A Prediction Method for Bank Customers' Transaction Reactions Based on Machine Learning

Zhang Dong

***Abstract*—Advertising is important for banks because it can be used to push more term deposit offers to existing customers and help banks to recommend packages or other deals to new customers in order to increase bank transactions. Bank advertisements are usually pushed through media pages such as the Internet or by phone, and in order to find more suitable business packages for their customers, many banks or telecommunication companies store data about their customers and analyze customer information to improve the success rate of ad push. The dataset used in this paper is the customer data stored by a Portuguese bank. With the rapid development of machine learning in recent years, artificial intelligence applied to data analysis is producing better and better prediction results. In this study, based on the above dataset, a predictive model is created using machine learning algorithms to classify the customer's response to the offers provided by the bank, so that the direction of the bank's subsequent advertising push can be improved. This classification is binary, i.e., it predicts whether the customer will accept these offers or not. Four classifiers including Support vector machine (SVM) algorithm, random forest algorithm, plain Bayes and extreme gradient boosted decision tree (XGBoost) were used in this paper, where the random forest classifier obtained the best results with an accuracy of 89.7**

***Index Terms*—heterogeneous defect prediction; variational autoencoders; feature representation**

I. Introduction

Bank advertisements mainly consist of advertisements from financial institutions. Besides targeting bank customers, such advertisements also include relevant business reports and information brochures. Moreover, statements regarding new share payments, reports on the results of investment plans, and some additional financial announcements may also be included [1]. Many banks or telecommunication companies store their customers' data to build relationships with them and offer preferential treatment. Meanwhile, they analyze this data to reasonably recommend packages to customers and retain customer loyalty. Therefore, a rational and effective analysis of the data to extract useful information can assist banks in more accurately pushing advertisements to the target audience, reducing the push to non-target groups, thus achieving the goals of cost reduction and sales increase. With the growth of the number of Internet users and enterprises, many e-commerce application clusters do not seem to be physically interconnected in the system but are interrelated in business. Online banking has also gradually emerged [2], which makes data easier to collect. However, in the Internet era, data is vast and complex, and traditional analysis methods have become inadequate, requiring more advanced data analysis methods for processing. Thus, the rise of machine learning in recent years has provided a good opportunity. Due to its ability to handle large amounts of data and achieve good prediction results, many commercial financial transactions have started to introduce it for prediction [3]. Most research on commercial financial transactions uses the logistic regression model and the SVM model for short-term prediction [4], and some also use the neural network model for prediction. With the development of machine learning and deep learning, more and more excellent methods and models have been applied to this field. In this process, many classic commercial financial transaction datasets have also been generated, such as the Boston Housing Price dataset, the Credit Card dataset, and the dataset used in this paper. This dataset can be downloaded from the UCI Machine Learning Repository [5]. It is composed of customer information stored by a Portuguese bank to increase sales and is associated with its direct marketing campaigns [6]. The focus of these campaigns was telephone calls, recording multiple calls to the same customer to ask whether they subscribed to their "term deposit" product. There are 41,188 rows in all examples of the dataset. Each example includes 16 input variables and one output variable (subscribed or not), and they are sorted by the recording date (from May 2008 to November 2010). The 16 input variables contain various customer details, including age, job, marital status, education, default, housing, loan, contact method, month, day, as well as some information about the last contact, such as duration, number of contacts, contact interval, and the success or failure of the last marketing campaign. However, the dataset is generally unbalanced [7], which may affect model prediction. Therefore, the aim of this study is to create a prediction model using machine learning algorithms based on this dataset to predict how target customers will respond to the preferential activities offered by the bank. This classification prediction is binary, that is, to predict whether customers will participate in these offers. The analysis results can better help banks determine the marketing target population. In this study, four classifiers are used: SVM [8], Random Forest [9], Naive Bayes [10], and XGBoost [11]. After comparison, the Random Forest has the best prediction performance in terms of evaluation metrics.

II. Prediction Methods

A. Data Preprocessing

First, the data needs to be processed in four steps. The first step is data cleaning; the second step is to convert categorical features into numerical features; the third step is to balance the data; and the fourth step is to select the best features and normalize them. Regarding the first step, none of the 45211 rows of data have missing values, but there is redundant

information. In this paper, the ”month” and ”day” columns of the recording time in the input data are removed. In the second step, since the remaining 15 input variables contain string data, which are finite discrete data, they are correspondingly converted into numerical representations. Examples of the data after cleaning and conversion in the first two steps are shown in Fig. 1.

	age	job	marital	education	default	balance	housing	loan	contact	duration	campaign	pdays	previous	poutcome
0	58	3	2	3	0	2143	1	0	0	261	1	-1	0	0
1	44	10	1	2	0	29	1	0	0	151	1	-1	0	0
2	33	5	2	2	0	2	1	1	0	76	1	-1	0	0
3	47	7	2	0	0	1506	1	0	0	92	1	-1	0	0
4	33	0	1	0	0	1	0	0	0	198	1	-1	0	0

Fig. 1: Partial data display after data cleaning and conversion

B. Dataset Balancing

When there are too many unbalanced data points in a dataset, some classification algorithms, such as Naive Bayes, may perform poorly during use, or it may lead to overfitting problems. Therefore, this paper examines the distribution of the input variables. The ideal variable distribution state is a symmetrical distribution rather than a skewed one. For variables with fewer selections, their data distribution should be balanced, that is, the number of selections for each category should be similar. However, as shown in Fig. 2, none of the input variables has a normal symmetrical distribution or a balanced quantity distribution. Moreover, for the output variable, that is, the predicted variable, as shown in Fig.3, the degree of distribution imbalance is even more severe. This indicates that the overall data is unbalanced. Thus, when the dataset is applied to the classifier, high precision will not be achieved. To address this issue, this paper will utilize two techniques to balance the dataset: random undersampling and the Synthetic Minority Over-sampling Technique (SMOTE).

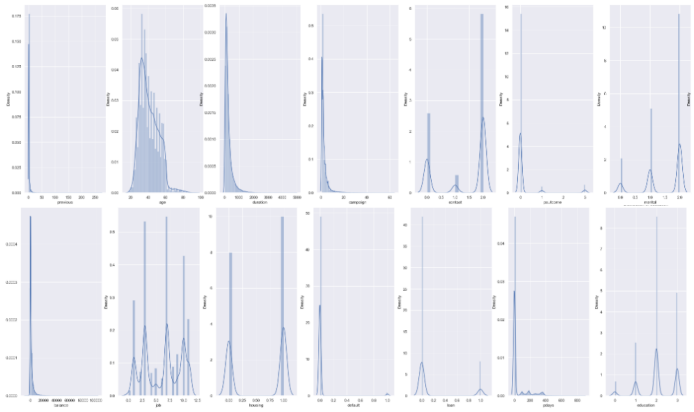


Fig. 2: Distribution Map of Input Variables

1) Random Undersampling Method: This method directly conducts ’undersampling’ on the majority samples in the training set. That is, some samples from the majority class

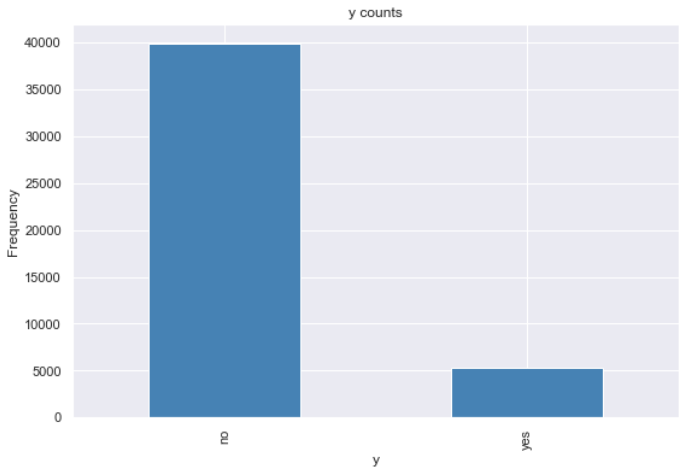


Fig. 3: Distribution Map of Predicted (Output) Variables

are removed to make the number of positive and negative examples close until the two classes are equal [12]. Therefore, this method modifies the sample distribution by changing the proportion of majority-class samples, to make the sample distribution more uniform. This method is simple and fast and is suitable for large data sets. However, there are also some problems. For random sampling, since the sampled sample set is smaller than the original one, it will cause data loss, preventing the model from obtaining a better prediction distribution.

2) Synthetic Minority Oversampling Technique (SMOTE): The Synthetic Minority Over-sampling Technique (SMOTE) attempts to select a point from the minority class and then uses the k-NN method to generate a new point as a point of the minority class [13]. Repetition of this process can achieve the goal of balancing the data set. This is a relatively reasonable strategy because, in terms of the feature space, the new samples are highly similar to the existing instances of the minority class, and at the same time, no additional information is provided to the model. However, its drawback lies in the relatively large computational load.

3) Method for Selecting a Balanced Dataset: After applying these two techniques, it was found that SMOTE is the optimal choice. The reason is that undersampling will significantly reduce the size of the original dataset, while SMOTE can generate new approximate samples, and the number of these samples is sufficient for model training. Moreover, experimental verification shows that the training results of the dataset generated by undersampling on the model are much worse than those of oversampling. There is approximately a 10% difference in accuracy.

4) Selecting Features from the Original Dataset: The final step is to improve the performance of the model and reduce the prediction time of the model by selecting certain features. First, analyze the correlations among the input variables and pick one of the variables with strong correlations. As shown in

Fig. 4, only ”pdays” and ”previous” among the input variables have relatively high correlations, but still below 0.5. Considering that the subsequent models are mainly tree models, no variables were deleted. Subsequently, it is necessary to evaluate the relationship between each input variable and the target variable and select the input variables with the most stable relationships. In this study, both filtering and wrapping methods were employed for analysis.

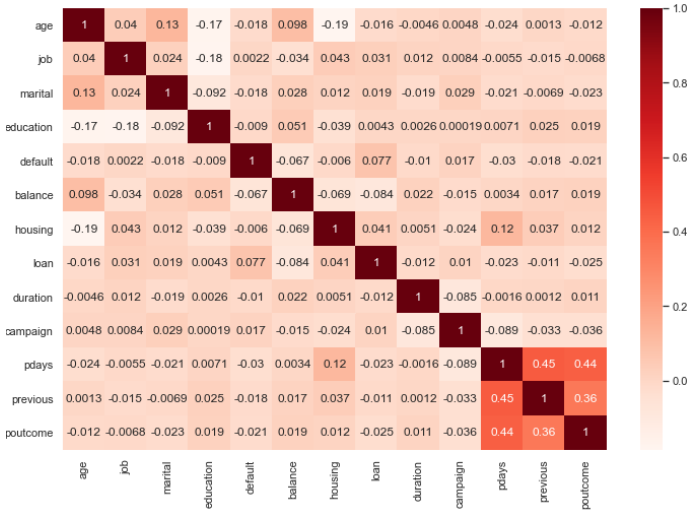


Fig. 4: Heatmap of the Correlation among Input Variables

The filtering method calculates the score for each feature associated with the target and then selects the best score among them. It is not a learning process; instead, it searches for relevant labeled features prior to the learning process [14]. The advantage of this algorithm is that it is computationally inexpensive and can avoid features that have the least impact on the target [15]. The wrapper method creates a subset of the dataset by training (a machine learning module). Then, it repeats this training process by adding or removing some features until the optimal combination is found to achieve the goal (using a greedy algorithm to find the optimal combination). The common techniques of this method are forward selection and backward elimination [16]. The advantage of this algorithm is that the selected features are of high quality, but its computational cost is also very high [17]. After conducting comparative experiments, it was found that the wrapper method achieves higher accuracy on the model. The characteristic values selected are shown in Table I . Therefore, the features selected by this method are used as the model inputs, and normalization processing is carried out before input.

III. Classifier Selection

After the previous steps, the data has been successfully pre - processed. Next, we will implement machine - learning algorithms, which can be divided into two categories: parametric and non - parametric classifiers. The aim of this process is to find the best model for each of the two parts and also to select the best model of each type as the final model. Non - parametric classifiers are also known as ”lazy learning” because they do not use assumptions during learning. Simply put, they use the samples collected in the training data [18]. The algorithms under this technique are XGBoost and Random Forest. After the processing in the previous steps, this paper applies the dataset to XGBoost and Random Forest, compares the results, and selects the best algorithm between them. The parametric classifier algorithm is the so - called metric algorithm, in which linear regression is often used [19]. Algorithms such as Naive Bayes and SVM belong to this technique. Following the processing of the previous steps, this paper will apply the dataset to the Naive Bayes classifier and the Support Vector Machine, compare the results, and select the best algorithm from them. Next, a brief review of the classification principles of the four algorithms will be presented.

TABLE I: The finally selected eigenvalues

No.	Features
0	age
1	job
2	education
3	loan
4	contact
5	duration
6	campaign
7	pdays
8	previous
9	poutcome

A. Random Forest Model

The random forest model is an algorithm that integrates multiple trees based on the concept of ensemble learning. Its fundamental unit is the decision tree, and essentially, it falls under a major branch of machine learning—ensemble learning methods. For classification problems, each decision tree within the random forest serves as a classifier. Consequently, for a given input sample, there will be N classification results from N trees. The random forest then aggregates all these classification voting results and designates the category with the most votes as the final output. This is also the simplest application of the Bagging approach.

B. XGBoost

XGBoost is an improved version of the GBDT (Gradient Boosting Decision Tree) [20]. It also belongs to a decision-tree-based model, of course. The basic components of XGBoost are decision trees, which can also be referred to as ”weak learners”. These ”weak learners” jointly form XGBoost, and they are generated in sequence. When generating the next decision tree, the prediction result of the previous one will be taken into account, that is, the deviation of the previous decision tree is considered. To predict a new sample, first, input it into each decision tree of XGBoost in turn. Then,

each decision tree will generate a predicted value, and finally, sum up all these predicted values to obtain the final prediction result. The advantages of XGBoost lie in its ability to automatically handle missing values and better process high-dimensional data. Besides, the XGBoost model can cope with noisy data more effectively and train the model more quickly and efficiently.

C. Naive Bayes Model

The Naive Bayes model is a classification method based on Bayes' theorem and the assumption of conditional independence of features. The Bayesian method classifies sample datasets using probability and statistics knowledge, with Bayes' principle as its foundation. Thanks to its solid mathematical footing, the misclassification rate of the Bayesian classification algorithm is quite low. Its distinctive feature is the combination of prior probability and posterior probability, which averts the subjective bias that comes with relying solely on prior probability and also avoids the overfitting issue that occurs when using sample information alone. The Bayesian classification algorithm can achieve a relatively high accuracy rate when dealing with large datasets, and the algorithm itself is rather straightforward. The Naive Bayes method simplifies the Bayesian approach accordingly. It assumes that, given the target value, the attributes are conditionally independent of one another. In other words, no single attribute variable holds greater or lesser weight in determining the decision result. Although this simplification does somewhat reduce the classification effectiveness of the Bayesian classification algorithm, it greatly simplifies the complexity of the Bayesian method in practical application scenarios.

D. Support Vector Machine

The principle of the support vector machine (SVM) is to search for the maximum margin among data points. It classifies data points into two categories and locates a maximum margin so that the two sets of data points are kept as far away from the margin as possible, thus achieving the goal of classification. The advantages of the SVM model are numerous. It can handle non-linear data, manage high-dimensional data more effectively, cope with noisy data better, and converge more rapidly. Moreover, the SVM model can utilize a variety of different kernel functions. These kernel functions enable it to fit the data more precisely, thereby enhancing the accuracy of the model.

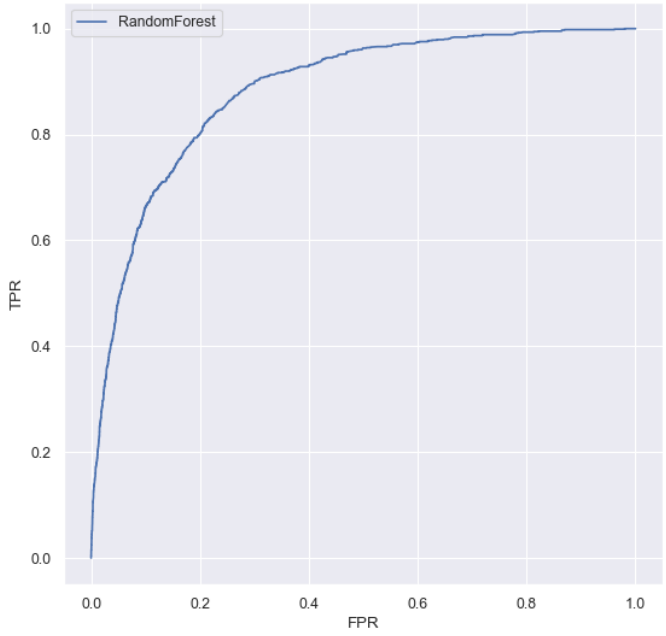
IV. Experimental Results and Analysis

All experiments in this study were conducted using Python. After preprocessing, the initial data contained 51211 rows. It was partitioned such that 80% of the data was used for the training set, and 20% was used for the testing set. A k-fold cross-validation with better performance was adopted, with k set to 5. In terms of result evaluation, accuracy (Accuracy), F1 value (F1-Score), the ROC curve, and the AUC value [21] were utilized to assess the models. Judging from the final evaluation results, the non-parametric model, random forest, achieved the highest score, with an accuracy rate of 89.58% and the highest AUC index of 0.88301. Next was the parametric model SVM, with an accuracy rate of 89.09% and an AUC index of 0.86887.

A. Non-parametric Classifiers

The results of the two non-parametric classifiers, XGBoost and random forest, are presented in Fig. 5 and Fig. 6. It is evident that each metric of the random forest model outperforms that of XGBoost, which indicates that the random forest has more accurate prediction results and better performance compared to XGBoost. In practical applications within the banking sector, this implies that the random forest model can provide more precise guidance for customer segmentation and targeted marketing strategies. It can identify with higher accuracy those customers who are most likely to respond positively to the bank's offers, thereby optimizing the allocation of marketing resources and maximizing the return on investment.

auc: 0.8830133414140171

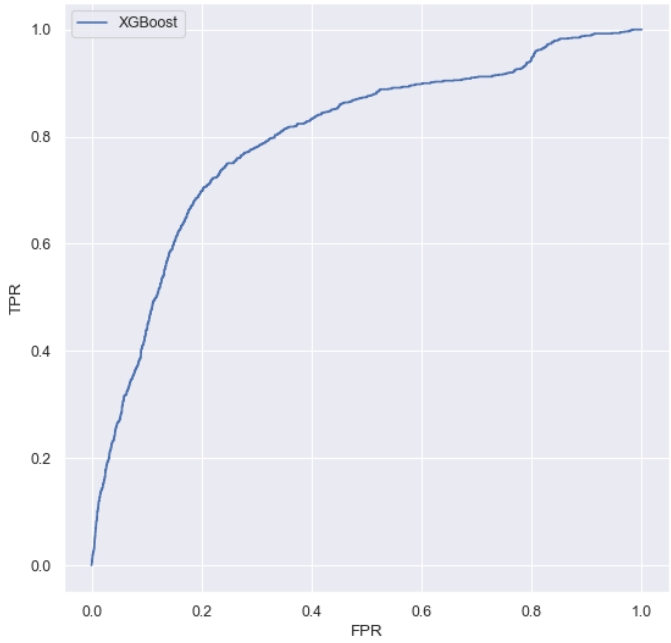


Model Accuracy : 89.58%
Model F1-Score : 89.96%

	precision	recall	f1-score	support
0	0.95	0.93	0.94	8147
1	0.48	0.56	0.52	896
accuracy			0.90	9043
macro avg	0.71	0.75	0.73	9043
weighted avg	0.90	0.90	0.90	9043

Fig. 5: Evaluation Results of the Random Forest

auc: 0.7942699153540488

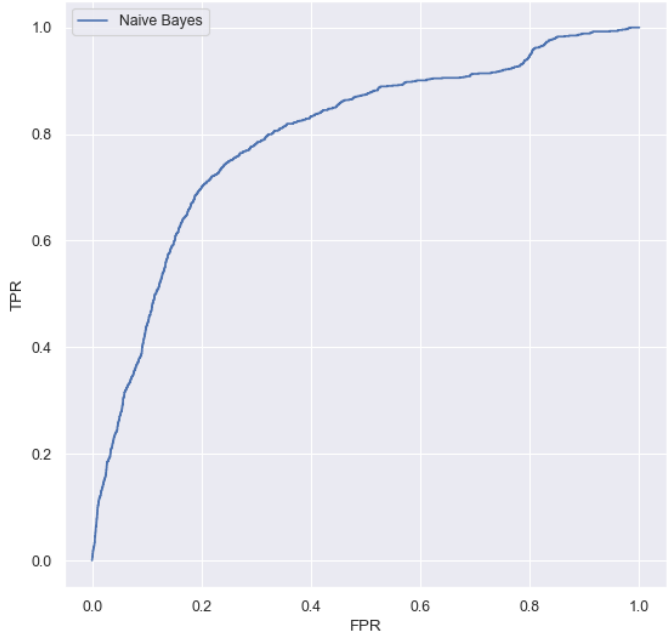


Model Accuracy : 82.62%
Model F1-Score : 80.87%
Cross Val Accuracy: 79.51 %
Cross Val Standard Deviation: 0.76 %

	precision	recall	f1-score	support
0	0.86	0.94	0.90	7295
1	0.58	0.35	0.44	1748
accuracy			0.83	9043
macro avg	0.72	0.65	0.67	9043
weighted avg	0.81	0.83	0.81	9043

Fig. 6: Evaluation Results of XGBoost

auc: 0.7940999369091147



Model Accuracy : 82.48%
Model F1-Score : 80.71%
Cross Val Accuracy: 79.62 %
Cross Val Standard Deviation: 0.78 %

	precision	recall	f1-score	support
0	0.86	0.94	0.90	7287
1	0.58	0.35	0.44	1756
accuracy			0.82	9043
macro avg	0.72	0.64	0.67	9043
weighted avg	0.80	0.82	0.81	9043

Fig. 7: Evaluation Results of Naive Bayes

B. Parametric Classifiers

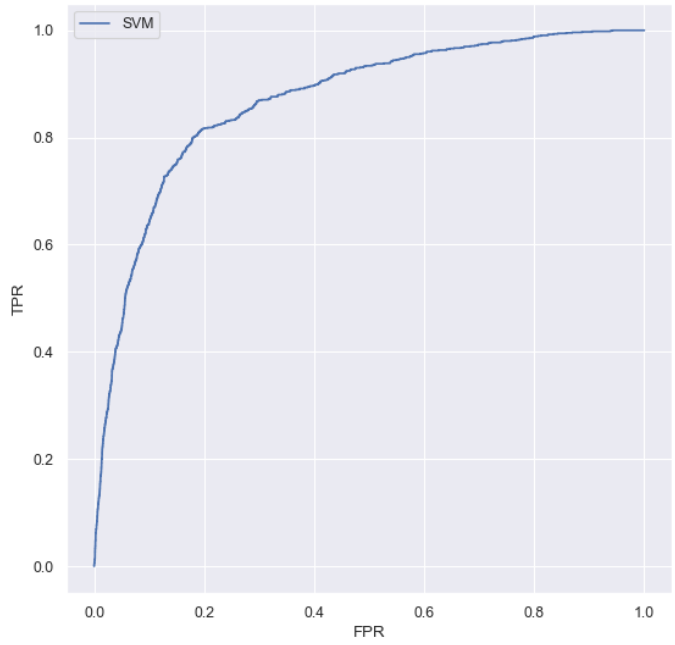
The evaluation results of the two parametric classifiers, the Naive Bayes model and the Support Vector Machine (SVM) model, are shown in Fig. 7 and Fig. 8. Judging from the results, the evaluation indicators of the SVM far exceed those of the Naive Bayes model. This demonstrates that, for the dataset processed by this method, the SVM has a better prediction effect.

This advantage is of great significance in the actual bank marketing scenario. For banks, a more accurate prediction model means being able to identify potential customers more effectively and reduce unnecessary waste of marketing resources. When dealing with complex customer data, the powerful nonlinear processing ability of SVM enables it to capture the hidden patterns and relationships in the data. In contrast, the Naive Bayes model, due to its relatively simple assumptions, appears to be inadequate when faced with complex data structures.

V. Conclusion

In this paper, a prediction model was established through a series of data processing and model comparisons using a given bank customer dataset. This model, which is a binary classification model, can predict how customers will respond to bank promotion activities based on given input variables. Due to the obvious imbalance in the dataset, during the preliminary data processing, conventional preprocessing methods were employed, such as data cleaning, discrete variable transformation, feature selection, and normalization. Additionally, two data-balancing methods were compared, and the SMOTE method, which yields better results, was selected. Regarding model selection, four classifiers, namely SVM, Random Forest, Naive Bayes, and XGBoost, were contrasted. The Random Forest model, with an accuracy rate of 89.6% and an AUC of 0.88, was chosen for its optimal evaluation results. Thus, the goal of the prediction model was achieved: to help banks predict the target population for marketing, boost sales, and cut costs.

auc: 0.8688722237939048



Model Accuracy : 89.09%
Model F1-Score : 89.33%
Cross Val Accuracy: 85.17 %
Cross Val Standard Deviation: 1.24 %

	precision	recall	f1-score	support
0	0.94	0.93	0.94	8090
1	0.48	0.54	0.51	953
accuracy			0.89	9043
macro avg	0.71	0.73	0.72	9043
weighted avg	0.90	0.89	0.89	9043

Fig. 8: Evaluation Results of SVM

In the future, the neural network approach in deep learning could be adopted to train the model, or further improvements could be made to the existing models to enhance the prediction effect and attain higher precision.

REFERENCES

[1] C. I. Mbama and P. O. Ezepue, "Digital banking, customer experience and bank financial performance," International Journal of Bank Marketing, vol. 36, no. 2, pp. 230–255, Apr. 2018, doi: 10.1108/IJBM-11-2016-0181.

[2] A. A. B. Ng and N. L. Abdullah, "Security challenges in designing an integrated web application for multiple online banking," in 2010 International Symposium on Information Technology, Jun. 2010, pp. 1–5, doi: 10.1109/ITSIM.2010.5561291.

[3] L. Sijia, T. Lan, Z. Yu, and Y. Xiuliang, "Comparison of the prediction effect between the logistic regressive model and SVM model," in 2010 2nd IEEE International Conference on Information and Financial Engineering, Sep. 2010, pp. 316–318, doi: 10.1109/ICIFE.2010.5609308.

[4] M. J. Cronin, Banking and finance on the internet. John Wiley & Sons, 1997.

[5] L.-L. Li, Z.-F. Liu, M.-L. Tseng, K. Jantarakolica, and M. K. Lim, "Using enhanced crow search algorithm optimization-extreme learning machine model to forecast short-term wind power," Expert Systems with Applications, vol. 184, Dec. 2021, doi: 10.1016/j.eswa.2021.115579.

[6] N. Roy, R. Ahmed, M. R. Huq, and M. M. Shahriar, "User-centric activity recognition and prediction model using machine learning algorithms," International Journal of Advanced Computer Science and Applications (IJACSA), vol. 12, no. 12, pp. 501–510, 2021.

[7] S. Abbas, "Deposit subscribe prediction using data mining techniques based real marketing dataset," International Journal of Computer Applications, vol. 110, no. 3, pp. 1–7, Jan. 2015, doi: 10.5120/19293-0725.

[8] B. Zhang, "Tactical decision system of table tennis match based on C4.5 decision tree," in 2021 13th International Conference on Measuring Technology and Mechatronics Automation (ICMTMA), Jan. 2021, pp. 632–635, doi: 10.1109/ICMTMA52658.2021.00146.

[9] N. Rochmawati et al., "Covid symptom severity using decision tree," in 2020 Third International Conference on Vocational Education and Electrical Engineering (ICVEE), Oct. 2020, pp. 1–5, doi: 10.1109/ICVEE50212.2020.9243246.

[10] K. R. Singh, K. P. Neethu, K. Madhurekaa, A. Harita, and P. Mohan, "Parallel SVM model for forest fire prediction," Soft Computing Letters, vol. 3, Dec. 2021, doi: 10.1016/j.socl.2021.100014.

[11] A. A. Supianto, A. Julisar Dwitama, and M. Hafis, "Decision tree usage for student graduation classification: A comparative case study in faculty of computer science Brawijaya University," in 2018 International Conference on Sustainable Information Engineering and Technology (SIET), Nov. 2018, pp. 308–311, doi: 10.1109/SIET.2018.8693158.

[12] M. Mimura, "Using fake text vectors to improve the sensitivity of minority class for macro malware detection," Journal of Information Security and Applications, vol. 54, Oct. 2020, doi: 10.1016/j.jisa.2020.102600.

[13] R. Geetha, S. Sivasubramanian, M. Kaliappan, S. Vimal, and S. Annamalai, "Cervical cancer identification with synthetic minority oversampling technique and PCA analysis using random forest classifier," Journal of Medical Systems, vol. 43, no. 9, Art. no. 286, Sep. 2019, doi: 10.1007/s10916-019-1402-6.

[14] A. Bommert, X. Sun, B. Bischl, J. Rahnenführer, and M. Lang, "Benchmark for filter methods for feature selection in high- dimensional classification data," Computational Statistics and Data Analysis, vol. 143, Mar. 2020, doi: 10.1016/j.csda.2019.106839.

[15] Y. Jiang, X. Liu, G. Yan, and J. Xiao, "Modified binary cuckoo search for feature selection: A hybrid filter-wrapper approach," in 2017 13th International Conference on Computational Intelligence and Security (CIS), Dec. 2017, pp. 488–491, doi: 10.1109/CIS.2017.00113.

[16] S. M. Kasongo and Y. Sun, "A deep learning method with wrapper based feature extraction for wireless intrusion detection system," Computers & Security, vol. 92, May 2020, doi: 10.1016/j.cose.2020.101752. Transactions on Cloud Computing, vol. 10, no. 4, pp. 2787-2803, 1 Oct.-Dec. 2022.

[17] H. Das, B. Naik, and H. S. Behera, "A Jaya algorithm based wrapper method for optimal feature selection in supervised classification," Journal of King Saud University - Computer and Information Sciences, vol. 34, no. 6, pp. 3851–3863, Jun. 2022, doi: 10.1016/j.jksuci.2020.05.002.

[18] A. Gupta et al., "A novel approach for classification of mental tasks using multiview ensemble learning (MEL)," Neurocomputing, vol. 417, pp. 558–584, Dec. 2020, doi: 10.1016/j.neucom.2020.07.050.

[19] D. Lestari, R. R. Bintana, and N. Budiman, "Online internship acceptance registration application at bank," Computer Science and Informatics Journal, vol. 3, no. 2, pp. 127–138.

[20] Friedman, J. H. (2001). Gradient boosting machines: A new machine learning method. Annals of Statistics, 29(5), 1189-1232.

[21] H. A.Younis, A. S. A. Mohamed, R. Jamaludin, and M. N. A. Wahab, "Survey of robotics in education, taxonomy, applications, and platforms during COVID-9," Computers, Materials and Continua, vol. 67, no. 1, pp. 687–707, 2021, doi: 10.32604/cmc.2021.013746.