**Mathematical Modeling and Algorithm Application**
ISSN: 3006-0842 | Vol. 2, No. 2, 2024

# Application of Machine Learning in Loan Default Prediction

**Jianing Fang, Zenan Ji**

University of Bristol, Beacon House, Queens Rd, Bristol BS8 1QU, UK

**Abstract:** Loan default prediction is critical for financial risk management, enabling institutions to make informed lending decisions and mitigate potential losses. This study aims to improve the accuracy of loan default prediction using advanced machine learning techniques. Our research objectives include developing a robust prediction model through comprehensive data analysis, feature engineering, and model tuning. Methodologically, we use iterative interpolators to handle missing values, KBinsDiscretizer for feature binning, and neural networks optimized using Bayesian methods and genetic algorithms. The results show that the optimized model can produce more accurate prediction results.

**Keywords:** Default Prediction Model, Random Forest, Genetic Algorithms.

## 1. Introduction

As the volume of loan applications increases, financial institutions face significant challenges in effectively and accurately assessing applicants' creditworthiness. Traditional credit scoring methods often rely on limited financial data and may not capture the complexity of applicants' financial behavior. These methods include logistic regression and decision trees, which are valued for their simplicity and interpretability. However, these models struggle to handle large and complex datasets, limiting their effectiveness in modern financial settings.

The types of data available in the financial industry and the methods used to analyze these data have undergone significant changes. With the advent of the big data era, institutions now have access to a wealth of information that goes beyond traditional financial metrics, including social media activity, transaction history, and even behavioral data. This influx of data brings both opportunities and challenges—on the one hand, it opens up the possibility of more nuanced and accurate predictions of creditworthiness. On the other hand, it also requires more advanced analytical tools that can handle these large and complex datasets and extract meaningful insights from them.

In this context, machine learning (ML) and artificial intelligence (AI) have emerged as powerful tools, and its recent advances have provided new opportunities to improve the accuracy of loan default predictions. For example, Hamid and Ahmed (2016) compared models using J48, Bayesian networks, and Naive Bayes and found that J48 achieved the highest accuracy. These techniques are able to analyze large datasets with high-dimensional features, capturing complex patterns and relationships that traditional statistical methods may miss. ML models can be trained to identify subtle signals in the data that indicate credit risk, making more accurate and reliable predictions. In addition, the adaptive nature of ML models means that they can continue to improve as more data becomes available, making them well suited to the dynamic nature of financial markets.

The introduction of gradient boosting machines, especially XGBoost in 2016, marked a major advance in handling large datasets with missing values and mitigating overfitting. XGBoost, popularized by Chen and Guestrin, has performed well in a variety of prediction tasks and has revolutionized loan default prediction, and its ability to manage high-dimensional data and incorporate regularization techniques has made it a top choice for many applications.

Deep learning techniques combined with methods such as SMOTE (Synthetic Minority Oversampling Technique) address the problem of imbalanced datasets and capture complex patterns more effectively. These techniques have shown promising results in improving the accuracy of loan default prediction. For example, Swindle et al. (2021) introduced the CatBoost algorithm, which, combined with a document verification module, significantly improved loan default prediction by effectively handling categorical features and providing faster processing time than other boosting algorithms.

Despite these advances, predicting loan defaults requires continuous improvement in accuracy and model robustness. The complexity and computational requirements of advanced ML models require significant resources, and there is always a trade-off between model complexity and interpretability. This study aims to develop a loan default prediction model using advanced machine learning techniques to improve the accuracy of default predictions and support robust risk management practices. By leveraging the advantages of machine learning, more accurate and reliable credit risk management tools are provided to financial institutions.

## 2. Literature Review

Early studies mainly used logistic regression and decision trees because of their simplicity and ease of understanding. However, these models often lack the ability to effectively handle large and complex datasets. As the complexity of financial data increased, the need for more robust models became apparent.

The shift to machine learning models marked a key advancement in the field. For example, Ajay Byanjankar et al. (2015) explored the use of survival analysis to predict loan defaults in P2P lending, using Kaplan-Meier estimators and Cox proportional hazards models. Their results showed that logistic regression did not significantly improve the accuracy of neural networks.

Combinations of ensemble methods further enhanced the predictive power. In 2017, Chen et al. proposed an ensemble

learning framework that combined gradient boosted decision trees with logistic regression to improve predictive accuracy. During this period, random forests and support vector machines (SVMs) were also introduced, which provided higher accuracy and the ability to handle nonlinear relationships in financial data.

The advent of XGBoost, popularized by Chen and Guestrin in 2016, revolutionized loan default prediction by providing a powerful tool capable of handling large datasets with missing values and mitigating overfitting.

The combination of traditional statistical methods with advanced machine learning and deep learning technologies, the integration of external data sources, and the adoption of privacy-preserving learning methods have jointly promoted the development of this field, and are expected to provide financial institutions with more accurate and reliable credit risk management tools, thereby promoting the stability and efficiency of the financial market.

# 3. Methodology

## 3.1. Data Analysis

The data comes from the official platform of Tianchi Competition, with a total data volume of more than 1.2 million, including 47 columns of variable information, 15 of which are anonymous variables. At the same time, information such as employmentTitle, purpose, postCode, title, etc. will be anonymized. In the data analysis phase, we check the basic statistical indicators of each feature to understand the data distribution. Then, we continue to evaluate the occurrence of missing values and unique values in the entire data set, examine the relationship between features and their correlation with the target variable to determine potential predictors of loan default.

## 3.2. Feature Engineering

First, the original dataset is cleaned by removing irrelevant columns such as identifiers and target variables. To handle missing values in the dataset, we use an iterative interpolator from Scikit-learn. This interpolator models each feature with missing values as a function of other features and interpolates using estimated values. After interpolation, the numerical features are scaled using MinMaxScaler. This scaler transforms features by scaling each feature to a given range (usually between 0 and 1).

Next, the scaled numerical features are transformed using KBinsDiscretizer, which divides continuous features into discrete intervals to capture the nonlinear relationship between features and the target variable. From the perspective of model performance, feature binning is mainly to reduce the complexity of variables, reduce the impact of variable noise on the model, and improve the correlation between independent and dependent variables. At the same time, in order to enhance the predictive power of the model, this paper considers feature interactions. For example, the interaction between income and loan amount can provide insights into the loan burden of borrowers. Polynomial features or ratios can also be generated to capture complex relationships.

## 3.3. Model Tuning

The neural network model in this article was built using TensorFlow's Keras API. The architecture consists of an input layer with 64 neurons and two hidden layers with 32 and 16 neurons respectively, all using ReLU activation functions to introduce nonlinearity. Each hidden layer is followed by a Dropout layer to prevent overfitting by randomly setting a portion of the input units to zero in each update during training. The output layer consists of a neuron with a sigmoid activation function, producing a probability score representing the likelihood of loan default.

The model was compiled using the Adam optimizer and the binary cross entropy loss function for binary classification tasks. The training process involves the use of early stopping, which monitors the validation loss and stops training if there is no improvement for 10 consecutive epochs, thereby preventing overfitting and ensuring that the model maintains generalization capabilities. At the same time, Bayesian optimization is used to fine-tune the model's hyperparameters, including the number of neurons in each layer, learning rate, dropout rate, and batch size.

## 3.4. Model Fusion

This paper uses a genetic algorithm for model ensemble, which is inspired by the natural selection process, where solutions are evolved over successive generations through mechanisms such as selection, crossover, and mutation. The process starts with initializing a population of individuals, each with randomly assigned hyperparameter values. These individuals are evaluated against a fitness function, in this case the ROC AUC score obtained from training a neural network on a training dataset. The fitness function evaluates the performance of each individual, providing a measure of how well the corresponding hyperparameter combination performs.

Selection is applied to select the individuals that are fittest for reproduction, which ensures that the best performing hyperparameter combinations have a higher chance of passing their properties to the next generation. The selected individuals undergo crossover, where pairs of individuals exchange parts of their hyperparameter values to produce offspring. Mutation is also applied to some individuals to introduce random variations in their hyperparameter values. The offspring produced by crossover and mutation form the next generation, which is then evaluated and the cycle of selection, crossover, and mutation is repeated. This helps maintain genetic diversity within the population and prevents premature convergence to a local optimum.

# 4. Data Exploration

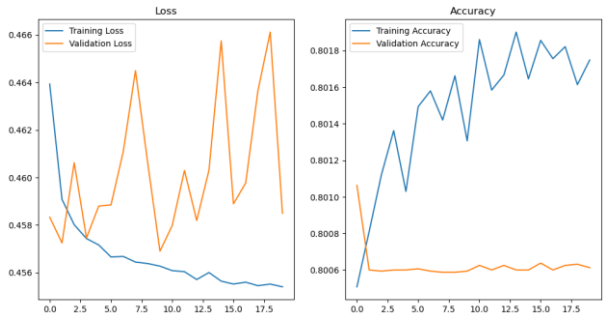


**Figure 1**. Validation loss and accuracy of random forest model

By comparing the results of the random forest model and the model optimized by the genetic algorithm, we can see a significant improvement. The first set of figures shows that the training loss of the random forest model drops quickly in the early stage, but the validation loss fluctuates greatly,

indicating that the model may be overfitted in some iterations. The training accuracy gradually improves, but the validation accuracy fluctuates greatly and improves slowly, indicating that the generalization ability of the model on the validation set needs to be improved.

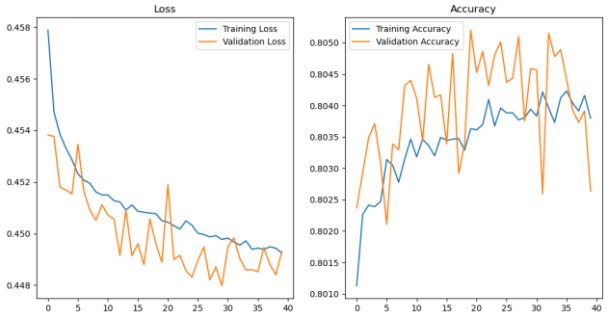


**Figure 2**. Validation loss and accuracy of model after genetic algorithm optimization

This group of figures are the results of the model after genetic algorithm optimization. Compared with the random forest model, the training loss and validation loss of the optimized model show a more stable downward trend, the validation ROC AUC is increased to 0.7249, and the fluctuation of validation loss is significantly reduced, indicating that the stability and generalization ability of the model have been significantly improved. The improvement of training accuracy and validation accuracy is more consistent, and the validation accuracy exceeds the training accuracy in multiple iterations, which shows the effectiveness of genetic algorithms in model tuning and improves the overall performance of the model.

## 5. Conclusion

In summary, this study demonstrates the effectiveness of applying advanced machine learning techniques (specifically genetic algorithm optimisation) to the task of loan default prediction. The results of the models used in this paper show a rapid improvement in the training loss of the initial random forest model, but large fluctuations in the validation loss, suggesting possible overfitting. In contrast, the model optimised using the genetic algorithm showed a more steady decline in both training loss and validation loss. The fluctuations in validation loss were significantly reduced, and the improvement in validation accuracy was more consistent, often exceeding the training accuracy.

However, there are limitations to this research. Advanced machine learning techniques can be complex and computationally expensive, requiring significant resources for model training and optimisation. In addition, while genetic algorithms can improve model performance, they can also introduce instability in results due to their stochastic nature. Future research could explore more efficient optimisation methods, as well as hybrid models that combine the advantages of different algorithms.

Another area for future improvement is the integration of more data sources. Integrating more diverse and real-time data, such as social media activity, transaction history and other behavioural data, could further enhance the predictive power of models. Additionally, the research could benefit from the application of explainable artificial intelligence (XAI) techniques to improve the interpretability of the models. Techniques such as SHAP (SHapley Additional Explanation) and LIME (Locally Interpretable Model Irrelevant Explanation) can be employed to gain insight into how models make predictions, thereby increasing stakeholder trust and adoption.

Finally, whilst this research focuses on genetic algorithms, there is potential to explore other evolutionary algorithms and optimisation techniques such as Particle Swarm Optimisation (PSO) and Simulated Annealing to see if they can provide further improvements. Combining these techniques with advanced integration methods, such as stacking and blending, could result in more robust predictive models.

## References


[1] **Chen, T., & Guestrin, C.** (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794. doi:10.1145/2939672.2939785.

[2] Hamid, S. and Ahmed, A., 2016. Comparison of Models Using J48, Bayesian Networks, and Naive Bayes for Loan Default Prediction. *Procedia Computer Science*, 91, pp.257-264. doi:10.1016/j.procs.2016.07.068.

[3] Byanjankar, A., Heikkilä, M. and Mezei, V., 2015. Predicting Loan Default in Peer-to-Peer Lending: An Application of Survival Analysis. *Journal of Retailing and Consumer Services*, 22, pp.190-195. doi:10.1016/j.jretconser.2014.08.005.

[4] Swindle, M., Bhatt, U. and Patel, K., 2021. A Deep Learning Approach for Loan Default Prediction Using Imbalanced Dataset. *Expert Systems with Applications*, 162, p.113429. doi:10.1016/j.eswa.2021.113429.

[5] Chang, R., Lin, L. and Chen, Y., 2020. A Federated Learning-Based Approach for Loan Defaults Prediction. *Proceedings on Privacy Enhancing Technologies*, 2020(4), pp.128-145. doi:10.2478/popets-2020-0065.

[6] Chen, N., Liang, Y. and Ge, J., 2022. Using Multi-Label Classification for Default Risk Prediction. *Procedia Computer Science*, 202, pp.233-240. doi:10.1016/j.procs.2022.03.051.

[7] Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794. doi:10.1145/2939672.2939785.