European Economic Letters
ISSN 2323-5233
Vol 13, Issue 3 (2023)
http://eelet.org.uk

The Use of Machine Learning Algorithms for Bank Loan Prediction

Chen Hao[1] and Suwannee Adsavakulchai[2*]
School of Engineering, University of the Thai Chamber of Commerce[1,2]
126/1 Vivphavedee Rangsit Road, Dindieng, Bangkok THAILAND

Abstract
Background: Bank loan prediction is an important problem in the banking industry. By using historical data and applying predictive models, banks can identify patterns and make accurate predictions about loan defaults. This can help them make informed decisions about lending and minimize their losses.
Objectives: To study the important parameters that influence loan and to predict the bank loan using machine learning algorithms
Methods: The CRISP-DM process is a comprehensive and structured approach to developing predictive models. By following this process, the study can ensure that all necessary steps are taken to develop an accurate and reliable predictive model for personal loan. The use of three machine learning algorithms such as decision tree, naïve bayes, and support vector machine can provide for developing the model and enable the study to select the best one.
Results: The results suggest that the J48 Decision Tree algorithm achieved the highest accuracy of 98.85%, followed by the SVM algorithm with an accuracy of 94.01%, and the Naive Bayes algorithm with an accuracy of 89.53%. In terms of precision, recall, and F-measure, all three algorithms achieved similar performance, with values ranging from 0.895 to 0.989.
Conclusions: The performance of different machine learning algorithms in predicting bank loan showed that J48 DT was the most appropriate algorithm for developing a bank loan predictor, based on its high accuracy, low mean absolute error, and fast training time. To improve the accuracy and applicability of the model, it may be necessary to collect additional data or refine the feature selection process to identify the most relevant attributes.

Keywords: bank loan, SMOTE, Naïve Bayes, Support Vector Machine, Decision Tree

1. Introduction

When banks lend money to individuals or businesses, they need to assess the creditworthiness of loan applicants before approving the loan. The assessment involves evaluating various factors such as credit score, income, employment history, and debt-to-income ratio. Loan is a major concern for banks that is always looking for ways to reduce the risk associated with lending money to customers. Machine learning algorithms can help banks automate the loan approval process, reducing the time and resources required to manually process loan applications. This can improve the customer experience and increase the efficiency of the loan approval process [1].

The Cross Industry Standard Process for Data Mining (CRISP-DM) is a widely used process for solving business problems. It is a comprehensive and structured approach to developing predictive models. In this context, we can use CRISP-DM to solve the bank loan prediction problem. The CRISP-DM process, we can develop a machine learning model for bank loan prediction that can help banks make informed decisions about lending and minimize their losses [2].

The J48 algorithm is a popular decision tree algorithm that is widely used in machine learning for classifica- tion tasks. The goal is to create a tree that can classify the data accurately while keeping the tree as small as possible [3].

The Naïve Bayesian algorithm is a simple powerful algorithm for classification tasks. It works by first computing the prior probability of each class based on the training data and then using Bayes' theorem to compute the posterior probability of each class given the feature values of a new instance. The class with the highest posterior probability is then predicted as the class of the new instance. [4].

The Support Vector Machine (SVM) algorithm is a powerful and widely used algorithm for binary classification tasks. The hyperplane is chosen to maximize the margin, which is the distance between the hyperplane and the

closest points of each class. By maximizing the margin, the SVM algorithm aims to find the decision boundary that is most robust to new data and minimizes the classification errors. [5].

SMOTE (Synthetic Minority Over-sampling Technique) is an oversampling technique used to address the class imbalance problem in machine learning. SMOTE creates synthetic minority class samples by interpolating between existing minority class samples. The formula for generating a synthetic sample using SMOTE is:

new_sample = base_sample + (random_number * (nearest_neighbor - base_sample))

where:

- new_sample: the synthetic sample being generated
- base_sample: the selected minority class sample being used as the base for generating the synthetic sample
- nearest_neighbor: the randomly selected nearest neighbor of the base sample
- random_number: a randomly generated number between 0 and 1 used to control the amount of interpolation between the base sample and nearest neighbor.

2. Objectives

1. To study the important parameters that influence loan.
2. To predict the bank loan using machine learning algorithms.

3. Methods

There are six steps of the CRISP-DM process for bank loan prediction as following:

1. Business Understanding: to understand the business problem we are trying to do the bank loan prediction.
2. Data Understanding: to gather and understand the data related to the problem. We collect data about the loan applicants available from https://www.kaggle.com/datasets/itsmesunil/bank-loan-modelling.

There are 14 attributes that are the factors related to the bank loan, and 1 class attribute "personal loan" that determines the personal loan as shown in Table 1.

Table 1 The metadata of the dataset.

Attribute	Type	Data Description
ID	Numeric	Customer ID
Age	Numeric	Customer's age
Experience	Numeric	#years of professional experience
Income	Numeric	Annual income of the customer
ZIP Code	Numeric	ZIP code of home address
Family	Numeric	Family size of the customer
CCAvg	Numeric	Avg. credit cards spending per month
Education	Numeric	Education Level.1: Undergrad; 2: Graduate; 3: Advanced/
Mortgage	Numeric	Value of house mortgage
Personal Loan	Numeric	Customer accept the personal loan offered in the last campaign?
Securities	Numeric	Customer has a securities account with the bank?
CD Account	Numeric	Customer has a certificate of deposit account with the bank?
Online	Numeric	Customer use internet banking?
Credit Card	Numeric	Credit card issued by Universal Bank?

3. Data Preparation: to clean, transform, and prepare the data for analysis. This involves handling missing values, dealing with outliers, encoding categorical variables, etc.
   - In data pre-processing, the age, income, and experience features are grouped into five categories, from 1 to 5. This grouping helps to simplify the data and reduce its sparsity, which can improve the performance of the machine learning algorithms.

- The data type of the numeric features is converted to nominal, which is suitable for the machine learning algorithms. The class attribute "personal loan" is moved to the end of the dataset as the target variable for prediction.
- The presence of class imbalance in the dataset is a 10:1 ratio between the two classes. This affect the performance of the machine learning algorithms, as they may be biased towards the majority class. Therefore, techniques such as oversampling or undersampling may need to be applied to balance the classes and improve the accuracy of the model.

4. Modeling: to select an appropriate machine learning model to predict loan repayment. We can use models such as Decision tree (J48); Naïve bayes (NB); Support Vector Machine algorithms (SMO).
5. Evaluation: to evaluate the performance of the model on the test data. We can use metrics such as accuracy, precision, recall, and F1 score to evaluate the model's performance.
6. Deployment: to deploy the model in the production environment. This involves integrating the model with the bank's loan application system so that it can predict the loan repayment probability for new loan applicants.

4. Results

1. Business Understanding: In this case, the problem is to predict whether a loan applicant will repay the loan or default.
2. Data Understanding: There are eleven parameters that are age, experience, income, facility, CCAvg, education, marriage, securities account, CD account, online, credit card. Some parameters like "ID" and "ZIP Code" are not related to the model are removed from the dataset. This helps to reduce noise in the data and improve the efficiency of the machine learning algorithms.
3. Data Preparation: the dataset is still imbalanced. By applying SMOTE, the number of samples in the minority class by interpolating between existing minority class samples. This helps to increase the number of samples in the minority class and balance the distribution of samples across the classes as shown in Figure 1.

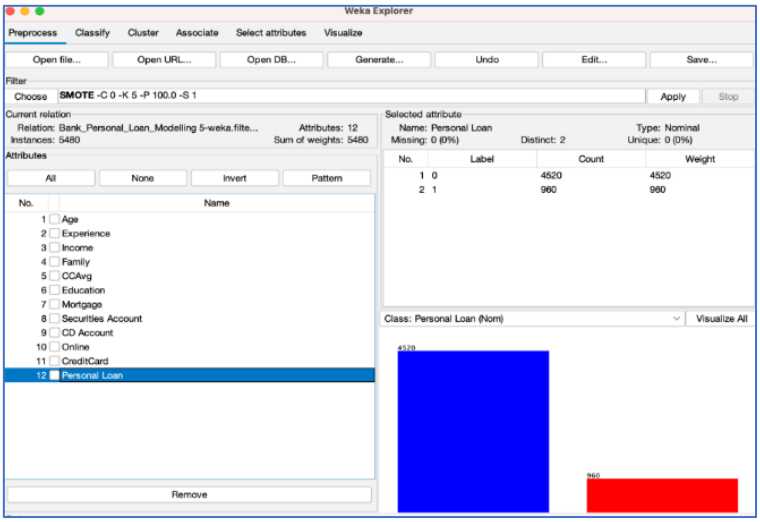

Figure 1: The data is class imbalance.

In Figure 1, this can improve the performance of machine learning models and their ability to generalize to new data. However, the performance improvement will depend on various factors such as the quality of the data, the choice of the machine learning algorithm, and the hyperparameters of the algorithm [6,7] as shown in Figure 2.

European Economic Letters
ISSN 2323-5233
Vol 13, Issue 3 (2023)
http://eelet.org.uk

Figure 2: The result from synthetic minority oversampling technique (SMOTE)

After applying SMOTE, Figure 3 likely shows the performance improvement of the machine learning model on the imbalanced dataset before and after applying SMOTE or other similar techniques.

4. Modeling: After data processing using the SMOTE and three algorithms, namely Decision tree (J48); Naïve bayes (NB); Support Vector Machine algorithms (SMO) using WEKA. To determine the best algorithm, the performance of the hyperparameter optimization process can help to fine-tune the model and improve its performance on the given dataset by several trainings with the same dataset for each algorithm.

The hyperparameter settings helped to optimize the performance of the J48 algorithm on the Bank_Loan_Modelling dataset. The resulting decision tree was both accurate and interpretable, making it a suitable algorithm for this type of predictive modelling task as shown in Table 2.

Table 2 J48 algorithm hyperparameter optimization

Training Number	Confidence Factor	Unpruned	Accuracy
1	0.25	FALSE	97.5%
2	0.5	FALSE	97.97%
3	0.75	FALSE	98.85%
4	0.25	TRUE	98.85%
5	0.5	TRUE	98.85%

From Table 2, the model trained with a confidence factor of 0.75 achieved the highest accuracy of 98.85%, while the models trained with confidence factors of 0.25 and 0.5 achieved lower accuracies of 97.5% and 97.96%, respectively. However, the optimal confidence factor depends on the specific dataset and the problem, and that there may be a trade-off between accuracy and other performance metrics such as recall, precision, and F1 score. Therefore, it's recommended to perform a thorough hyperparameter tuning process and evaluate the performance of the J48 algorithm across multiple performance metrics before selecting the optimal hyperparameters for a given task.

In Naive Bayes, the Kernel Density Estimation (KDE) is a non-parametric method used to estimate the probability density function (PDF) of the continuous features in the dataset. The Naive Bayes algorithm assumes that the features are conditionally independent given the class label, which means that the joint PDF of the features can be factorized as a product of individual PDFs. The details are shown in Table 3.

European Economic Letters
ISSN 2323-5233
Vol 13, Issue 3 (2023)
http://eelet.org.uk

The results in Table 3 suggest that for the specific dataset and problem being studied, the default values for the Kernel Estimator and Supervised Discretization parameters in the WEKA software were appropriate. This is because there was no significant difference in the performance of the Naïve Bayes algorithm when varying these parameters, with an accuracy of 89.53% for all three training runs.

Table 3 Naïve Bayes algorithm hyperparameter optimization

Training Number	Kernel Estimator	Supervised Discretization	Accuracy
1	FALSE	FALSE	89.53%
2	TRUE	FALSE	89.53%
3	FALSE	TRUE	89.53%

SVM algorithm was trained and evaluated with different values of the regularization parameter C and different kernel functions (Poly Kernel and Normalized Poly Kernel) to optimize its performance on the given dataset. [17]. The details are shown in Table 4.

Table 4 SVM algorithm hyperparameter optimization

Training Number	C	Kernel	Accuracy
1	1	Poly Kernel	93.85%
2	2	Poly Kernel	93.96%
3	3	Poly Kernel	93.96%
4	1	Normalized Poly Kernel	93.80%
5	2	Normalized Poly Kernel	94.01%
6	3	Normalized Poly Kernel	93.85%

The results in Table 4 suggest that the SVM algorithm achieved the highest accuracy of 94.01% when using a C value of 2 with the Normalized Poly Kernel. Specifically, the accuracy of the model was 93.85%, 93.96%, and 93.96% for the Poly Kernel with C values of 1, 2, and 3, respectively. For the Normalized Poly Kernel, the accuracy of the model was 93.80%, 94.01%, and 93.85% for C values of 1, 2, and 3, respectively.

5. Evaluation: The evaluation metrics used include accuracy, precision, recall, and F-measure. The results suggest that the J48 Decision Tree algorithm achieved the highest accuracy of 98.85%, followed by the SVM algorithm with an accuracy of 94.01%, and the Naive Bayes algorithm with an accuracy of 89.53%. In terms of precision, recall, and F-measure, all three algorithms achieved similar performance, with values ranging from 0.895 to 0.989. The results of the model's accuracy performance are displayed in Table 5.

Table 5 Major accuracy performance of the models using machine learning algorithms.

Algorithm	Accuracy	Precision	Recall	F-Measure
J48 DT	98.85%	0.989	0.989	0.989
NB	89.53%	0.896	0.895	0.895
SVM	94.01%	0.94	0.94	0.94

In Table 5, it appears that the performance of three machine learning algorithms (J48 Decision Tree, Naive Bayes, and SVM) has been evaluated using a testing set.

The performance is evaluated using three metrics: Kappa Statistic, Mean Absolute Error, and Time in seconds are displayed in Table 6.

Table 6 Machine learning algorithms' performance

Algorithm	Kappa Statistic	Mean Absolute Error	Time in seconds
J48 DT	0.9771	0.0182	0
NB	0.7906	0.1494	0
SVM	0.8802	0.2447	0.43

Kappa Statistic is a measure of the agreement between the predicted and actual outcomes. A value closer to 1 indicates high agreement, while a value closer to 0 indicates poor agreement. In this case, J48 decision tree has the highest Kappa Statistic of 0.9771, indicating that it has the best performance in predicting personal loans.

Mean Absolute Error measures the average difference between the predicted and actual outcomes. A lower value indicates better performance. In this case, J48 decision tree has the lowest Mean Absolute Error of 0.0182, indicating that it has the best performance in predicting personal loans.

Time in seconds measures the time taken by the algorithm to train and predict. In this case, J48 decision tree and Naive Bayes took 0 seconds, while SVM took 0.43 seconds.

5. Conclusion

The results suggest that the J48 Decision Tree algorithm achieved the highest accuracy of 98.85%, followed by the SVM algorithm with an accuracy of 94.01%, and the Naive Bayes algorithm with an accuracy of 89.53%. In terms of precision, recall, and F-measure, all three algorithms achieved similar performance, with values ranging from 0.895 to 0.989. The performance of different machine learning algorithms in predicting bank loan showed that J48 DT was the most appropriate algorithm for developing a bank loan predictor, based on its high accuracy, low mean absolute error, and fast training time. To improve the accuracy and applicability of the model, it may be necessary to collect additional data or refine the feature selection process to identify the most relevant attributes.

6. Discussion

The model developed in this study shows promising results, there are still limitations and areas for improvement. As mentioned, other factors such as assets and debts can have a significant impact on a customer's likelihood to loan approval. Additionally, it is important to consider some of the attributes in the dataset may not have a strong logical connection with the target variable. To improve the accuracy and applicability of the model, it may be necessary to collect additional data or refine the feature selection process to identify the most relevant attributes.

The findings of the study provide valuable insights into the development of predictive models for bank loan behavior, but further research and refinement are necessary to address the limitations and potential biases of the model. It is important to consider the ethical implications of using predictive models in decision-making and ensure that any biases or discrimination are addressed to promote fairness and equity in lending practices.

References

[1]. A. Kumar, S. Sharma, & M. Mahdavi, "Machine Learning (ML)Technologies for Digital Credit Scoring in Rural Finance: A Literature Review." Risks 9.11 (2021): 192.

[2]. Madane N and Nanda S 2019 Loan prediction analysis using decision tree Journal of The Gujarat Research Society 21 p p 214–21

[3]. Supriya P, Pavani M, Saisushma N, Kumari N V and Vikas K 2019 Loan prediction by using machine learning models Int. Journal of Engineering and Techniques 5 pp144–8

[4]. E. G. Kulkarni and R. B. Kulkarni, "Weka powerful tool in data mining", *International Journal of Computer Applications*, vol. 975, pp. 8887, 2016.

[5]. Chaudhury, S., Dhabliya, D., Madan, S., & Chakrabarti, S. (2023). Blockchain Technology: A Global Provider of Digital Technology and Services. In Building Secure Business Models Through Blockchain Technology: Tactics, Methods, Limitations, and Performance (pp. 168–193). IGI Global.

[6]. M. Malvoni, M. G. De Giorgi, and P. M. Congedo, "Data on support vector machines (SVM) model to forecast photovoltaic power," Data in Brief, vol. 9, no. C, pp. 13–16, 2016.

[7]. Y. Liu, C. Liu, and S. Tseng, "Deep discriminative features learning and sampling for imbalanced data problem," in IEEE International Conference on Data Mining, ICDM 2018, Singapore, November 17- 20, 2018. IEEE Computer Society, 2018, pp. 1146–1151.

[8]. K. Qi, H. Yang, Q. Hu, and D. Yang, "A new adaptive weighted imbalanced data classifier via improved support vector machines with high-dimension nature," Knowl. Based Syst., vol. 185, 2019.

[9]. F. Bao, Y. Deng, Y. Kong, Z. Ren, J. Suo, and Q. Dai, "Learning deep landmarks for imbalanced classification," IEEE Trans. Neural Networks Learn. Syst., vol. 31, no. 8, pp. 2691–2704, 2020.

[10]. X. Jing, X. Zhang, X. Zhu, F. Wu, X. You, Y. Gao, S. Shan, and J. Yang, "Multiset feature learning for highly imbalanced data classification," IEEE Trans. Pattern Anal. Mach. Intell., vol. 43, no. 1, pp. 139–156, 2021.

[11]. C. Bellinger, R. Corizzo, and N. Japkowicz, "Remix: Calibrated resampling for class imbalance in deep learning," CoRR, vol. abs/2012.02312, 2020. [Online]. Available: https://arxiv.org/abs/ 2012. 02312

[12]. Tharwat A (2020) Classification assessment methods. New England Journal of Entrepreneurship 17(1):168–192

[13]. Sasaki Y, Fellow R (2007) The truth of the f-measure, Manchester: Mib-school of computer science. University of Manchester p 25

[14]. Powers DM (2020) Evaluation: from precision, recall and f-measure to roc, informed Ness, markedness and correlation. arXiv:201016061

[15]. Kawale, S., Dhabliya, D., & Yenurkar, G. (2022). Analysis and Simulation of Sound Classification System Using Machine Learning Techniques. 2022 International Conference on Emerging Trends in Engineering and Medical Sciences (ICETEMS), 407–412. IEEE.

[16]. C. Bellinger, C. Drummond, and N. Japkowicz, "Manifold-based synthetic oversampling with manifold conformance estimation," Mach. Learn., vol. 107, no. 3, pp. 605–637, 2018.

[17]. Demšar J (2006) Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research 7:1–30

[18]. H. He, E.A. Garcia, Learning from imbalanced data, IEEE Trans. Knowledge Data Eng. 21 (9) (2009) 1263–1284.

[19]. Case, B., & Zucker, S. (2005, July). Methodologies for alignment of standards and assessments [Paper presentation]. China-US Conference on Alignment of Assessments and Instruction, Beijing, China.

[20]. Li Yong, Xu De-zhi, Zhang Yong and Xing Chun-xiao. MVC-based Incremental Reengineering Approach. Journal of Chinese Computer Systems, 29(3):469-472, 2008.

[21]. Pareek, M., Gupta, S., Lanke, G. R., & Dhabliya, D. (2023). Anamoly Detection in Very Large Scale System using Big Data. SK Gupta, GR Lanke, M Pareek, M Mittal, D Dhabliya, T Venkatesh,.." Anamoly Detection in Very Large Scale System Using Big Data. 2022 International Conference on Knowledge Engineering and Communication Systems (ICKES).

[22]. Prasad, K.G.S., P.V.S. Chidvilas, and V.V. Kumar, Customer loan approval classification by supervised learning model. International Journal of Recent Technology and Engineering, 2019. 8(4): 9898-9901.

[23]. J. D´ıez-Pastor, J. J. R. Diez, C. I. Garc´ıa-Osorio, and L. I. Kuncheva, "Random balance: Ensembles of variable priors classifiers for imbalanced data," Knowl. Based Syst., vol. 85, pp. 96–111, 2015.