Comparative Analysis of Generalization for Machine Learning Application in Loan Default Binary Classification

Jiayuan Ma[a]
Faculty of Engineering and Computer Science, The University of Sydney, Sydney, Australia

Keywords: Machine Learning, Generalization, Loan Default.

Abstract: Predicting loan defaults is critical for banks because it enables financial institutions to assess the risks associated with loan approval. While traditional methods of evaluating loan applicants rely on subjective assessments, Machine Learning (ML) has emerged as a powerful alternative to predicting defaults. This study conducted a comparative analysis of six ML models on a dataset of 255,347 loan applicants. The goal is to evaluate the generalization ability of each model when exposed to different data distributions. The results show significant performance degradation across all models when tested on unseen data, highlighting the issue of distribution shifts between training and testing sets. Some techniques such as domain adaptation and distribution alignment are discussed as potential solutions to improve model robustness. These findings provide valuable insights that could guide financial institutions in selecting more reliable, adaptable, and robust models for accurately predicting loan defaults, thus improving decision-making processes and reducing financial risk.

1 INTRODUCTION

In banks, predicting loan default is a critical component of deciding to accept a loan application, which is a need of becoming more and more commonplace in society because although it is a win-win financial activity for both lenders and borrowers, defaulting on a loan may lead to severe consequences, such as decrease of profit in banks, obstacles in the academic careers of the uninformed student population (Looney, 2022; Looney, 2019; Lakshmanarao, 2023; Baesens, 2003).

Traditionally, banks assess loan applicants based on educational background, occupation, and income. However, quantifying the importance of these factors is complex, often relies on the expertise and subjective judgment of professional bankers, and requires tons of labour-intensive and time-consuming (Lakshmanarao, 2023; Wu, 2019). In contrast, applying Machine Learning (ML) algorithms enhances the efficiency of solving the binary classification problem of loan approval, facilitating the selection of appropriate predictive models (Athey, 2018). For instance, Extensive research in this field provides evidence that multivariable regression often outperforms traditional methods (Serrano-Cinca, 2016). In addition, while although logistic regression is designed for distinguishing binary targets by using probability between 0 and 1, there is a study showing that it may not be suitable for situations involving nonlinear data, decision trees have demonstrated strong results, albeit with some risk of overfitting due to the model's principle that tries to divide the dataset repeatedly into nodes base on the level of entropy in different classes causing better performance in training set and worse in testing set (Lakshmanarao, 2023; Jin, 2015; Shih, 2014). In addition, random forests are a powerful and popular ensemble learning technique that consists of multiple decision trees. Each decision tree is trained on a different subset of the data set, providing individual estimates that are combined together to form a result. Random forest is more stable and accurate than single decision tree. It can effectively solve the overfitting problem of decision tree and enhance the generalization ability of the model to the unknown data in the future (Baesens, 2003). Moreover, the Naive Bayes algorithm performs well with small datasets, and K-Nearest Neighbours, a simple but effective model, tends to excel when the features in the dataset exhibit high

[a] https://orcid.org/0009-0009-8252-3713

Ma, J.
Comparative Analysis of Generalization for Machine Learning Application in Loan Default Binary Classification.
DOI: 10.5220/0013224800004568
In *Proceedings of the 1st International Conference on E-commerce and Artificial Intelligence (ECAI 2024)*, pages 319-323
ISBN: 978-989-758-726-9

correlation (Zareapoor, 2015). Furthermore, the advent of neural networks and deep learning has introduced even more promising outcomes (Wu, 2019; Chong, 2017).

Many researchers have compared the performance of various ML models in predicting loan defaults. Despite these efforts, there remains a gap regarding the generalization ability of these models across diverse datasets. Addressing this gap is essential for developing more reliable predictive tools, which is crucial for banks to make better-informed decisions when selecting predictive models. This study will utilize the Loan Default Prediction dataset from Coursera, which includes 255,347 samples and 16 features related to borrowers, providing a comprehensive basis for analysis. The results of this study could provide insights for banks, enabling them to adopt more robust and generalizable models for loan default prediction, ultimately making to better risk management and decision-making in the financial industry. Therefore, the primary objective of this study is to evaluate and compare the generalization ability of different models from ML in the context of predicting loan default. Various techniques of machine learning will be employed to achieve this objective, and performance indicators will be used.

2 METHOD

2.1 Dataset Setup

The dataset used in this study is sourced from Coursera's educational materials and contains information on bank loan defaults. It includes 255,347 samples, each representing 16 borrower features, such as age, income, education, marital status, and employment, along with a record of whether the borrower defaulted. The dataset underwent preprocessing before the experiments to enhance the prediction accuracy of loan defaults. This included normalizing quantitative features and encoding categorical variables. After data preprocessing, K-means clustering was employed to divide the dataset into two categories, labelled Group A and Group B. Group A represents the subset of data available for model training in banks. At the same time, Group B serves as an unseen sample set. Therefore, to address the class imbalance, only the Group A dataset was subjected to Synthetic Minority Over-sampling Technique (SMOTE) for oversampling. This approach ensures a more robust model performance evaluation in predicting loan defaults.

2.2 Machine Learning Models-based Prediction

This study employs six fundamental machine learning models. These models are integrated into the sklearn library, providing easy access for researchers to implement and use. Various performance evaluation metrics are also available to assess the models' effectiveness. This setup allows for comprehensive comparisons and evaluations of model performance in predicting outcomes based on the selected dataset.

2.2.1 K Nearest Neighbour

The K nearest Neighbour (KNN) algorithm is a simple and is commonly used for non-parametric classification (this study) and regression tasks. It calculates the distance between samples to be classified and all the samples in the training set select the nearest number of K neighbours and predicts the class of the new sample by voting or weighted average according to the class of these neighbours. The advantage of KNN is interpretable and adaptable to multi-class problems. However, it is computationally inefficient in high dimensional Spaces and sensitive to noise (Zhang, 2021).

2.2.2 Logistic Regression

Logistic regression (LR) is a widely used classification algorithm, especially for binary classification problems. It predicts the class by mapping a linear combination of the input variables to an interval from 0 to 1 and setting a threshold. LR assumes there is a linear relationship between input variables and output variables. It is easily interpretable and computationally efficient. However, LR has limited performance when facing complex nonlinear problems (Fernandes, 2020).

2.2.3 Support Vector Machine

Support Vector machines (SVM) are also supervised for classification and regression that perform classification by finding the optimal decision boundary (maximizing the margin between two classes). SVM performs well in dealing with small samples and high-dimensional data and can solve nonlinear problems through kernel functions. However, its computational cost is high on large-scale datasets (Abdullah, 2021).

Table 1: Performance Comparison on 6 Models between Group A and B.

MODEL	INDICATOR	GROUP A	GROUP B
KNN	Precision	0.7570	0.1512
	Recall	0.9654	0.4504
	F1-score	0.8486	0.2264
Logistic Regression	Precision	0.6827	0.2054
	Recall	0.6857	0.6937
	F1-score	0.6842	0.3170
SVM	Precision	0.7038	0.1949
	Recall	0.6994	0.6711
	F1-score	0.7016	0.3020
Naïve Bayes	Precision	0.6806	0.2128
	Recall	0.6943	0.6665
	F1-score	0.6874	0.3226
Decision Tree	Precision	0.8870	0.1823
	Recall	0.9990	0.2331
	F1-score	0.9397	0.2046
Random Forest	Precision	0.9790	0.4365
	Recall	0.9984	0.1291
	F1-score	0.9886	0.1993

2.2.4 Naïve Bayes

Based on Bayes theorem, Naïve Bayes is an algorithm of probabilistic classification. It assumes that the features are independent. Although this assumption is often not true, Naive Bayes still shows good classification results in practice. The advantage of the proposed algorithm is that it is computationally efficient and insensitive to noise, making it suitable for processing large-scale data sets (Saritas, 2019).

2.2.5 Decision Tree

The decision tree (DT) is another classification and regression algorithm that generates a DT by recursively selecting the best features to partition the data. Its advantage is that the model has good interpretability and can deal with multi-class problems and nonlinear data. However, DT is prone to overfitting and sensitive to data noise (Charbuty, 2021).

2.2.6 Random Forest

Random forest (RF) is an ensemble learning algorithm that builds multiple DTs and combines their predictions to improve the accuracy of classification or regression. It reduces the overfitting problem of single DT by introducing randomness and has strong adaptability to high-dimensional data. RF performs well in classification accuracy and stability (Speiser, 2019).

3 RESULTS AND DISCUSSION

After analysing 6 model performances across groups A and B, a notable decline in F1 scores is observed for all models, indicating a significant reduction in the ability of models to balance precision and recall when applied to group B, as shown in Table 1 below.

The reason why the performance of the six machine learning models deteriorates severely when moving from Group A (training and testing) to Group B (testing) can be attributed to several key factors. A major reason may be that there is a distribution shift

between the two groups; that is, the data distribution in Group B is significantly different from that in Group A. KNN depends on the distance between data points, and when the distribution changes, it is challenging to find suitable neighbouring points in Group B, causing its performance to slide. Similarly, logistic regression, as a linear model, is very sensitive to nonlinear relationships and complex category boundaries in Group B and cannot maintain high accuracy. The ability of SVM to find the optimal hyperplane is also frustrated when the boundary data points of Group B deviate from Group A. Naive Bayes assumes feature independence, and performance suffers when feature relationships in Group B are more complex than in Group A. Decision trees, due to their tendency to overfit, may have learned some specific patterns in Group A that do not generalize well to Group B, resulting in poor performance. Finally, while random forests enhance generalization by integrating multiple decision trees, they fail to fully meet this challenge in the face of significant distribution shifts.

To address these issues, an effective technique to improve model robustness in the presence of distribution shifts is domain adaptation distribution alignment. The focus of the method is to align the feature distribution between the source domain (group A) and the target domain (Group B). By minimizing the discrepancy between these distributions, models can generalize better to unseen data. Domain adaptation techniques can be particularly effective when the distribution differences are significant, as seen in this study. Using this approach, the model will be better to capture the underlying structure of both datasets, improving its ability to handle new, unseen data from Group B. As can be seen from the sharp drop in recall, the model struggles to identify the true positive class in the new group. Solving these problems may require regularization and data augmentation to improve the robustness of the model to distribution shifts.

4 CONCLUSIONS

This study compares the generalization performance of six machine learning models in predicting loan defaults. The results reveal that model performance, as measured by the F1 score, significantly declines when applied to unseen data due to distribution shifts between training and testing sets. Among the models, Random Forest showed the highest performance in the training set but experienced a sharp decline in unseen data, indicating overfitting. Techniques like domain adaptation and distribution alignment are suggested to address these issues. In conclusion, while machine learning models offer enhanced predictive capabilities over traditional methods, their generalization ability remains challenging. Future research should focus on improving model robustness, particularly in the face of distribution shifts, to ensure better risk management in financial settings.

REFERENCES

Abdullah, D. M. & Abdulazeez, A. M. 2021. Machine learning applications based on svm classification a review. Qubahan Academic Journal.

Athey, S. et al. 2018. The impact of machine learning on economics. The economics of artificial intelligence: An agenda, pp. 507–547.

Baesens, B., Setiono, R., Mues, C., & Vanthienen, J. 2003. Using neural network rule extraction and decision tables for credit-risk evaluation. Manag. Sci., vol. 49, pp. 312–329.

Charbuty, B. & Abdulazeez, A. M. 2021. Classification based on decision tree algorithm for machine learning. Journal of Applied Science and Technology Trends.

Chong, E., Han, C., & Park, F. C. 2017. Deep learning networks for stock market analysis and prediction: Methodology, data representations, and case studies. Expert Systems with Applications, vol. 83, pp. 187–205.

Fernandes, A. A. T., Filho, D. B. F., da Rocha, E. C., & Nascimento, W. 2020. Read this paper if you want to learn logistic regression. Revista de Sociologia e Política.

Jin, Y. & Zhu, Y. 2015. A data-driven approach to predict default risk of loan for online peer-to-peer (p2p) lending. In 2015 Fifth international conference on communication systems and network technologies, pp. 609–613, IEEE.

Lakshmanarao, A., Gupta, C., Koppireddy, C. S., Ramesh, U., & Dev, D. 2023. Loan default prediction using machine learning techniques and deep learning ann model. 2023 Annual International Conference on Emerging Research Areas: International Conference on Intelligent Systems (AICERA/ICIS), pp. 1–5, 2023.

Looney, A. & Yannelis, C. 2022. The consequences of student loan credit expansions: Evidence from three decades of default cycles. Journal of Financial Economics, vol. 143, no. 2, pp. 771–793.

Looney, A. & Yannelis, C. 2019. How useful are default rates? borrowers with large balances and student loan repayment. Economics of Education Review, vol. 71, pp. 135–145.

Saritas, M. M. & Yas¸ar, A. B. 2019. Performance analysis of ann and naive bayes classification algorithm for data classification. International Journal of Intelligent

Systems and Applications in Engineering, vol. 7, pp. 88–91.

Shih, J.-Y., Chen, W.-H., & Chang, Y.-J. 2014. Developing target marketing models for personal loans. In 2014 IEEE International Conference on Industrial Engineering and Engineering Management, pp. 1347–1351, IEEE.

Speiser, J. L., Miller, M. I., Tooze, J. A., & Ip, E. H.-S. 2019. A comparison of random forest variable selection methods for classification prediction modeling. Expert systems with applications, vol. 134, pp. 93–101.

Wu, M., Huang, Y., & Duan, J. 2019. Investigations on classification methods for loan application based on machine learning. In 2019 International Conference on Machine Learning and Cybernetics (ICMLC), pp. 1–6, IEEE.

Zareapoor, M., Shamsolmoali, P., et al. 2015. Application of credit card fraud detection: Based on bagging ensemble classifier. Procedia computer science, vol. 48, no. 2015, pp. 679–685.

Zhang, S. 2021. Challenges in knn classification. IEEE Transactions on Knowledge and Data Engineering, vol. 34, pp. 4663–4675.