Prediction of Long-term Deposit Customers Using SVM Optimized with Borderline-SMOTE

Yunyi Gao

Corresponding author.Email:2911458297@qq.com

Nanjing Agricultural University, Nanjing, China

Abstract. As the economic situation and market environments continually evolve, the financial sector faces numerous challenges, one of which is effectively enhancing the capacity to attract long-term deposits. Long-term deposits are crucial for the stability of a bank's capital and are also a critical factor in enabling banks to offer loans at lower costs. This paper aims to identify existing customers who are highly likely to subscribe to long-term deposits by employing machine learning techniques, thereby helping banks to more accurately target their marketing strategies and improve the efficiency of resource utilization.

Keywords: Economic Situation; Long-term Deposits; Capital Stability; Machine Learning

1 Introduction

Customer relationship management (CRM) is a pivotal business function across various industries. Among its many facets, enhancing customer engagement and loyalty poses significant challenges. Research has consistently demonstrated that acquiring a new customer can be up to five times more costly than retaining an existing one. Given the high expenses associated with customer acquisition, established companies increasingly prioritize customer retention over new customer acquisition. Within the domain of customer retention, accurately predicting customer churn is crucial. Even a marginal improvement in churn prediction accuracy can translate into significant financial savings for a company[1].

Over time, various industries, including banking, telecommunications, online retail, and others, have adopted diverse strategies to predict and prevent customer churn. The intricacies of customer behavior, coupled with the sheer volume of data, present significant challenges in this regard. However, advancements in machine learning and artificial intelligence now enable businesses to harness predictive models for accurate customer churn forecasting[2].This capability allows banks to pinpoint potential long-term depositors with greater precision, thus gaining a competitive edge in the marketplace.

In long-term customer forecasting, data imbalance issues are frequently encountered..There has been significant research attention given to imbalanced classification in the field of machine learning. Imbalanced data is prevalent in various practical applications, including communication fraud detection , medical diagnosis ,and industrial fault detection . Imbalanced data refers to a scenario where the number of majority samples far exceeds that of the minority

MSEA 2024, May 24-26, Jinan, People's Republic of China

DOI 10.4108/eai.24-5-2024.2350156

samples. This severe imbalance adversely affects the learning performance of classifiers. In order to increase the accuracy of the classifier when facing unbalanced data, Chawla et al.[3] proposed the synthetic minority oversampling technique (SMOTE) method. The basic idea of this method is to artificially increase the number of minority samples according to certain rules, so that the ratio of the two types of samples reaches a balance. Compared with the traditional oversampling algorithm, the SMOTE algorithm can effectively alleviate the overfitting problem in the classification process and improve the classification performance of the classifier to a certain extent. H He et al.[4] innovatively proposed an adaptive synthetic sampling method. H Han et al.[5]proposed the Borderline-SMOTE sampling method. The core idea of this method is to oversample the minority samples near the classification boundary, so that the classification boundary between the majority class samples and the minority class samples is clearer[6]. According to the study, SVM, a relatively new algorithm, has required characteristics for the decision family's control[7].Notably, techniques such as SMOTE (Synthetic Minority Over-sampling Technique) and SVM (Support Vector Machines), along with their extensions, offer particular advantages in addressing challenges related to high dimensionality, overfitting, and local optima[8].

This paper adopts a method that combines theory and empirical analysis, using a dataset from the Tianchi website.In this study This paper is organized as follows. In section 2, the methodology of this study is explained including the proposed algorithm. Section 3 delineates the experimental methodology and presents the empirical findings. The concluding section encapsulates the contributions of this paper and delineates avenues for future research.

2 Related Theoretical Analysis

2.1 Data Set Processing and Analysis

The dataset used in this paper was directly downloaded from Tianchi. The first step involved preprocessing the data:

Initially, the data was cleansed by removing duplicate records to ensure each piece of data was unique. This helps prevent the model from overfitting on certain repeated samples. Then, a careful examination of each feature was conducted to identify and address potential outliers. Subsequently, features were standardized, scaling all numeric features to a uniform range. Through standardization, it is ensured that the model considers all features equally, without being dominated by those with a broader range[9].

After data cleaning, the remaining data were processed by Isolation Forest outliers. Isolation Forest was proposed by Liu, Ting, and Zhou (2008)[10], which directly identifies outliers instead of finding outliers by modeling normal data points. It applied a tree structure to isolate each observation. Outliers will be isolated first, while normal points tend to be hidden deep in the tree. They call each Tree an Isolation Tree or iTree. On iTree, observations with short path lengths are considered outliers.

Figure 1 employs partition diagrams and trees to elucidate the process by which iTree isolates data points. The partitioning begins with the most distant point, represented by a orange dot, which is isolated with a single cut. Subsequent cuts are made for the green and blue dots in order of their distances. The number of cuts required to isolate a point correlates directly with its depth in the tree, which inversely determines the anomaly score. Similarly, Figure 2's tree structure demonstrates that isolating the red dot requires one split, the green dot two, and the blue dot three, suggesting that the tree depth serves as an effective measure for outlier scores. By convention, higher outlier scores indicate greater deviation, thus scores are defined as the inverse of tree depth.

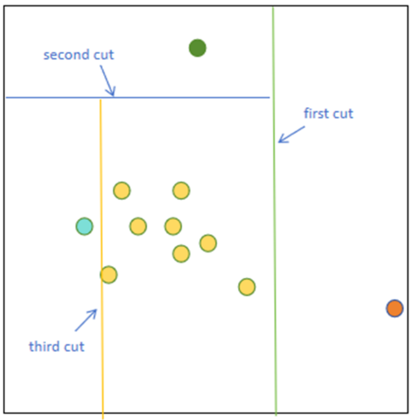

Fig. 1. iTree Isolation Partition Diagram

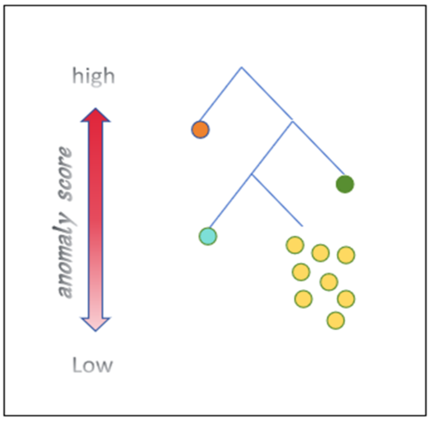

Fig. 2. iTree Isolation Partition Tree

Conclusion can be reached by svm after isolation forest treatment and Borderline Smote treatment of unbalanced data set.

2.2 Theoretical Foundation of Support Vector Machine

2.2.1 Concept of Support Vector Machine Algorithm.

1990′s SVM theory had been developed by Vapnik and his group. SVM concept was inherited from neural network or may be saying that SVM is mathematical extension of Neural Network[11]. It demonstrates many unique advantages in solving problems with small samples, non-linearity, and high-dimensional pattern recognition. It can be extended to other machine-learning problems, such as function fitting. The essence of SVM is to map vectors into a higher-dimensional space, where a maximum-margin hyperplane is constructed. Two parallel hyperplanes are established on both sides of the hyperplane to separate the data. The orientation of the separating hyperplane is chosen to maximize the distance between these two parallel hyperplanes.

2.2.2 Implementation Steps of Support Vector Machine Algorithm

SVM aims to determine an appropriately directed separating hyperplane so that the gap between the closest data points from different categories (i.e., the support vectors) to this hyperplane is maximized as much as possible.

The geometric margins of the hyperplane are shown in equation(1):

$$\gamma_i = \frac{y_i(\varpi \cdot x_i + b)}{\|\varpi\|} \tag{1}$$

Determining the hyperplane that maximises the geometric margins for optimal classification involves an optimisation problem, equation(2). The goal of the optimisation is to maximise the spacing between data points of different categories, thus ensuring the most efficient classification:

$$\begin{cases} \max\limits_{\varpi,\mathrm{b}} \gamma \\ s.t. \frac{y_i(\varpi \cdot x_i + b)}{\|\varpi\|} \geq \gamma \end{cases} \tag{2}$$

2.2.3 Advantages and Disadvantages of the Support Vector Machine Algorithm

SVM has demonstrated high accuracy in many practical problems, especially when the dimensionality of the data is higher than the number of samples. By appropriately choosing the kernel function and adjusting regularization parameters, SVM can effectively deal with the overfitting problem and has good generalization ability. It can efficiently handle high-dimensional data, working well even when the number of features exceeds the number of samples.

However, the SVM classifier has shortcomings in dealing with imbalanced datasets, where minority classes are generally more important in the sample data. Although the SVM model has high classification accuracy and performs well on small sample datasets, in many real-world problems, achieving a higher global classification accuracy by sacrificing the classification accuracy of minority class samples does not meet the classification needs of practical issue[12].

2.3 Theoretical Basis of SMOTE

2.3.1 Concept of SMOTE Algorithm

SMOTE, an acronym for Synthetic Minority Over-sampling Technique, represents a methodology aimed at addressing the issue of class imbalance in datasets through the generation of synthetic samples. Oversampling techniques enhance the representation of minority class instances by generating new samples through interpolation among existing ones. This method effectively increases the count of minority class samples, promoting a more balanced dataset. This method addresses the overfitting problem that may arise from duplicating minority class samples, a common issue with traditional oversampling methods.

2.3.2 The significance of using SMOTE

Machine learning models not only struggle with the issue of overfitting due to insufficient feature analysis but also face the challenge of biased predictions. Such bias commonly originates from data imbalance, which can occur either inherently or as a result of external factors. Intrinsically, many real-world datasets are characterized by significant class imbalances, where one class significantly outnumbers the other. For example, in the context of predicting diabetes, it is often the case that the diabetic class represents the minority, as the actual prevalence of diabetes in the population is typically lower compared to the number of individuals who are diabetes-free. This disparity in class representation can lead to biased model outcomes, affecting the model's ability to generalize accurately to the entire population[13].

2.3.3 Extensions of the SMOTE Algorithm

Borderline-SMOTE is a variant of SMOTE specifically designed to strengthen the minority class samples located at the borders of the dataset. In imbalanced datasets, border samples are the hardest to classify because they are close to or mixed with the majority class area. Borderline-SMOTE focuses on these border minority class samples and oversamples them, with the goal of enhancing the model's ability to recognize boundary areas.

Borderline-SMOTE is available in two versions, BSO1 and BSO2. BSO1 adopts the original SMOTE's concept, generating new samples solely between two adjacent minority class samples using similar class neighbor information from the primary sample. In contrast, BSO2 expands this approach by utilizing neighborhood information from the entire training set, identifying K neighbors among both majority and minority class samples. To avoid placing new samples too close to the decision regions of the majority class, the algorithm limits the random number to the range [0, 0.5], ensuring that new samples are closer to the primary sample rather than the nearest neighbor. This method effectively minimizes the introduction of noise and enhances the quality of the synthetic samples, so This article employs the BSO2 method.The pseudo-code for the Borderline-SMOTE2 method is shown in Table 1:

Table 1. Pseudocode for the BSO2 method

```
BSO2 method
Input:   S={(x_i, y_i)},i=1 to N, y_i in {+, -},N^-, N^+, IR = N^- / N^+, SR, K
Output:   S' = {(x_i, y_i)}
for x in S^+:
    N_near = KNN(x, S, K)
    N_major = majority_class(N_near, y)
      if K/2 <= N_major < K:
          DANGER.add(x)
for i in range(N^+ * SR):
    x = random(DANGER)
    N_near = KNN(x, S, K)
    x' = random(N_near)
    rand = random(0, 0.5)
    x_new = x + rand * (x' - x)
    S^New.add(x_new)
```

In summary, Borderline-SMOTE, with its targeted oversampling strategy, not only improves the classification performance of imbalanced datasets but also reduces the risk of overfitting and enhances the model's generalization ability. It is an effective strategy for dealing with imbalanced data.

3 Research on SVM and Its Improved Models in the Prediction of

Long-term Deposit Customers

3.1 Data Preprocessing and Analysis

The dataset used in this study was obtained from the Tianchi platform, comprising 32,950 entries. Predictions were made on whether a customer would make a long-term deposit (y) based on 15 attributes. The attributes and their meanings are presented in Table 2 as follows:

Table 2. Dataset Variables and Descriptions

Variable Name	Explanation	Description
Age	An individual's age	
Occupation	Job category	"Administrative","Manual Labor", "Business Owner", "Domestic Worker", "Executive", "Pensioner", "Independent Business", "Service Industry", "Scholar", "Technical Staff", "Jobless", "Undefined"
Material	Marital status	"Divorced", "Married", "Single", "Unknown"; Note: "Divorced" refers to divorced or widowed
Education	Educational Attainment	"Basic 4y", "Basic 6y", "Basic 9y", "High school", "Illiterate", "Professional course", "University degree", "Unknown"
Default	History of Credit Default	"No", "Yes", "Unknown"
Housing	Residential Loan	
Loan	Individual Loan	
Contact	Communication Method	
Month	Month of Last Contact	
Dayofweek	Day of the Week of Last Contact	
Duration	Duration of Last Contact	This attribute greatly affects the output target
Campaign	Total interactions with this client in the current campaign	
Days since prior contact	Days elapsed since the last outreach to this client from a prior campaign	A value of 999 indicates that there was no prior contact with the client.
Previous contact count	Contact count with this client before the current campaign	
Prior campaign outcome	Results from the last marketing effort	"Unsuccessful", "Not Applicable", "Achieved"
y	Final result	

This paper conducts a comprehensive analysis of each attribute within the dataset, presenting the correlation coefficients of each attribute with the outcomes, as depicted in Figure 3. These correlation coefficients serve to elucidate the linear relationships between the features. By meticulously analyzing these coefficients, one can discern whether positive, negative, or no correlations exist between the features, thereby deepening their understanding and insight into the dataset.

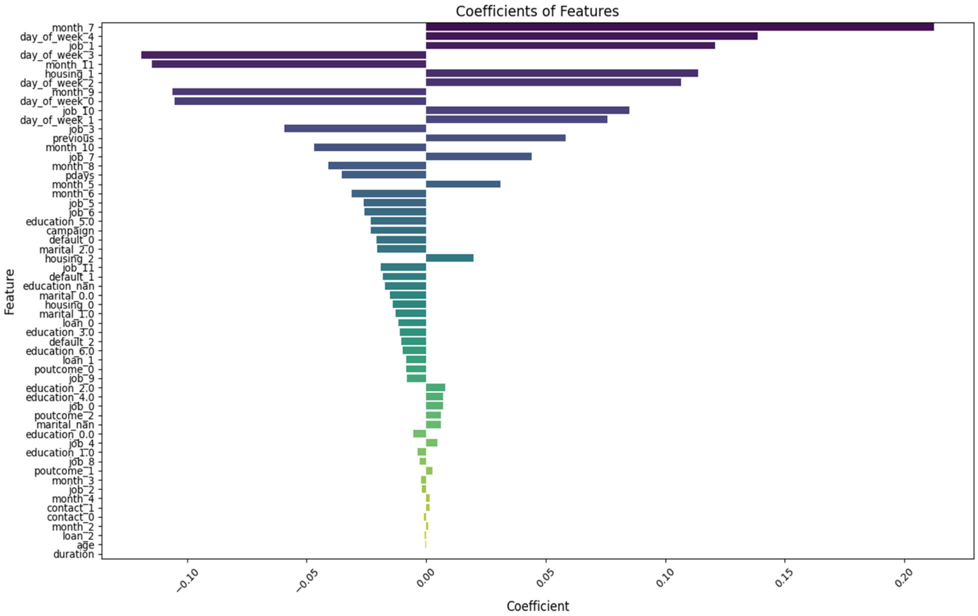

Fig.3. Coefficients of Features

The dataset was divided into a training set and a test set with a ratio of 7:3.Given the total data volume of 32,950 entries, which is relatively large, outlier detection was conducted before training the data. This study employed the isolation-forest outlier detection method, which performs excellently with high-dimensional and large datasets; in datasets with highly imbalanced labels, such as when there is a significant difference between normal and abnormal data, the Isolation Forest shows good performance.

After outlier preprocessing, there was an improvement in the dataset's accuracy, as shown in Table 3:

Table 3. Accuracy Comparison

Processing Method	Accuracy
SVM Prediction	90.081%
SVM Prediction after Isolation-forest Preprocessing	92.297%

Subsequent analysis is conducted on data post-outlier treatment via Isolation Forest. As indicated in Figure 4, there is a significant imbalance between the positive and negative instances in the dataset, with the ratio of long-term deposit customers to non-long-term deposit customers being approximately 7.8:1. To achieve a balanced state, the Borderline-SMOTE oversampling algorithm was introduced for processing.

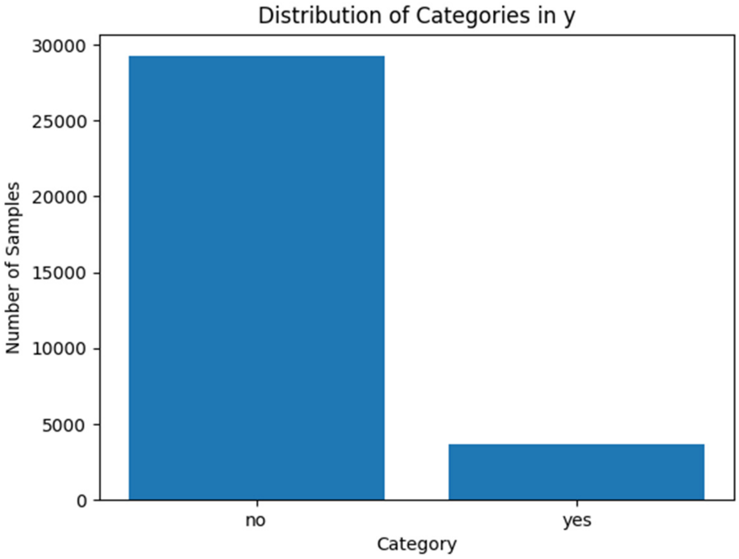

Fig. 4. Distribution of Positive and Negative Instances

3.2 Model Evaluation Metrics

3.2.1 Classification Accuracy

Classification accuracy, also known as accuracy, is a metric used to measure the performance of a model. It represents the proportion of samples correctly predicted by the model out of the total number of samples. It can be calculated according to equations(3).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{3}$$

3.2.2 ROC and AUC Metrics

The ROC curve serves as an analytical tool to depict the efficacy of a classification model across all conceivable decision thresholds. This curve plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at varying thresholds, facilitating an evaluation of the model's discriminative capacity.

True Positive Rate (TPR): Commonly referred to as sensitivity, recall, or hit rate, the TPR quantifies the model’s proficiency in accurately identifying positive instances.

False Positive Rate (FPR): This metric measures the fraction of negative instances erroneously classified as positive, thus reflecting the occurrence of false alarms.

Additionally, the Area Under the Curve (AUC) metric, which spans from 0 to 1, encapsulates the overall classification performance of the model. A higher AUC value indicates a more effective classifier.The relationship between AUC value and model classification ability is shown in Table 4:

Table 4. Relationship between AUC Value and Model's Classification Ability

AUC Value	Interpretation
AUC=1	Indicates perfect classification ability of the model to distinguish between positive and negative instances.
0.5< AUC < 1	"The proximity of the Area Under the Curve (AUC) to 1 serves as an indicator of superior model performance. An AUC approaching 1 denotes a robust classification capability, distinguishing effectively between the classes under consideration."
AUC = 0.5	Indicates the classification ability of the model is equivalent to random guessing.
AUC < 0.5	Indicates the model's classification performance is worse than random guessing.

3.2.3 KS Metric

The KS (Kolmogorov-Smirnov) statistic measures the maximum difference between the cumulative distribution functions (CDF) of the predicted probabilities of the positive and negative classes. Specifically, it calculates the maximum gap between the CDFs of the positive and negative classes across all possible thresholds.

$$\mathrm{KS} = \max(\mathrm{TPR} - \mathrm{FPR}) \tag{4}$$

Different ranges of KS metric values categorize model performance as follows, shown in Table 5[14]:

Table 5. Relationship Between KS Metric and Model Performance

KS Value Range	Model Performance Classification
KS<0.2	Model lacks discriminative ability
0.2⩽KS<0.3	Model has weak discriminative ability
0.3⩽KS<0.5	Model has strong discriminative ability
0.5⩽KS<0.75	Model has very strong discriminative ability
0.75⩽KS	Model may have anomalies

3.2.4 Weighted Avg Metric.

In classification tasks, particularly with imbalanced datasets, a range of performance metrics are employed. The term "weighted average" refers to a holistic metric that factors in the support, which is the number of samples for each class, and appropriately weights performance metrics such as precision, recall, and F1 score to provide a balanced assessment.

In imbalanced datasets, some classes may have significantly more samples than others. Suppose the average of all class metrics is calculated. In that case, the performance of minority classes will contribute less to the overall performance, possibly failing to reflect the model's ability to handle minority classes accurately. Using weighted averages ensures that the model's overall performance across all classes is evaluated more fairly, matching each class's contribution to its representation in the dataset.

3.3 Improvement Process of SVM Model with Borderline-SMOTE

3.3.1 Parameter Tuning Process.

① By traversing all data in the training subset and dividing the training samples according to the majority and minority class criteria, "yes" is marked as the minority class, and "no" as the majority class. This step results in a minority class sample size of 3638 and a majority class sample size of 29013.

② Taking into account accuracy and classification performance, the parameter 'desired_samples' is set at 3000. Thus, after processing with Borderline-SMOTE, the minority class samples increased to 6638.

3.3.2 Comparative Analysis of Metrics.

After oversampling with the Borderline-SMOTE2 algorithm and integrating with a single algorithm, the performance of the combined model surpasses that of the original single model. The performance comparison is illustrated in Figure 5, Figure 6, and Table 6.

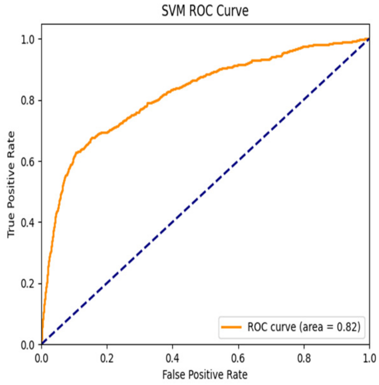

Fig. 5. ROC Curve of the SVM Model

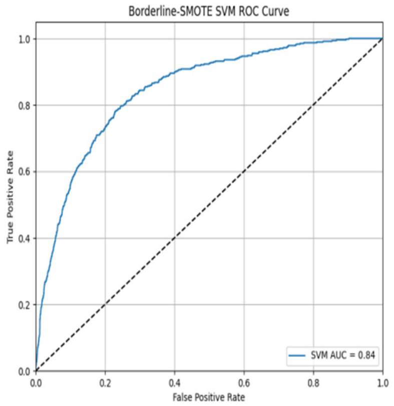

Fig. 6. ROC Curve of the SVM Model after Borderline-SMOTE Data Processing

Table 6. Comparison of Metrics Before and After Applying Borderline-SMOTE

Model Type Evaluation Metric	Accuracy	KS	AUC	Recall (weight avg)	f1-score (weight avg)	Precision (weight avg)
Borderline-SMOTE2 SVM Model	0.92390	0.552	0.8445	0.92	0.92	0.91
SVM Model	0.92297	0.523	0.8206	0.89	0.92	0.90

4 Conclusions

The study focused on developing and implementing a long-term deposit customer identification prediction system based on Support Vector Machine (SVM) and the Borderline-SMOTE technique. By leveraging SVM's strong classification capability and Borderline-SMOTE's oversampling strategy for imbalanced datasets, the paper successfully enhanced the predictive performance of long-term deposit customer identification. Specifically, the implementation of the Borderline-SMOTE algorithm, tailored for targeted oversampling of minority class samples, effectively addressed the imbalance inherent in the original dataset, providing a more balanced and information-rich training foundation for the SVM model.

Assessing the reliability of classifier algorithms is crucial to ensuring data quality. This study utilized comprehensive metrics such as the area under the ROC curve, KS index, and others to yield meaningful results[15].Experimental findings demonstrated that the SVM model coupled with Borderline-SMOTE technology significantly improved key performance metrics such as accuracy and AUC compared to the traditional SVM model. This enhancement strengthened the model's capability to identify potential long-term deposit customers, offering banks and financial institutions more precise and reliable decision-support tools.

These results underscore the importance of addressing data imbalance issues in predictive modeling tasks, particularly in the banking and finance sectors where accurate customer identification is essential for business success. The demonstrated effectiveness of the SVM-Borderline-SMOTE approach highlights its potential for wider adoption in similar applications, suggesting avenues for future research to explore its applicability in other domains and its potential integration with emerging technologies. Additionally, the study emphasizes the significance of robust evaluation methodologies in assessing the performance of predictive models, serving as a guideline for future studies in this field.

References

[1] Vo, N. N., Liu, S., Li, X., & Xu, G. (2021). Leveraging unstructured call log data for customer churn prediction. Knowledge-Based Systems, 212, 106586. https://doi.org/10.1016/j.knosys.2020.106586

[2] Haddadi, S. J., Farshidvard, A., dos Santos Silva, F., dos Reis, J. C., & da Silva Reis, M. (2024). Customer churn prediction in imbalanced datasets with resampling methods: A comparative study. Expert Systems with Applications, 246, 123086. https://doi.org/10.1016/j.eswa.2023.123086

[3] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321 – 357. https://doi.org/10.1016/j.eswa.2023.123086

[4] Haibo He, Yang Bai, E. A. Garcia and Shutao Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), Hong Kong, 2008, pp. 1322-1328, doi: 10.1109/IJCNN.2008.4633969.

[5] Han, H., Wang, WY., Mao, BH. (2005). Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning. In: Huang, DS., Zhang, XP., Huang, GB. (eds) Advances in Intelli-

gent Computing. ICIC 2005. Lecture Notes in Computer Science, vol 3644. Springer, Berlin, Heidelberg. https://doi.org/10.1007/11538059_91

[6] Han, M., Wu, Y., Huang, Y., & Wang, Y. (2021). A fault diagnosis method based on improved synthetic minority oversampling technique and svm for unbalanced data. In IOP Conference Series: Materials Science and Engineering (Vol.1043, No.5, p.052034). IOP Publishing. DOI:10.1088/1757-899X/1043/5/052034

[7]Kurani, A., Doshi, P., Vakharia, A., & Shah, M. (2023). A comprehensive comparative study of artificial neural network (ANN) and support vector machines (SVM) on stock forecasting. Annals of Data Science, 10(1), 183-208.https://doi.org/10.1007/s40745-021-00344-x

[8]Guo, J., Wu, H., Chen, X., & Lin, W. (2024). Adaptive SV-Borderline SMOTE-SVM algorithm for imbalanced data classification. Applied Soft Computing, 150, 110986.https://doi.org/10.1016/j.asoc.2023.110986

[9]Peng, W. (2019). Overdue Prediction of Microfinance Users Based on Combined Random Forest-Logistic Regression Model Master (Dissertation, Chongqing University). Master https://kns.cnki.net/kcms2/article/abstract?v=CNKoHtoL3RHYkCbuTvwsYvygNGYkS7UJ4wBCebTNeTXNvrIzB7cvV96kOSGcO2sjze8EmJv-OrBD4AxHgkradCnZBD2tXt_C--gkKSbEEh-F6ALV642UZeIiCtzFjZoAmYtOHoPai8RQmHv4OHuFRw==&uniplatform=NZKPT&language=CHS

[10] Liu, F. T., Ting, K. M., & Zhou, Z. H. (2008, December). Isolation forest. In 2008 eighth ieee international conference on data mining (pp. 413-422). doi: 10.1109/ICDM.2008.17.

[11] Chandra, M.A., Bedi, S.S. Survey on SVM and their application in image classification. Int. j. inf. tecnol. 13, 1–11 (2021). https://doi.org/10.1007/s41870-017-0080-1

[12] Wan Yibin. Research on employee departure prediction based on SMOTE-SVM under unbalanced data[D]. Donghua University, 2023.DOI:10.27012/d.cnki.gdhuu.2022.000479.

[13] Alex, S. A., Nayahi, J. J. V., & Kaddoura, S. (2024). Deep convolutional neural networks with genetic algorithm-based synthetic minority over-sampling technique for improved imbalanced data classification. Applied Soft Computing, 156, 111491. https://doi.org/10.1016/j.asoc.2024.111491

[14] Ji Peiling. Research on credit assessment of bank personal credit loans based on SMOTE-Logistic regression algorithm [D]. Hui University of Technology, 2021. doi:10.27790/d.cnki.gahgy.2021.000144.

[15] Drosou, K., Georgiou, S., Koukouvinos, C., & Stylianou, S. (2014). Support Vector Machines Classification on Class Imbalanced Data: A Case Study with Real Medical Data. Journal of Data Science, 12(4), 727-753. https://doi.org/10.6339/JDS.201410_12(4).0008