty parameter.

2. Imbalance Data Technique

Imbalanced data usually refers to a situation in classification problems where the classes are not equally represented. In binary classification, the class with more instances is called the majority class, while the class with fewer instances is called the minority class.

2.1 Synthetic Minority Oversampling Technique (SMOTE)

Oversampling is a sampling technique used to balance a dataset by increasing the number of examples in the minority class. SMOTE is a method used to address class imbalance in datasets, particularly in binary classification problems where one class is significantly underrepresented compared to the other. SMOTE works by generating synthetic samples for the minority class to augment the dataset, rather than simply duplicating existing samples (Ramyachitra and Manikandan, 2014). The procedure for SMOTE includes the following steps: First, Identify Neighbors, for a given point M in the minority class, use the k-nearest neighbors algorithm to find k points that are closest to M within the minority class and then select one of the k-nearest neighbors at random. The selected neighbor, along with the original point M, is used to generate a synthetic point.

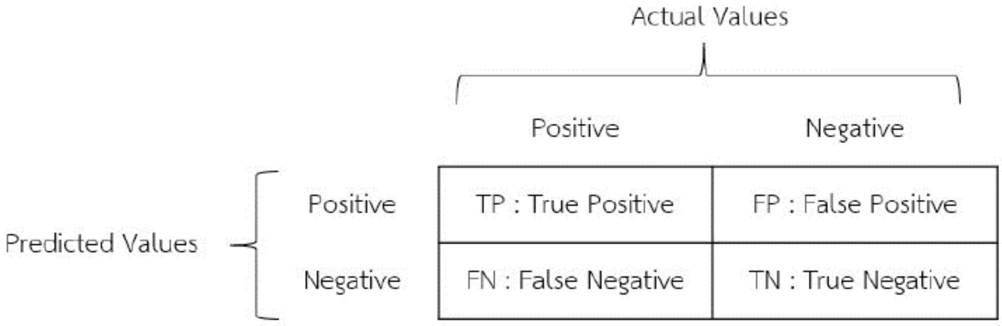
2.2 Random Undersampling

This method involves removing some examples from the majority class. Random Undersampling is a straightforward method where samples from the majority class are randomly eliminated until the dataset is balanced.

3. Model Evaluation

3.1 Confusion matrix

confusion matrix is a performance measurement tool used in machine learning for classification problems where the output can have two or more classes. It is a type of table that provides to know the performance of a classification model on a set of test data with known true values (Xie *et al.*, 2019) as shown in Figure 1.



		Actual Values	
		Positive	Negative
Predicted Values	Positive	TP : True Positive	FP : False Positive
	Negative	FN : False Negative	TN : True Negative

Figure 1 Confusion matrix.

Where True Positive (TP) is the prediction is positive, and the actual result is also positive, False Positive (FP) is the prediction is positive, but the actual result is negative, True Negative (TN) is the prediction is negative, and the actual result is also negative, and False Negative (FN) is the prediction is negative, but the actual result is positive.

Accuracy, Precision, Recall and F-measure can be calculated as following equations,

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{9}$$

$$Precision = \frac{TP}{TP+FP}, \tag{10}$$

$$Recall = \frac{TP}{TP+FN}, \tag{11}$$

$$F - measure = \frac{2(Precision \cdot Recall)}{Precision + Recall}, \tag{12}$$

$$False\ Positive\ rate = \frac{FP}{FP+TN}, \tag{13}$$

$$False\ Negative\ rate = \frac{FN}{FN+TP} \tag{14}$$

3.2 Receiver operation characteristic curve (ROC)- Area under the ROC curve (AUC)

ROC is a probability curve commonly used in binary classification tasks to evaluate classifier output. This curve is plotted with the TP rate on the Y-axis and the FP rate on the X-axis. The Area Under the ROC Curve, called AUC, quantifies the ability of the model to discriminate between classes (Narkhede, 2016). The AUC score ranges from 0 to 1, where a higher value indicates better model performance, with 0.5 or lower suggesting poor performance.

MATERIALS AND METHODS

This section outlines the process used in this research. The empirical part of the study was conducted using RapidMiner Studio version 9.6 (education license).

1. Data Collection

In this research, we utilized data from the Kaggle dataset store. (Shrutimechlearn, 2019).

2. Data Selection

The dataset used in this research was extracted from the Kaggle dataset and contains customer details from a bank, with a binary target variable indicating whether a customer has left the bank (closed their account: churn customer) or remains a customer. The dataset comprises 8,166 instances, with 203 churn and 7,963 non-churn instances, and includes 13 attributes. For this study, we extracted 11 attributes to build the necessary and targeted features and the meaning of each variable is shown in table 1.

3. Data Preparation

3.1 Normalizing data

For continuous variables, they were normalized by using Min-Max normalization to change the values to a common scale. Min-Max normalization operates by measuring how much greater a field value is compared to the minimum value min (X), and then scaling this difference by the range (Larose and Larose, 2014).

$$X^* = \frac{X - \min(X)}{\max(X) - \min(X)} \quad (15)$$

where X is our original field values,

X^* is the normalized field value.

3.2 Encoding

The dataset includes both continuous and categorical variables. However, the SVM algorithm only accepts numeric data, so the categorical variables were converted into 0 and 1 using label encoding. In this dataset, we identified five categorical variables, which were then transformed into numerical format.

Table 1 Variables of bank customer data

Variable	Meaning	Type of variable
Age	Age of the customer	continuous
Balance	Bank balance of the customer	continuous
CreditScore	Credit score of the customer	continuous
EstimatedSalary	Estimated salary of the customer in Dollars	continuous
Gender	Gender of the customer	nominal 0=female(3,522), 1=male(4,644)
Geography	The country to which the customer belongs	nominal 0=France(4,288), 1=Spain(2,108), 2=Germany(1,770)
HasCrCard	Binary Flag for the customer holds a credit card or not	nominal 0=no(2,392), 1=yes(5,774)
IsActiveMember	Binary Flag for the customer is member or not	nominal 0=no(3,686), 1=yes(4,480)
NumOfProducts	Number of bank products the customer is utilising	Discrete 1 (3,819), 2 (4,276), 3 (65), 4(6)
Tenure	Number of years for which the customer has been	Discrete 0 (326), 1 (841), 2 (864), 3 (814), 4 (804), 5 (825), 6 (790), 7 (872), 8 (844), 9 (788), 10 (398)
Exited	Binary flag for the customer closed or retained	nominal 0=customer is retained (7,963) 1=customer closed account (203)

3.3 Data sampling

The dataset consisted of 8,166 instances, with 203 churned customers and 7,963 non-churned customers. This imbalance meant that churned customers represented only 2.48% of the total. To address this class imbalance, we employed random undersampling (No. of churned 203 : No. of non – churned 203) and SMOTE techniques (No. of churned 7963 : No. of non - churned 7963) in our research.

4. Model Selection

After normalizing data, encoding, and sampling the data, we split the final dataset into two proportions: 70% for training 30% for testing, and 80% for training and 20% for testing. Moreover, we enhance the performance of the classification model by using the optimized weight function in RapidMiner Studio to identify important features. Decision Tree, Naïve Bayes, and SVM were utilized to predict the bank customer churn.

5. Model Evaluation

A confusion matrix is a table that helps evaluate the performance of a classification model. In this research, we used precision, AUC, recall, accuracy, false positive rate, false negative rate, and F-score to evaluate the model.

RESULTS

1. Sampling

Figure 2 and 3 show the results of the Exited variable distribution after using sampling

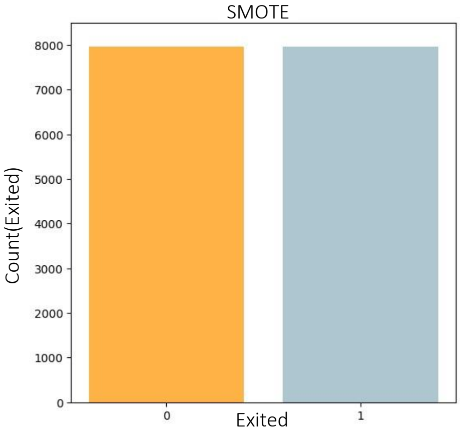


Figure 2 SMOTE.

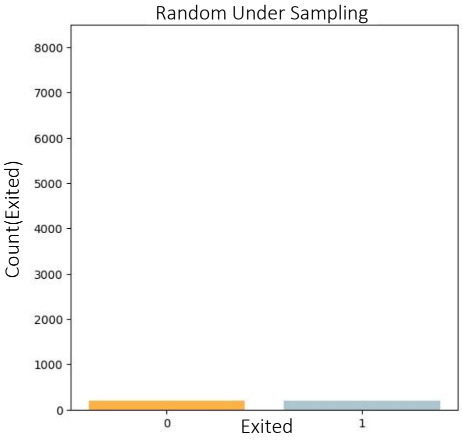


Figure 3 Random under sampling.

2. Feature selection

In this research, we used the optimized weight function in RapidMiner Studio to identify the most important features, enhancing the performance of the classification model.

3. Modeling

3.1 Decision tree

The following parameters were optimized; minimal gain, confidence, criterion, and maximum depth of trees, and the Gain Ratio was selected. Table 2 and Table 3 shows results obtained from the decision tree.

Table 2 Result of Decision tree model for imbalance dataset.

Proportion	No. of Variable	Parameter	ACC	Precision	Recall	AUC	FPr	FNr	F.
70%, 30%	6	MD=6, MG=0 Confi=0.04	97.43	57.14	6.25	0.693	0.13	93.75	11.27
80%, 20%	6	MD=6, MG=0 Confi=0.04	97.43	66.67	4.65	0.555	0.06	95.35	8.63

Table 3 Result of Decision tree model for balanced dataset.

Technique	Proportion	No. of Variable	Parameter	ACC	Precision	Recall	AUC	FPr	FNr	F.
SMOTE	70%,30%	6	MD=21, MG=0 Confi=0.13	83.42	82.92	83.13	0.898	16.30	16.87	83.02
	80%,20%	6	MD=31, MG=0 Confi=0.19	83.33	82.44	83.76	0.902	17.09	16.24	83.09
Random under sampling	70%,30%	5	MD=11, MG=0 Confi=0.07	76.23	73.68	86.15	0.801	35.02	13.85	79.43
	80%,20%	7	MD=31, MG=0 Confi=0.19	74.07	70.45	79.49	0.777	30.95	20.51	74.70

Remark: Confi is confidence, MG is minimun gain and MD is maximum depth. ACC is accuracy, FNr is false negative rate, FPr is false positive rate and F is F-measure, all expressed as percentages (%) and AUC stands for the area under the ROC curve.

3.2 Naïve Bayes

Table 4 and Table 5 shows results obtained from the Naïve Bayes.

Table 4 Result of Naïve Bayes model for imbalance dataset.

Proportion	No. of Variable	ACC	Precision	Recall	AUC	FPr	FNr	F.
70%, 30%	9	97.59	100	7.81	0.843	0	92.19	14.49
80%, 20%	9	97.37	0	0	0.820	0	100	-

Table 5 Result of Naïve Bayes model for balanced dataset.

Technique	Proportion	No. of Variable	ACC	Precision	Recall	AUC	FPr	FNr	F.
SMOTE	70%, 30%	9	77.98	76.40	79.36	0.847	23.33	20.64	77.85
	80%, 20%	9	77.65	76.37	78.63	0.847	23.29	21.37	77.48
Random under sampling	70%, 30%	6	79.51	88.46	70.77	0.866	10.53	29.23	78.63
	80%, 20%	7	77.65	81.58	78.63	0.847	23.29	20.51	80.52

Remark: ACC is accuracy, FNr is false negative rate, FPr is false positive rate and F is F-measure, all expressed as percentages (%) and AUC stands for the area under the ROC curve.

3.3 Support vector machine

In this research, we selected the radial basis function (RBF) kernel for the SVM model due to the significance of the dataset in this study. The following parameters were optimized; epsilon (ξ), kernel gamma (γ), and penalty parameter (C). Table 6 and Table 7 shows results obtained from the Support vector machine.

Table 6 Result of Support vector machine model for imbalance dataset.

Proportion	No. of Variable	Parameter	ACC	Precision	Recall	AUC	FPr	FNr	F.
70%,30%	3	GM=1, C=0.5 Epsilon=0	97.55	100	6.25	0.540	0	93.75	11.76
80%,20%	4	GM=1, C=1.5 Epsilon=0	97.61	100	9.30	0.546	0	90.70	17.02

Table 7 Result of Support vector machine model for balanced dataset.

Technique	Proportion	No. of Variable	Parameter	ACC	Precision	Recall	AUC	FPr	FNr	F.
SMOTE	70%, 30%	9	GM=2, C=1 Epsilon=0	90.87	89.92	91.55	0.961	9.76	8.45	90.73
	80%, 20%	9	GM=2, C=1.5 Epsilon=0	90.99	89.55	92.36	0.964	10.33	7.64	90.93
Random under sampling	70%, 30%	6	GM=1, C=0.5 Epsilon=0	82.79	85.48	81.54	0.840	15.79	18.46	83.46
	80%, 20%	7	GM=1, C=1 Epsilon=0	85.19	84.62	84.62	0.833	14.29	15.38	84.62

Remark: GM is kernel gamma and C is penalty parameter. ACC is accuracy, FNr is false negative rate, FPr is false positive rate and F is F-measure, all expressed as percentages (%) and AUC stands for the area under the ROC curve.

4. Comparative results of all models.

In this comparative section, the values of AUC, precision, recall, FN rate, and F-measure were utilized to evaluate the performance of the models. Table 8 presents the results for each model, with 70% of the dataset used for training and 30% for testing. Table 9 presents the results for each model, with 80% of the dataset used for training and 20% for testing.

Table 8 Evaluation results for training on 70% of the dataset and testing on 30% of the dataset.

Proportion	Model	Precision	Recall	AUC	FNr	F.
70%, 30%	Original DT	57.14	6.25	0.693	93.75	11.27
	Original NB	100	7.81	0.843	92.19	14.49
	Original SVM	100	6.25	0.540	93.75	11.76
	SMOTE DT	82.92	83.13	0.898	16.87	83.02
	SMOTE NB	76.40	79.36	0.847	20.64	77.85
	SMOTE SVM	**89.92**	**91.55**	**0.961**	**8.45**	**90.73**
	RD under DT	73.68	86.15	0.801	13.85	79.43
	RD under NB	88.46	70.77	0.866	29.23	78.63
	RD under SVM	85.48	81.54	0.840	18.46	83.46

Table 9 Evaluation results for training on 80% of the dataset and testing on 20% of the dataset.

Proportion	Model	Precision	Recall	AUC	FNr	F.
80%, 20%	Original DT	66.67	4.65	0.555	95.35	8.63
	Original NB	0	0	0.820	100	-
	Original SVM	100	9.30	0.546	90.70	17.02
	SMOTE DT	82.44	83.76	0.902	16.24	83.09
	SMOTE NB	76.37	78.63	0.847	21.37	77.48
	SMOTE SVM	**89.55**	**92.36**	**0.964**	**7.64**	**90.93**
	RD under DT	70.45	79.49	0.777	20.51	74.70
	RD under NB	81.58	78.63	0.847	20.51	80.52
	RD under SVM	84.62	84.62	0.833	24.14	84.62

Remark: AUC stands for the area under the ROC curve, FNr is false negative rate and F is F-measure. DT is decision tree, NB is naïve Bayes and SVM is support vector machine. RD is random under sampling.

CONCLUSIONS

As the ratio of closed status to retained customers is 203: 7963 and from the results of all models with original data, it shows high accuracy but ignoring the smaller data group results in low recall and AUC and results in a high false negative rate. This demonstrates that data imbalance can affect model validity, Therefore, it is crucial to account for this disparity in modeling. To enhance efficiency, imbalanced data techniques are employed in this research.

The use of SMOTE and random undersampling to address imbalanced data results in improved modeling outcomes as indicated by increased recall, precision, and AUC values, along with a marked reduction in the false negative rate. According to the comparative results, we found that the support vector machine (SVM) technique with a balanced dataset using the SMOTE technique outperformed the other algorithms in predicting customer churn for the bank customers dataset, by the splitting of 70% training data: 30% test data, it provided recall = 91.55%, AUC = 96.1%, F-score = 90.73% and false negative rate = 8.45%. With an 80% training and 20% test data split, the SVM demonstrated even better performance, achieving a recall of 92.36%, an AUC of 96.4%, an F-score of 90.93%, and a false negative rate of 7.64%. This indicates that the model performs better with an 80:20 data split. For the future work, this research may be extended by developing the sampling technique for the other imbalance problem and improving the model for the churn prediction problem.

REFERENCES

Amin, A., Anwar, S., Adnan, A., Nawaz, M., Howard, N., Qadir, J., Hawalah, A. and Hussain, A. (2016). Comparing oversampling techniques to handle the class imbalance problem: A customer churn - prediction case study. IEEE Access 4: 7940 - 7957.

Burez, J. and Van den Poel, D. (2009). Handling class imbalance in customer churn prediction. Expert Systems with Applications 36(3): 4626 - 4636.

Expert System Team. (2017). What is Machine Learning? A definition. Source: http://expertsystem.com/machine-learning-definition/. Retrieved from 16 May 2024.

Haddadi, S.J., Farshidvard, A., Silva, F.D.S., dos Reis J.C. and da Silva Reis, M. (2024). Customer churn prediction in imbalanced datasets with resampling methods: A comparative study. Expert Systems with Applications 246: 123086.

Han, J., Pei, J. and Kamber, M. (2011). Data Mining: Concepts and Techniques. (Third Edition). Waltham, USA.: Elsevier.

Kimura, T. (2022). Customer churn prediction with hybrid resampling and ensemble learning. Journal of Management Information & Decision Science 25(1).

Larose, D.T. and Larose, C.D. (2014). Discovering Knowledge in Data: An Introduction to Data Mining. New Jersey: John Wiley & Sons, Inc.

Ling, C. and Li, C. (1998). Data mining for direct marketing problems and solutions. In: Proceedings of the fourth international conference on knowledge discovery and data mining (KDD-98). NY: AAAI Press, New York.

Narkhede, S. (2016). Understanding AUC - ROC Curve. Source: https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5. Retrieved from 16 May 2024.

Ramyachitra, D. and Manikandan, P. (2014). Imbalanced dataset classification and solutions: a review. International Journal of Computing and Business Research (IJCBR) 5(4).

Shrutimechlearn. (2019). Churn Modelling classification data set. Source: https://www.kaggle.com/datasets/shrutimechlearn/churn-modelling. Retrieved from 24 March 2024.

Srinivasan, R., and Subalalitha, C.N. (2023). Sentimental analysis from imbalanced code-mixed data using machine learning approaches. Distributed and Parallel Databases 41: 37 – 52. doi: 10.1007/s10619-021-07331-4.

Wadikar, D. (2020). Customer Churn Prediction. Masters Dissertation, Technological University Dublin. Dublin, Ireland.

Xie, W., Liang, G., Dong, Z., Tan, B. and Zhang, B. (2019). An Improved Oversampling Algorithm Based on the Samples' Selection Strategy for Classifying Imbalanced Data. Mathematical Problems in Engineering 2019. doi: 10.1155/2019/3526539.

❑❑❑❑❑