German Credit Risk Prediction Using Machine Learning Models

Rongfei Ma

Department of English, Nankai University, Tianjin, 300071, China
2310043@mail.nankai.edu.cn

Abstract. Management of credit risk plays a vital role in the financial industry, allowing institutions to mitigate losses, optimize capital allocation, and make informed decisions. This study investigates the predictive efficacy of five machine learning algorithms (Decision Trees, Logistic Regression, Random Forest, k-Nearest Neighbor (KNN), and Support Vector Machine (SVM) and three ensemble methods (voting, gradient boosting and stacking) in a German credit dataset. The results show that models have a better performance when data preprocessing is refined. For example, the accuracy of KNN is increased from 0.69 to 0.74. Besides, ensemble models outperform the best performance of the single algorithm. For example, the best-performing Xgboost reaches a higher F1 score (0.61) compared with Random Forest (0.6). However, to reach better performance, handling data imbalance and redundant noise should be taken into consideration. In general, by systematically comparing the boundary conditions of a single model and an integrated framework, this paper verifies the important role of data preprocessing and ensemble methods in credit risk assessment and provides a reproducible benchmark process for the construction of lightweight risk control systems.

Keywords: German Credit Risk Prediction, Machine Learning Models, ensemble methods

1 Introduction

In many domains, including credit risk management, machine learning (ML) is gaining more and more attention. Traditional approaches, such as manual audits and statistical methods (e.g., linear regression), used to be popular tools in credit risk prediction. However, limited by linear assumptions and the ability to process structured data, it is difficult for traditional approaches to capture the complex nonlinear risks caused by sudden policy changes and black swan events in the financial market. Its static modeling method is even less capable of adapting to the rapid evolution of the market structure. The advent of machine learning (ML) has introduced transformative potential, expanding the coverage of risk factors by

Y. Sun (ed.), *Proceedings of the 2025 3rd International Conference on Image, Algorithms, and Artificial Intelligence (ICIAAI 2025)*, Advances in Computer Science Research 122,
https://doi.org/10.2991/978-94-6463-823-3_27

integrating multimodal data, while its generalization ability is particularly prominent when handling high-dimensional and sparse data. Existing studies have extensively validated individual ML algorithms as well as ensemble models.

Barboza et al. compared the bankruptcy prediction performance of ML methods (SVM, bagging, boosting, random forests, and neural networks) against statistical models like discriminant analysis and logistic regression [1]. By using data from North American firms (1985–2013) sourced from the Salomon Center and Compustat databases, the study analyzed over 10,000 firm-year observations. Besides the original Altman Z-score, six extra financial indicators were used to enhance model accuracy. The key finding was that ML algorithms are roughly 10% more efficient than traditional models. Similarly, Le and Viviani compared statistical approaches with three ML methods (neural networks, SVM, and KNN) by predicting a five-year dataset of US banks [2]. It contained 3,000 records with 31 financial ratios as predictors, among which 1,438 were defaulted and 1,562 were from active banks. The results indicated that neural networks and KNN significantly outperformed statistical models, while SVM did not show superiority. Moscatelli et al. explored the predictive performance of tree algorithms versus traditional classifiers (discriminant analysis, logistic regression, etc) [3]. They concluded that ML models can provide more precise predictions based on their analysis of around 300,000 observations of financial and credit behavior indicators for non-financial firms in Italy (2011–2017).

By performing a credit risk analysis based on German credit data, this research aims to compare the performance across several machine learning algorithms in the classification problem and to demonstrate the importance of data preprocessing and ensemble methods in elevating the model's performance.

2 Methodology

2.1 Data Resource

The German credit risk dataset was originally compiled by Professor Hans Hofmann in 1994 and is now publicly available through the UCI Machine Learning Repository. It contains 1,000 loan application records with 20 features and 1 binary target variable (credit risk: good/bad). The data is derived from anonymized credit archives of a German commercial bank, stripped of personally identifiable information (PII) to comply with GDPR and other privacy regulations.

2.2 Data Preprocessing

First, different preprocessing methods are applied to numerical and categorical features. For numerical features, missing values are filled with the median to avoid

deviations caused by extreme values affecting the mean. Then, StandardScaler is used to give data a mean of zero and a standard deviation of one. This improves model convergence speed and performance, especially in distance-based algorithms (e.g., KNN) and gradient-descent-based algorithms (e.g., logistic regression, SVM). For categorical features, missing values are filled with the most frequent value, and OneHotEncoder is used for one-hot encoding to convert them into dummy variables. This prevents the model from misinterpreting the magnitude of category labels.

Second, the label processing is optimized. By deducting the mean and dividing by the standard deviation, the "credit amount" is standardized to make the data distribution more symmetric and stable. Then, samples are classified into high and low-risk categories based on whether the standardized value is greater than the mean, ensuring objective and reasonable label classification.

2.3 Exploratory Data Analysis

The correlation matrix was displayed as a heatmap to demonstrate the intercorrelation among the features of the bank account users who obtained loans in Germany. The stronger the positive correlation between two features, the lighter the color block that intersects them (Fig. 1).

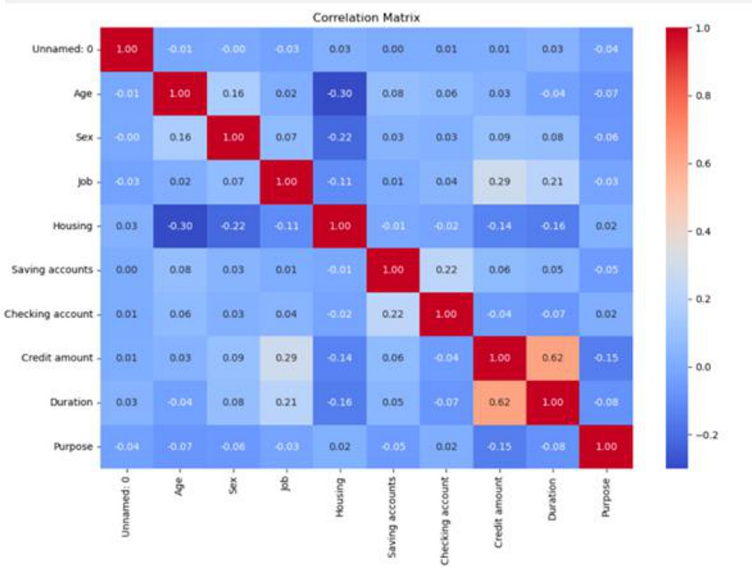

Fig. 1. The structure of the autoencoder (Picture credit: Original).

From the correlation matrix, it can be observed that there are significant

differences in the correlations among different features. Among them, the correlation between Credit amount and Duration is the highest, reaching 0.62. This strong positive correlation between loan amount and loan term suggests that when the loan amount is larger, the loan term tends to be longer, and vice versa. While the correlation between Age and Housing is the lowest, at -0.30, suggesting a certain negative correlation between age and whether there is a housing loan. The correlations among other variables are relatively low, indicating that these features are to some extent independent.

2.4 Machine Learning Algorithms

Logistic Regression. Logistic regression is an efficient algorithm that is fast and easy to train. First, a value of a weighted sum of the input features plus a bias term would be calculated. Then, the value would pass through a sigmoid function (i.e., S-shaped) and output a number between 0 and 1. The instance would respectively categorized as positive class (labeled “1”) and negative class (labeled “0”) based on whether the estimated probability is greater than 50%, which makes it a binary classifier.

Decision Tree. Decision Trees are naturally explainable and require very little data preparation. It functions based on its tree structure, traversing branches downward step by step from the root node that adheres to the feature values and predefined judgment conditions of the node, such as comparing feature values with thresholds. Each node judgment acts as a filter for data features, continuing this process until a leaf node is reached. In classification tasks, the leaf node’s category label directly becomes the final predicted category.

Random Forest. Lantz's research demonstrates that the Random Forest algorithm leverages an ensemble of decision trees to achieve optimal performance [4]. During the training, it strategically selects the splitting point and variable that is most relevant for each tree and node to minimize errors, enabling its ability to maintain analytical efficacy across high-dimensional datasets. The method uses only a small random part of the complete set of observations, and can handle large data sets. By integrating multiple decision trees through bagging and feature randomness, the methodology exhibits exceptional versatility in handling large-scale data while preserving model accuracy.

KNN. The principle of KNN is to predict the category of a new value x based on the category of the K nearest points to it. To implement KNN, the selection of the K value

and the calculation of the distance between points are important. When choosing the K point, a relatively large critical K point whose error rate will rise no matter whether it increases or decreases can be selected. Usually, Euclidean distance is used to calculate the distance.

SVM. Support Vector Machine (SVM) algorithm identifies an optimal hyperplane within the feature space that maximizes the margin between different classes, ensuring samples from distinct classes are separated as clearly as possible. For non-linearly separable data, SVM employs kernel functions to map data into higher-dimensional spaces, transforming non-linear relationships into linear ones. In datasets with fewer than 1000 samples, Support Vector Machines (SVM) with RBF kernels consistently outperformed neural networks (MLP) and other classifiers in terms of both accuracy and stability [5].

Voting Classifier. Voting creates a classifier with better performance in a simple way. It aggregates the predictions of each single algorithm and predicts the class that gets the most votes. This type of majority-vote classifier is called a hard voting classifier (Fig. 2).

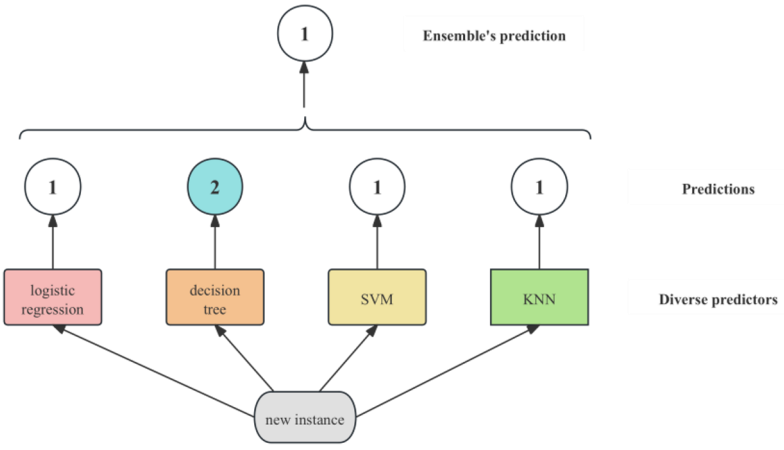

Fig. 2. Voting algorithm (Picture credit: Original).

This voting classifier often achieves a higher accuracy than the best classifier in the ensemble. In fact, the ensemble still can be a strong learner even if each classifier is a weak learner, as long as it is provided sufficient number and diversity of weak learners.

Stacking Classifier. The core idea of stacking is to aggregate the prediction results of multiple base learners by training a meta-learner, rather than using simple methods like Voting. Each of the base predictors predicts a different value, and then the final predictor (called a blender, or a meta learner) takes these predictions as inputs and makes the final prediction, and the number of features is equal to the number of base classifiers [6].

Gradient Boosting. Gradient Boosting works by sequentially adding predictors to an ensemble, each one correcting its predecessor. It enhances predictive accuracy through the bagging technique while demonstrating exceptional performance in handling skewed datasets, rendering it particularly suitable for developing credit risk assessment models in the banking sector. It can handle outliers and missing values that are common in real-world datasets [7].

2.5 Evaluation Metrics

In credit risk assessment, the performance evaluation of classification models requires a comprehensive analysis using multiple metrics. The confusion matrix serves as a fundamental tool, visually reflecting the classification performance by statistically cross-distributing true and predicted classes (TP, FP, TN, FN).

Accuracy, as an overall performance metric, calculates the proportion of correctly classified samples, but it may overestimate model performance in imbalanced scenarios (e.g., low default rates). Precision evaluates the proportion of true positive samples among all predicted positives, avoiding false positives. While recall reflects how many actual positives are correctly identified, it minimizes false negatives. The F1-score is a harmonic mean of precision and recall, which balances these two metrics for scenarios requiring both error control, such as credit scoring models.

3 Results and Discussion

3.1 Model Results Analysis

As shown in Fig. 3, the ratio between low-risk samples and high-risk samples of German credit risk data is 65% to 35%, which means it is a slightly skewed dataset. As a result, besides accuracy, the F1 score should be the main measure to evaluate models' performance.

Fig. 3. Credit Risk Distribution (Picture credit: Original).

As shown in Table 1, Random Forest and Xgboost show the best performance in accuracy (0.78) and F1 score (0.61). Random Forest does well in dealing with high-dimensional data and non-linear relationships because it reduces overfitting by integrating multiple decision trees. On the other hand, Xgboost incrementally optimizes the model through gradient boosting, excelling in imbalanced data scenarios. Logistic Regression and SVM achieve comparable accuracy (0.77) to Random Forest. But low F1 scores (0.56) reveal their insufficient recall in class-imbalanced settings, which may lead to missing high-risk clients.

Among ensemble methods, Stacking Classifier matched Xgboost's accuracy (0.78) but showed a slightly lower F1 score (0.59), suggesting that its meta-classifier may not fully leverage the predictive strengths of base classifiers. Voting Classifier only delivered moderate performance (F1=0.57), indicating that a simple majority voting strategy offers limited improvement.

KNN achieved relatively high accuracy (0.74) but the lowest recall (0.47), reflecting its weakness in identifying minority classes (high-risk clients). Decision Trees perform badly due to overfitting (F1=0.53) and necessitate pruning or depth constraints to enhance generalization. Although SVM maintained stable performance with small sample sizes (accuracy=0.77), its RBF kernel exhibited poor generalization (recall=0.48), shown by its inability to capture complex features of high-risk clients.

Table 1. Model Evaluation

Algorithm	Accuracy	Precision	Recall	F1 score
Logistic Regression	0.77	0.67	0.48	0.56
Decision Tree	0.7	0.52	0.55	0.53
Random Forest	0.78	0.67	0.54	0.6
KNN	0.74	0.6	0.47	0.52
SVM	0.77	0.67	0.48	0.56
Voting	0.76	0.63	0.52	0.57
Stacking	0.78	0.7	0.51	0.59
Xgboost	0.78	0.68	0.55	0.61

3.2 Limitations of the Study

First, the German Credit dataset contains only 1,000 historical samples, which is limited in size and rather outdated. Additionally, imbalanced data distribution (65% low-risk vs. 35% high-risk samples) is likely to weaken high-risk client identification. Second, variations in algorithm sensitivity (e.g., KNN relies heavily on standardization) were not exclusively analyzed. The Stacking classifier underperformed XGBoost in F1-score, and the Voting classifier achieved only moderate performance, indicating limitations in effectively leveraging base learners' strengths. Last, the feature engineering lacks sufficient exploration of advanced feature selection, nonlinear relationships, and dynamic variables (e.g., repayment trends). The models would learn to make more accurate predictions and better measure metrics if more features were included.

3.3 Model Improvement Suggestions

To address the challenge of imbalanced datasets, SMOTE could be implemented to help balance the data and improve model performance, particularly in terms of recall [9]. SMOTE addresses class imbalance by generating artificial samples for the underrepresented class, which improves the model's capacity to detect and predict these instances more effectively. The additional examples of the underrepresented class can lead to a more balanced class distribution and a more accurate classification model.

Adopting hyperparameter tuning techniques can enhance the predictive performance of both single and ensemble algorithms. For instance, adjusting the tree depth and number of trees of the Random Forest model helps reduce overfitting. For SVM, optimizing kernel functions and regularization parameters enhances the generalization capability. While for XGBoost models, combining hyperparameters, gridsearch, and optuna can improve their accuracy and the ability to handle imbalanced data [9].

3.4 Future Research Direction

Future research could consider incorporating explainable AI (XAI) techniques. For example, Chang et al. applied SHAP values to credit risk models to measure the interaction effects of key features like “credit amount” and “loan duration” [10]. Building on this, Chang's study expanded the dataset by including not only more historical data but also factors like industry trends and macroeconomic indicators. Consequently, the model has better performance in capturing complex market dynamics and customer-behavior patterns. Its prediction accuracy increases by 15%. Moreover, developing lightweight ensemble methods, such as Stacking with a meta-learner, can enhance risk prediction as well.

4 Conclusion

This paper systematically explored the predictive capabilities of machine learning models and ensemble methods in credit risk assessment using the German credit dataset. By evaluating five individual algorithms and three ensemble methods, the research underscores the critical role of data preprocessing and model integration in enhancing classification performance. Key findings reveal that refined data preprocessing, including feature scaling and categorical encoding, significantly improved model accuracy, as exemplified 7% increase in accuracy of KNN. Furthermore, ensemble methods such as XGBoost demonstrated superior performance (F1-score: 0.61) compared to standalone models, highlighting their efficacy in handling imbalanced datasets and complex feature relationships.

The practical implications of this work extend to financial institutions seeking lightweight, scalable risk management frameworks. Ensemble methods (particularly gradient boosting) offer a solution for identifying high-risk clients while balancing precision and recall, which is a critical requirement in credit scoring systems. However, limitations such as the dataset’s small size, class imbalance, and outdated nature constrain the generalization ability of results. Additionally, the underperformance of certain ensemble strategies (e.g., Voting Classifier) suggests opportunities for optimizing meta-learner architectures.

Future research could prioritize expanding datasets to include dynamic variables (e.g., macroeconomic indicators) and adopting advanced techniques such as SMOTE for imbalance mitigation and explainable AI (XAI) for model transparency. Combining lightweight ensemble methods with hyperparameter optimization could further enhance predictive accuracy and operational efficiency. Ultimately, this study provides a reproducible benchmark for developing adaptive credit risk systems, bridging the gap between theoretical innovation and practical financial applications.

References

1. Barboza, F., Kimura, H., Altman, E.: Machine learning models and bankruptcy prediction. Expert Systems with Applications 83, 405-417 (2017)
2. Le, H.H., Viviani, J.-L.: Predicting bank failure: An improvement by implementing a machine learning approach to classical financial ratios. Research in International Business and Finance 44, 16-25 (2018)
3. Moscatelli, M., Narizzano, S., Parlapiano, F., Viggiano, G., et al.: Corporate default forecasting with machine learning. Bank of Italy, Economic Research and International Relations Area Technical Report (2019)
4. Lantz, B.: Machine learning with R. Packt Publishing Ltd, Birmingham (2013)
5. Fernández-Delgado, M., Cernadas, E., Barro, S., Amorim, D.: Do we need hundreds of classifiers to solve real-world classification problems? Journal of Machine Learning Research 15(1), 3133-3181 (2014)
6. Li, Y., Zhao, R., Sha, M.: A Hybrid Credit Risk Evaluation Model Based on Three-Way Decisions and Stacking Ensemble Approach. Computational Economics (2024)
7. Natekin, A., Knoll, A.: Gradient boosting machines, a tutorial. Frontiers in Neurorobotics 7, 21 (2013)
8. Ruchita, M., Bhargavi, M., Rakshita, M., Nandini, B.C., Aziz, I., Gopi, J.: Leveraging SMOTE and Random Forest for Improved Credit Card Fraud Detection. In: 2024 International Conference on Sustainable Communication Networks and Application (ICSCNA), pp. 795-800. IEEE (2024)
9. Sumaya, S.S., Ibraheem, N., Sarab, M.H.: Credit Card Fraud Detection Using Improved Deep Learning Models. CMC-Computers Materials & Continua 78(1), 1050-1069 (2024)
10. Chang, V., Xu, Q.A., Akinloye, S.H., et al.: Prediction of bank credit worthiness through credit risk analysis: an explainable machine learning study. Annals of Operations Research (2024)