Dean&Francis
ISSN 2959-6130

Credit Card Default Prediction: a Systematic Machine Learning Pipeline and Model Comparison

Zhenning Zhang

Detroit Green Technology Institute, Hubei University of Technology, Wuhan, China
Corresponding author:zhenning.zhang66@outlook.com

Abstract:

Predictions of credit card defaults are inextricably linked with the risk management functions of financial institutions. Traditional statistical science falters with these kinds of problems involving complicated nonlinear structures, while machine learning studies have not performed systematic comparisons among several algorithms backed up by business interpretations. This study created a complete machine-learning pipeline around the University of California, Irvine (UCI) credit card dataset, comparing five models: logistic regression, decision trees, random forests, gradient boosting machines, and multi-layer perceptrons. Hyperparameter tuning was carried out using grid search, while the Area Under the ROC Curve (AUC) was considered the main evaluation metric. The results indicate logistic regression could discriminate best (AUC on test: 0.764), outperforming more complex ensemble and deep learning models. Permutation importance has identified the bill amount four months prior and the credit limit as the most important predictors and shown the greater value of dynamic financial behavior data against static demographic data. The study provides a high-performing model framework that is interpretable and reproducible for the financial industry, challenging the argument that complexity equals performance and showing the practical worth of simple models in certain scenarios.

Keywords: Credit card default prediction; Machine learning; Model comparison; Interpretability; AUC

1. Introduction

The rapid growth of credit card services necessitates advanced risk management, where accurate default prediction is critical to mitigate financial losses and systemic risks [1]. To this end, empirical analyses comparing statistical and machine learning methods are vital for enhancing predictive performance [2]. Recent comparative studies continue to demonstrate the effectiveness of machine learning and deep learn-

ISSN 2959-6130

ing-based models in credit scoring tasks.

This study builds upon foundational work. The benchmark University of California, Irvine (UCI) credit card default dataset, established by Yeh and Lien, enables comparative evaluation of predictive techniques [3]. Expansive benchmarking studies, such as that by Lessmann et al., have systematically compared state-of-the-art classification algorithms for credit scoring [4]. Subsequent research has addressed specific challenges like class imbalance in datasets, investigating methods to improve prediction accuracy [5]. The integration of diverse data sources is recognized as a promising direction for obtaining deeper insights into borrower behavior [6]. Continuing this empirical tradition, comparative studies of ensemble learning techniques remain central to advancing credit risk management. Simultaneously, the pursuit of model interpretability, including through advanced techniques like Explainable AI (XAI), is a key focus of contemporary research in credit risk analysis [7]. Underpinning these efforts is the broader paradigm of employing interpretable frameworks for financial risk analysis [8]. Furthermore, recent comparative studies of machine learning models for credit card default prediction continue to highlight the importance of model interpretability [9]. Meanwhile, the application of interpretable machine learning techniques in financial risk management has gained significant attention [10].To address the gap in systematic comparisons that prioritize both performance and business interpretability, this study conducts a comprehensive empirical evaluation of five machine learning paradigms using a complete, reproducible pipeline. The primary contributions are twofold:First,it is demonstrated that a simple, interpretable model (logistic regression) can achieve superior discriminative power on the UCI benchmark dataset, challenging the assumption that model complexity necessarily leads to better performance.Second,beyond mere performance metrics, actionable risk insights are provided through permutation importance analysis, identifying historical bill amounts and credit limit as the most predictive factors, thereby offering a practical and interpretable framework for financial institutions.

2. Methodology and Experimental Design

This chapter describes the global methodology of the study, beginning with the data sources and going through the preprocessing and the feature engineering pipelines, the machine learning algorithms, model training and tuning procedures, and evaluation metrics. To ensure that the experiments were repeatable, all processes were placed into a reproducible code pipeline that guaranteed their rigor and the reliability of the results.

2.1 Data Sources and Preprocessing

The study uses the credit card default data set in UCI, a classical academic benchmark, that consists of 30,000 credit card clients from Taiwan with 23 variables divided into four main groups: demographic attributes (gender, education level, marital status, age), credit attributes (credit limit) and behavioral history attributes (payment status for the last six months, bill statement amounts for the last six months, and past payment amounts for the last six months). The dependent feature “default next month” is a binary variable that shows if the client defaulted in the ensuing month (1=default,0=non-default). Anomalously coded education and marital status in the original data set were merged into another category to reduce sparsity-related impacts on model stability.

Therefore, a systematic data-preprocessing workflow was set in place in order to avoid data leakage and ensure consistency throughout the analysis pipeline. For numerical features (for example, credit limit, monthly bill amounts), missing values were imputed with median values and standardized to a mean of zero-based and unit-variance distribution. For categorical features (for example, gender, repayment status), missing values were filled with the most frequent category, then converted into binary indicator variables using one-hot encoding. The data was split into training and testing data with a 70:30 ratio using stratified sampling that kept the same ratio of default to non-default observations as in the original data.

2.2 Model Selection and Training Framework

Five classical algorithms from different paradigms of machine learning were selected for extensive comparative analysis. Logistic regression serves as a linear model baseline with strong interpretability. Decision trees are non-linear, rule-based, interpretable models that are subject to overfitting. Random forests are bagging-based ensemble learning methods that build multiple decision trees and aggregate their results to lower the variance and improve generalization. Boosting-based ensemble methods, especially GradientBoostingClassifier, implement iterative correction of errors from previous models in achieving higher accuracy. Multi-layer perceptrons are simple feed-forward neural networks that were used to exploit the capabilities of deep learning for this given task.

Class imbalance was treated in that defaulting costs made up just 22% of all observations by utilizing class_weight = ‘balanced’ during training of the models for logistic regression, decision trees, and random forests. This thereby gives priority to minimizing the misclassification of the minority class. All models were tied up with the above-mentioned preprocessing pipeline to form an end-to-end “preprocessing-modeling” entity. Each model’s hyperparameter tuning was performed utilizing grid search with stratified 3-fold cross-validation. Two grids were set up for “debug” and “final” scenarios, balancing computational efficiency while not restricting search breadth; thus, for instance, C was tuned for logistic regression while n_

estimators and max_depth were tuned for tree-based models.

Assessment of model performance was done on a fully independent test set calculated by complementary metrics. Primary in assessing model performance was the AUC, measuring how well the model discriminated between defaults and non-defaults across thresholds. Precision, recall, and F1-score were provided to carefully assess the model's performance on the tested default class (positive class) of interest to us, particularly recall (the power to catch true defaulters). Confusion matrices contrasted the predicted and actual labels visually.

Importance analysis for features was determined using the built-in. feature_importances_ property of tree models (decision trees, random forests, and gradient boosting). For non-tree algorithms, such as logistic regression, permutation importance is used. In the methodology, it randomly permuted the values of a given feature in the test set and measured the drop in model performance (AUC)-a higher drop corresponds to a higher importance of that feature.

3. Experimental Results, Analysis, and Discussions

This chapter presents and discusses model testing results in the order specified by the experimental outline in Chapter 1. The analysis moves from a performance-centered model comparison through to business interpretation of the best model, together with its systematic identification of key risk drivers, to address the research questions set up in the study.

3.1 Comparison and Analysis of Model Performances

After parameter tuning and test set evaluation, model performance rankings are shown in Table 1. A notable finding is that the simplest model, logistic regression, achieved the highest test AUC (0.764), slightly outperforming the powerful gradient boosting machine (0.762) and random forest (0.748). The decision tree, as a single-tree model, performed significantly worse than its ensemble counterparts (AUC = 0.722), confirming its theoretical flaws of high variance and overfitting susceptibility. The multi-layer perceptron underperformed expectations (AUC = 0.650), and a "Convergence Warning" during training suggests it requires more complex structures, additional iterations, or finer learning rate tuning.

Table 1. Comparison of Performances of Various Models on the Testing Set.

Model	Testing AUC	Testing F1 Score	Testing Precision	Testing Recall
Logistic Regression	0.764	0.525	0.477	0.586
Gradient Boosting Machine	0.762	0.447	0.671	0.336
Random Forest	0.748	0.509	0.487	0.534
Decision Tree	0.722	0.508	0.502	0.514
MLP	0.650	0.379	0.382	0.377

The logistic regression model's superior discriminative ability is visually confirmed by its Receiver Operating Characteristic (ROC) curve (Fig. 1), with an AUC of 0.764-well above the random classifier line (AUC = 0.5).

ISSN 2959-6130

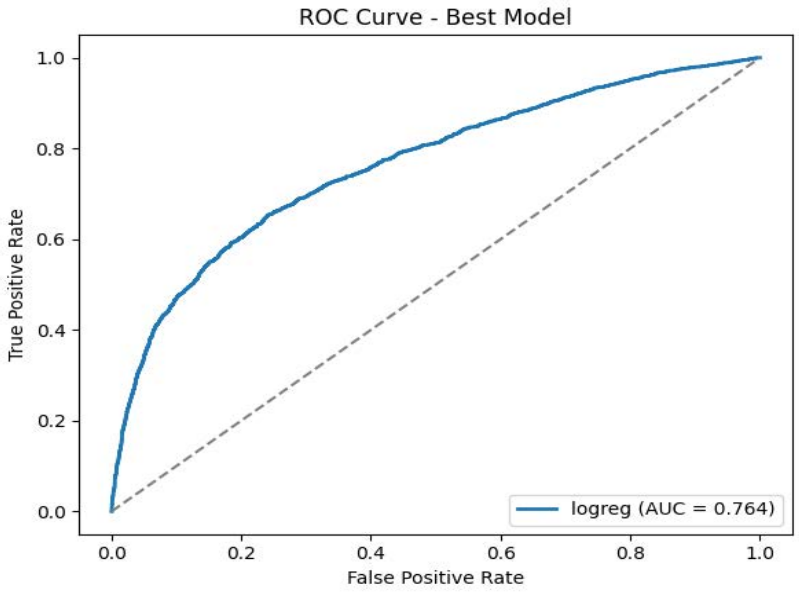

Fig. 1. ROC curve of the best-performing (Logistic Regression) model (Photo/Picture credit: Original).

Logistic regression's strong performance suggests two possibilities: first, after careful feature engineering, key predictive patterns in the dataset are approximately linearly separable; second, sophisticated nonlinear models (e.g., GBDT and MLP) are more sensitive to hyperparameters and may not have reached their optimal state under the study's tuning constraints. This finding challenges the conventional assumption that "increased complexity equals improved predictive power," emphasizing simple, interpretable models as robust baselines in real-world applications.

3.2 Business Interpretation and Trade-offs of the Best Model

Further analysis of the logistic regression model's confusion matrix and classification report revealed its business relevance. The model reliably identifies non-defaulting customers (Class 0) with an accuracy of 0.874. The confusion matrix (Fig. 2) shows that among 7,009 actual non-defaulters, 5,731 were correctly identified, and among 1,991 actual defaulters, 1,166 were successfully flagged.

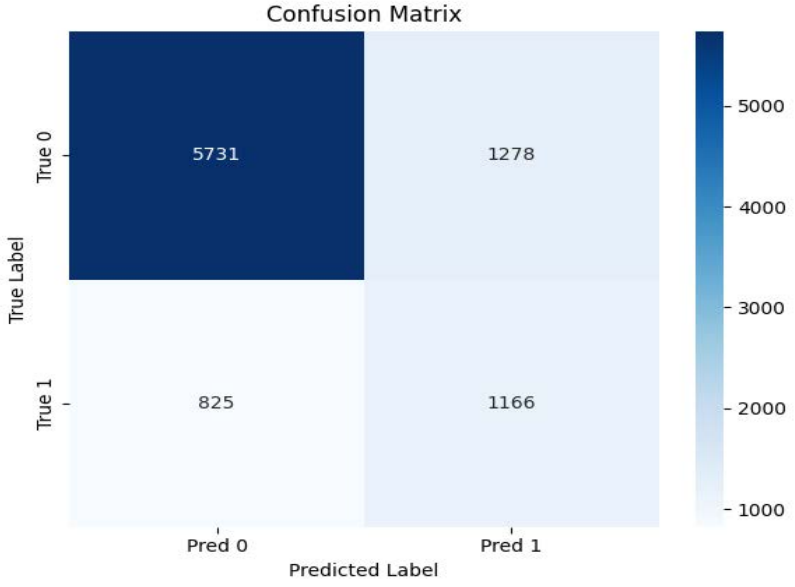

Fig. 2. Confusion matrix of the Logistic Regression model on the test set (Photo/Picture credit: Original).

But not really impressive in predicting class 1 customers that default at a probability of 0.477. There are many customers flagged as high risk who are actually not defaulters. This is a really good trade-off for financial risk mitigation, because such false positives only mean added costs for manual review, while their counterparts, false negatives, incur a loss of money directly. Thus, it can serve as a highly automated first filtration to narrow down thousands of high-risk clients, who can most probably undergo secondary deep audits by risk control specialists, optimizing risk management resource allocation.

3.3 Analysis of Feature Importance and Insights on Risk

The analysis of permutation importance (Fig. 3) indicated how much each feature contributed. These insights were similar to what would be expected by intuition based on finance, in that historical bill amounts are much more important: the bill amount from 4 months ago is the most predictive, suggesting that customers' financial situations and consumption levels months previously consistently foreshadow what will be their ability to repay in the future. Possibly, these suggest income behaviors or resilience. The second most important predictor is credit limit, negatively correlated with the probability of default-that is, higher limits are provided by banks to customers with better assessed credit quality. Features of immediate repayment behavior, like whether the payment has been timely or derived features thereof, are also ranked high because, next to on-time payments, timely payment indicates the greatest positive signal against future creditworthiness. The demographic attributes, such as gender and educational level, have little importance; thus, there is a consensus regarding risk control that dynamic, behavior-based financial data would be more predictive than static personal background information.

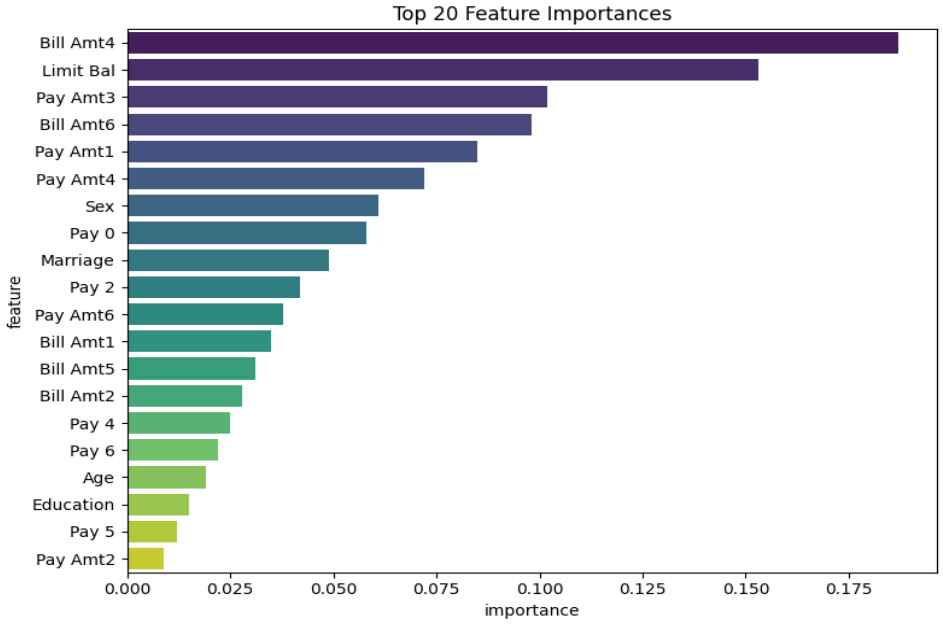

Fig. 3. Top 20 feature importances based on permutation importance for the Logistic Regression model (Photo/Picture credit: Original).

4. Conclusion

This study enacts an elaborate end-to-end machine learning pipeline to evaluate algorithms and facilitate interpretative analysis systematically for credit card default prediction. The major evidence from the empirical research indicates that logistic regression on the UCI dataset had the highest discrimination power in default prediction (AUC: 0.764) and performed better than existing more complex models like gradient boosting machines. Most importantly for practitioners and researchers is the reminder that within a particular data context, simple, highly interpretable linear models might give the best trade-off of performance for transparency, thus strengthening the empirical evidence found in other forms of financial prediction. Such a similar argument corroborates broader applicability for the general principle of interpretability, one of the cornerstones of the burgeoning field of Explainable Artificial Intelligence (XAI) in finance.

An in-depth assessment of the shortlisted models exposes the principal drivers of default risk in their respective contexts. Past financial conduct trajectories that define repay-

ISSN 2959-6130

ment behavior (average bill amounts from past months) and beginning credit history (ultimate credit limit) remain the most predictive variables in the global sense, whereas repayment behavior in the near past does procure significant signaling in the short run. This renders some support to the behavioral risk assessment paradigm from the point of view of data science, thus furnishing direct empirical evidence to financial institutions for the enrichment of their risk control models' feature ecologies.

An obvious limitation this study faces is that the data used are local and warrant validation across wider geographic domains and economic cycles; hyperparameter searches for the more complex models, mainly the multi-layer perceptron, are constrained due to computational power; and global feature importance analysis constitutes the backbone of the explanation from an interpretability perspective. In the future, an attempt will be made to validate the findings on larger and more heterogeneous datasets, implement state-of-the-art hyperparameter tuning techniques, e.g. Bayesian Optimization, to unleash the full potential of the complex models, provide local explanation techniques for individual predictions to establish trustworthy AI risk control systems, and accommodate online learning so that the model ramifications can adapt to the dynamic landscape of financial markets. The ultimate aim of this research is the demonstration of the applicability of interpretable, robust, and simple machine learning methods to the solution of financial risk management while simultaneously increasing operational efficacy and controlling risk.

References

[1] Mestiri S, Hiboun S M. Credit scoring using machine learning and deep learning-based models. Data Science in Finance and Economics, 2024, 2: 236-248.

[2] Li Y, Chen W. A comparative performance assessment of ensemble learning for credit scoring. Mathematics, 2020, 8(10): 1756.

[3] Yeh I C, Lien C H. The comparison of data mining techniques for the predictive accuracy of probability of default of credit card clients. Expert Systems with Applications, 2009, 36(2): 2473-2480.

[4] Lessmann S, Baesens B, Seow H V, Thomas L C. Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. European Journal of Operational Research, 2015, 247(1): 124-136.

[5] Alam T M, Shaukat K, Hameed I A, Luo S, Sarwar M U, Shabbir S, Khushi M. An investigation of credit card default prediction in the imbalanced datasets. IEEE Access, 2020, 8: 201173-201198.

[6] Wang L, Zheng J, Yao J, Chen Y. A multi-stage integrated model based on deep neural network for credit risk assessment with unbalanced data. Kybernetes, 2025, 54(9): 4626-4657.

[7] Golec M, AlabdulJalil M. Interpretable LLMs for credit risk: A systematic review and taxonomy. arXiv preprint, 2025: arXiv:2506.04290.

[8] Sowmiya M N, Jaya Sri S, Deepshika S, Hanushya Devi G. Credit risk analysis using explainable artificial intelligence. Journal of Soft Computing Paradigm, 2024, 6(3): 272-283.

[9] Lin S H, Nguyen T, Lai H H, Huang M H. Credit card default prediction: A comparative study of machine learning models based on accuracy, sensitivity, and specificity. Journal of Namibian Studies, 2023, 33.

[10] Crosato L, Liberati C, Repetto M. Lost in a black-box? Interpretable machine learning for assessing Italian SMEs default. Applied Stochastic Models in Business and Industry, 2023, 39(6): 829-846.