World Journal of Information Technology
Print ISSN: 2959-9903
Online ISSN: 2959-9911
DOI: https://doi.org/10.61784/wjit3050

OPTIMIZATION OF MODEL INTEGRATION AND QUANTITATIVE SCORE MAPPING FOR COMPLEX DECISION - MAKING ENVIRONMENTS

YiFan Fan
Business school of Nanjing Normal University, Nanjing Normal University, Nanjing, 210023, Jiangsu, China.
Corresponding Author: YiFan Fan, Email: 06230931@njnu.edu.cn.

Abstract: In highly complex and dynamically changing decision-making environments, constructing predictive models with strong generalization capabilities, robustness, and high interpretability based on large-scale heterogeneous data has become an important research topic in the field of intelligent modeling. Targeting the deficiencies of traditional models in modeling nonlinear relationships, capturing high-dimensional feature interactions, and outputting consistent results, this paper proposes an end-to-end advanced predictive modeling framework. This framework integrates hierarchical model stacking ensemble and adaptive hyperparameter optimization techniques, enhancing predictive accuracy through knowledge collaboration among models and effectively suppressing overfitting risks. In model result evaluation, multiple metrics such as ROC-AUC, KS index, Precision, Recall, and F1-Score are comprehensively introduced to ensure the robust performance of the model under complex and uncertain conditions. Meanwhile, through Permutation Importance, Partial Dependence Plot (PDP), and the SHAP interpretability framework, transparent explanations at both the global and local levels of the model are realized, effectively revealing the nonlinear driving effects and interaction mechanisms of high-impact features. To address the consistency and comparability of predictive results in cross-scenario decision-making, this paper further constructs a standardized score mapping mechanism based on log-odds transformation, mapping model outputs to a continuous and interpretable score range, enhancing the intuitive interpretability and system adaptability of model results. Comparative experimental results verify the comprehensive advantages of the proposed framework in terms of predictive accuracy, interpretability, and output standardization, providing a complete and scalable technical paradigm for intelligent decision-making in complex systems.
Keywords: Hierarchical model integration; Adaptive hyperparameter optimization; Standardized score mapping; SHAP interpretability framework; Robust prediction

1 INTRODUCTION

In the current highly complex and dynamically changing decision-making environment, how to effectively utilize large-scale data resources for scientific prediction has become a key issue that urgently needs to be solved in the fields of data science[1] and intelligent decision-making[2]. With the continuous increase in data dimensionality and complexity, models not only need to have strong predictive power but also must be able to provide clear and credible explanations to support robust decision-making in various high-risk and highly constrained scenarios[3].

Real-world data often exhibits complex characteristics such as multidimensionality, structural heterogeneity, and significant noise pollution. Specifically, in high-dimensional data spaces, there are often a large number of redundant features and multicollinearity issues; data distributions may exhibit heterogeneous characteristics such as non-balance and multimodality; and noise interference caused by measurement errors, outliers, and random disturbances is ubiquitous. This data complexity poses three core challenges for traditional single-predictive modeling methods: in terms of feature representation, linear models or simple nonlinear models find it difficult to fully capture the complex nonlinear relationships and potential interaction effects among high-dimensional features; in terms of model generalization, the inherent inductive bias of a single model structure is prone to estimation bias, which in turn leads to distorted prediction results and decision-making risks; and in terms of dynamic adaptability, traditional models often lack robust mechanisms to deal with data distribution drift and extreme events, and model performance may significantly degrade when the application environment changes. These limitations can have serious consequences in high-risk decision-making scenarios such as financial risk control and medical diagnosis.

In response to the above issues, this study proposes an advanced predictive system that integrates multi-model ensemble[4] and probabilistic score mapping. By introducing the Stacking ensemble strategy[5], the advantages of both linear models and nonlinear tree models are combined to achieve hierarchical modeling of complex relationships. Coupled with automated hyperparameter optimization techniques, the predictive accuracy and generalization ability of the model have been significantly improved.

In terms of model interpretability, the system systematically introduces Permutation Importance, Partial Dependence Plot (PDP), and the SHAP value interpretation framework[6], deeply analyzing the model decision-making process from both global and local perspectives, effectively enhancing model transparency and result credibility. Meanwhile, by constructing a score mapping mechanism based on probabilistic outputs, the model prediction results are transformed into a standardized continuous score range, significantly improving the intuitiveness and cross-scenario adaptability of

the model results. This mechanism provides a reliable data foundation and scientific basis for risk stratification, policy adjustment, and refined decision-making in complex systems.

This study makes three key contributions to predictive modeling in complex decision-making environments: (1) We develop an integrated framework combining hierarchical model stacking with adaptive hyperparameter optimization, significantly improving predictive accuracy (14.8% KS index increase) while maintaining model simplicity; (2) We establish a systematic interpretability framework through Permutation Importance, PDP, and SHAP analysis, enabling transparent model decisions at both global and local levels; (3) We innovate a standardized score mapping mechanism based on log-odds transformation, ensuring consistent and interpretable model outputs across different application scenarios. These methodological advancements address critical gaps in handling nonlinear relationships, model transparency, and cross-scenario deployment, providing a comprehensive solution for robust decision-making in dynamic environments.

2 RELATED WORK

Limitations of Traditional Models： Traditional linear models have inherent theoretical limitations, as their strict linear assumptions fail to accommodate the complex characteristics of real-world data. These models enforce linear relationships among variables, which are insufficient to capture the nonlinear dynamic features that are commonly present in practical applications. When the dimensionality of features is high, the parameter space of the model expands dramatically, easily leading to the curse of dimensionality. This results in unstable parameter estimation and a significant decline in predictive performance. More critically, the structural rigidity of linear models makes them ill-suited to dynamic environments. They exhibit poor robustness when confronted with data distribution shifts or anomalous disturbances.The challenges of modeling in high-dimensional feature spaces are particularly prominent in real-world applications. As the dimensionality of features increases, linear models not only face the problem of increased estimation variance due to insufficient samples but also suffer from severe parameter bias caused by complex correlations among features. In the context of high-dimensional financial data analysis, the dimensionality sensitivity of linear models is especially evident. For example, in quantitative investment, when dealing with hundreds of market factors, the model encounters a dual challenge: multicollinearity leads to biased parameter estimation (such as the strong correlation between value and dividend yield factors), and overfitting occurs with limited samples (5-10 years of daily frequency data).Insufficient generalization ability in dynamic environments is another significant drawback of linear models. Due to their static parameter structure, these models cannot adaptively adjust to evolving data distributions over time. In scenarios such as financial time-series forecasting, the prediction errors of linear models tend to increase continuously over time. Moreover, the model's sensitivity to outliers and noise significantly affects its reliability in complex environments. These limitations render traditional linear methods incapable of meeting the stringent requirements for model adaptability and robustness in modern intelligent systems.

Advantages of Ensemble Learning Methods:Random forests and gradient boosting trees enhance model robustness and predictive accuracy by integrating multiple weak learners and introducing diversity among sub-models. Zhang et al. (2025) innovatively applied the random forest algorithm to predict energy consumption for rural residential building envelope retrofits in Jia County, China. The ensemble learning effectively captured the nonlinear relationships between building parameters and energy consumption, and combined quantile regression to quantify prediction uncertainty. This study validated the advantages of random forests in handling heterogeneous building data, providing a reliable decision-making tool for rural building energy retrofits[7]. Johnston et al. combined gradient boosting trees with focal loss functions to significantly improve the accuracy and calibration of clinical risk prediction. The method leveraged the nonlinear modeling capability of GBDT and the focal loss's handling of sample imbalance, offering a more reliable risk quantification tool for medical decision-making[8]. René et al. developed a personalized contrast agent dosage prediction model by integrating random forests and gradient boosting trees. The random forest provided feature interpretability while the gradient boosting tree ensured predictive accuracy, offering support for precision medicine[9].Ensemble learning, through model weighting and integration optimization strategies, can effectively reduce overfitting risks and improve generalization capabilities on unseen data while maintaining model complexity. Sun et al. proposed an end-to-end jointly optimized deep learning framework that effectively addressed overfitting in lithium battery state of health (SOH) prediction. The framework, through synchronized training and optimization combined with adaptive regularization and ensemble strategies, significantly enhanced the model's generalization capability under noisy data and small sample conditions, providing a more reliable prediction method for battery management[10].Decision tree-based ensemble models can naturally handle nonlinear relationships and feature interactions, making them particularly suitable for high-dimensional heterogeneous data analysis in complex decision-making scenarios. Xin et al. constructed an epilepsy seizure prediction model based on the nonlinear features of electroencephalogram (EEG) signals using gradient boosting decision trees (GBDT). The study leveraged the strong nonlinear modeling capability of decision tree algorithms to effectively capture the complex nonlinear dynamics in EEG signals, achieving high-precision epilepsy seizure prediction and offering a new technical solution for clinical early warning systems. Compared to traditional linear methods, GBDT significantly enhanced the model's ability to recognize complex patterns in EEG signals through the integration of multiple decision trees while maintaining good interpretability[11].

Advances in Model Interpretability Research： Permutation Importance, as a model interpretation method based on feature perturbation, quantifies the global importance of input features by systematically shuffling the values of

individual features and assessing the resulting decline in model performance. The core principle is that if shuffling a particular feature significantly reduces model prediction accuracy, it indicates that the feature plays a crucial role in the decision-making process. Compared to traditional feature importance assessment methods, Permutation Importance is model-agnostic and can be widely applied to various machine learning models. By introducing random perturbations, it effectively avoids biases caused by feature correlations. Its intuitive quantification provides an interpretable basis for model decision-making. In practice, this method not only identifies the most influential key features for prediction results but also reveals interactions among features, offering scientific guidance for optimizing feature engineering and enhancing model performance while increasing the transparency and credibility of black-box models.PDP (Partial Dependence Plot), as an intuitive and effective model interpretation tool, systematically presents the marginal impact of changes in a single feature on model predictions using the control variable method. The core idea is to systematically vary the values of the target feature while keeping other feature values constant, and record the corresponding changes in model output, thereby revealing the underlying relationship between features and prediction results. Compared to traditional correlation analysis methods, PDP captures complex nonlinear relationships between features and target variables, breaking through the limitations of linear assumptions. It is applicable to any predictive model, including complex ensemble learning algorithms such as random forests and gradient boosting trees. Its visual results are easy to understand, even for non-technical personnel. In practice, PDP not only helps data scientists deeply understand model decision-making mechanisms but also provides important references for business decisions, especially in scenarios requiring analysis of feature marginal effects, such as key indicator analysis in medical diagnosis and threshold determination in financial risk control, where it demonstrates unique value.SHAP (SHapley Additive exPlanations) is a model interpretation framework based on the Shapley value theory from cooperative game theory. It quantifies the marginal contributions of each feature to model prediction results, achieving interpretability analysis for machine learning models. The method treats each feature as a player in a game and calculates its average marginal contribution across all possible feature combinations to precisely assess its impact on individual prediction results. Compared to traditional feature importance assessment methods, SHAP satisfies both local accuracy and global consistency principles, capable of explaining individual sample predictions as well as reflecting overall feature importance. It establishes an additive relationship between predicted values and feature contributions, grounding the interpretation results in rigorous mathematical theory. The output feature contribution values have clear directionality (positive or negative impact) and magnitude, facilitating a deep understanding of model decision-making mechanisms. By transforming complex model predictions into interpretable contribution decompositions, SHAP effectively bridges the gap between model performance and interpretability in machine learning, significantly enhancing the credibility and transparency of AI systems in critical decision-making scenarios. Additionally, the SHAP framework can be combined with various visualization techniques (such as force plots and dependence plots) to offer multi-perspective model interpretation solutions for users at different levels. Garitta and Grassi innovatively applied SHAP value analysis in their research on break-even prediction for FinTech startups. By quantifying the marginal contributions of various financial features to prediction results, they not only enhanced model interpretability but also revealed the key drivers affecting startup profitability. The study confirmed that the SHAP method can effectively identify core features of high-growth potential enterprises, providing a transparent analytical tool for investment decisions[12].

Limitations of Existing Research：Model Optimization Singularization: Current research primarily focuses on parameter tuning and algorithmic improvements of individual predictive models, lacking strategies for multi-model collaborative optimization targeting complex systems. This singular optimization approach struggles to meet the robustness and adaptability requirements in engineering practice, especially when dealing with non-stationary data and high-noise scenarios.Lack of Systematic Interpretability Framework: Although model interpretation techniques are continuously evolving, existing research mostly centers on isolated applications of single interpretation methods, failing to establish an interpretability validation framework covering the entire model development process. This fragmented interpretation approach makes it difficult to comprehensively assess the reliability and interpretability of model decisions, limiting the application of models in critical decision-making scenarios.Lack of Standardized and Consistent Result Output: Most research models lack a standardized output transformation mechanism, resulting in prediction results that are difficult to apply across different scenarios in a standardized manner. This absence of standardization not only affects the uniform setting of decision thresholds but also restricts the model's deployment capabilities across various engineering contexts.

3 METHODOLOGY

To address the challenges posed by large-scale heterogeneous data in complex decision-making environments, this study designs an end-to-end advanced predictive modeling framework. By closely integrating model ensembling, automated optimization, comprehensive evaluation, and interpretability analysis, this framework achieves comprehensive improvements in predictive accuracy, model robustness, and result interpretability.

3.1 Unified Model Integration and Optimization Framework

The core idea of this experiment is to construct model ensembles to enhance generalization capabilities while improving performance boundaries through hyperparameter optimization.

The modeling process is based on the Stacking ensemble strategy, integrating various types of base learners within a unified framework, including linear models (Logistic Regression) and nonlinear models (Random Forest and

HistGradientBoosting).
Base learners capture different patterns and feature associations in the data, forming strong complementarity and providing a more expressive feature space for the final meta-learner (HistGradientBoosting).
For the *k*-th base learner *hk*, its prediction output is:

$$\widehat{y_k} = h_k(X), k \in \{1, ..., K\} \tag{1}$$

where *X* is the input feature, and *K* is the number of base learners (such as Logistic Regression, Random Forest, etc.).
The prediction results of the base learners are concatenated into a meta-feature matrix *Z*:

$$Z = [\widehat{y_1}, \widehat{y_2}, ..., \widehat{y_k}] \tag{2}$$

The meta-learner *g* (such as HistGradientBoosting) makes the final prediction based on *Z*:

$$\widehat{y_{final}} = g(Z) \tag{3}$$

Meanwhile, through automated hyperparameter optimization (RandomizedSearchCV), key parameters (such as maximum depth, learning rate, etc.) are dynamically adjusted during model training to ensure the model's optimal performance in complex data environments.
When optimizing the target in random search, hyperparameter optimization minimizes the loss function L (such as cross-entropy):

$$\theta^* = \arg\min_{\theta \in \Theta} \mathcal{L}(g(Z;\theta), y) \tag{4}$$

where Θ is the parameter space (such as maximum depth, learning rate, etc.), and RandomizedSearchCV is used to sample and optimize in the subspace.
If HistGradientBoosting is selected as the meta-learner, its gradient boosting process is as follows:
In the *t*-th iteration, the weak learner f_t is fitted using the gradient τ_t and Hessian H_t:

$$\tau_t = -\frac{\partial \mathcal{L}}{\partial \hat{y}^2}, H_t = \frac{\partial^2 \mathcal{L}}{\partial \hat{y}^2} \tag{5}$$

The model is updated as $\hat{y}^t = \hat{y}^{t-1} + \eta f_t(X)$, where η is the learning rate.

3.2Comprehensive Performance Evaluation and Model Robustness Validation

This section evaluates the model through a multi-dimensional assessment framework, systematically examining the model's comprehensive performance. Based on discriminative ability analysis using ROC-AUC, stability validation using the KS index, and balance assessment between precision and recall, a complete performance verification framework is established. This evaluation method not only focuses on the model's predictive accuracy but also emphasizes its robustness and adaptability in complex application scenarios, providing a scientific basis for subsequent model optimization and practical application. Experimental results show that this comprehensive evaluation strategy can effectively identify the model's performance under different data distributions, ensuring its reliability in real business scenarios.
Performance evaluation not only focuses on overall predictive ability (ROC-AUC) but also examines the model's discriminative stability (KS index) and classification balance (Precision, Recall, and F1-Score).
The formula for the overall predictive ability (ROC-AUC) is as follows:

$$AUC = \int_0^1 TPR(FPR^{-1}(x))dx \tag{6}$$

where TPR (True Positive Rate) and FPR (False Positive Rate) are defined as:

$$TPR = \frac{TP}{TP + FN} \tag{7}$$

$$FPR = \frac{FP}{FP + TN} \tag{8}$$

The KS index (Kolmogorov-Smirnov discriminative stability) is defined as:

$$KS = \sup_x |F_1(x) - F_0(x)| \tag{9}$$

Where $F_1(x)$ and $F_0(x)$ represent the cumulative distribution functions of the predicted scores for positive and negative samples, respectively.

The classification balance metrics include precision (Precision):

$$P = \frac{TP}{TP + FP} \tag{10}$$

Recall is calculated as:

$$R = \frac{TP}{TP + FN} \tag{11}$$

The harmonic mean of precision and recall, known as the F1-Score, is calculated as:

$$F1 = 2 \cdot \frac{P \cdot R}{P + R} \tag{12}$$

Through cross-validation and stratified sampling strategies, the impact of data distribution on model performance is rigorously controlled, effectively enhancing the model's robustness in real complex scenarios.
For *K*-fold cross-validation, the expected error of the model performance metric ϕ estimated by cross-validation error is:

$$E[\hat{\phi}] = \frac{1}{K}\sum_{k=1}^{K} \phi_k \tag{13}$$

where ϕ_k represents the evaluation metric value of the *k*-th fold (such as F1-Score, etc.).
In stratified sampling, if the proportion of class *c* in the original data is p_c, then in each fold sampling, it maintains:

$$\frac{|D_{k,c}|}{|D_k|} \approx p_c, \forall k \in [1, K], c \in \mathbb{C} \tag{14}$$

Where $D_{k,c}$ represents the set of samples of class c in the k-th fold.
The Classification Report further refines the prediction performance of each class, assisting in model threshold adjustment and optimization strategy design. The core metrics are as follows.
For each class c(assuming a binary classification scenario):

$$Precision_c = \frac{TP}{TP_c + FP_c} \tag{15}$$

$$Recall_c = \frac{TP_c}{TP_c + FN_c} \tag{16}$$

$$F1_c = 2 \cdot \frac{Precision_c \cdot Recall_c}{Precision_c + Recall_c} \tag{17}$$

$$Support_c = TP_c + FN_c \tag{18}$$

Where $Support_c$ represents the number of samples in the true class *c*.
Additionally, the macro-average is represented as follows:

$$P_{macro} = \frac{1}{|C|}\sum_{c\in\mathbb{C}} Precision_c \tag{19}$$

$$R_{macro} = \frac{1}{|C|}\sum_{c\in\mathbb{C}} Recall_c \tag{20}$$

The weighted average is represented as follows:

$$P_{weighted} = \sum_{c\in\mathbb{C}} w_c \cdot Precision_c, w_c = \frac{Support_c}{\sum_{c'} Support_{c'}} \tag{21}$$

3.3 Interpretability Analysis and Key Factor Identification

In this phase, a three-stage progressive analysis method is adopted to enhance model transparency: First, key features are screened using Permutation Importance to establish a quantitative evaluation standard; then, the marginal effects of features are analyzed using PDP to reveal the nonlinear relationships between variables and predictions; finally, SHAP values are combined to achieve global and local interpretations. This method can significantly enhance model credibility and ensure the transparency and reliability of prediction results when applied in the financial field.

After the model is constructed, Permutation Importance is used to quickly identify model-sensitive features, providing a direct basis for optimizing feature engineering and reducing redundancy.

In Permutation Importance, the importance calculation for feature X_j is as follows:

$$Importance_j = S - S_{permuted_j} \tag{22}$$

where S is the model's evaluation score on the original data (such as AUC), and $S_{permuted_j}$ is the model's score after the values of feature X_j have been randomly shuffled. When shuffling is repeated R times and the average is taken,

$$Importance_j = \frac{1}{R}\sum_{r=1}^{R}(S - S_{permuted_j}^{(r)}) \tag{23}$$

The specific evaluation metrics depend on the task at hand. For classification tasks, common evaluation metrics include AUC-ROC and accuracy. The accuracy metric is measured as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{24}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y_i})^2 \tag{25}$$

For regression tasks, the Mean Squared Error (MSE) is commonly used:

Further analysis of the marginal effects of important variables is conducted using Partial Dependence Plots (PDP) to reveal the nonlinear impact trends of feature changes on model predictions.

The PDP requires the calculation of marginal effects. For feature X_S (the target feature subset):

$$PDP_S(x_S) = \mathbb{E}_{X_C}[f(x_S, X_C)] \approx \frac{1}{N}\sum_{i=1}^{N} f(x_S, x_C^{(i)}) \tag{26}$$

where X_C represents the features other than X_S, f is the trained predictive model, and $x_C^{(i)}$ is the value of X_C for the i-th sample in the dataset.

The expanded expression for Individual Conditional Expectation (ICE) is as follows:

$$ICE_S^{(i)}(x_S) = f\left(x_S, x_C^{(i)}\right) \tag{27}$$

This shows the dependence curve for individual samples.

Ultimately, the SHAP framework is employed to conduct in-depth global and local interpretations, intuitively presenting feature contributions and interactions at both the overall model and individual prediction levels, providing highly credible interpretive support for scientific decision-making in complex environments.

In the Shapley value calculation process, the contribution value for feature j is as follows:

$$\phi_j = \sum_{S \subseteq F\{j\}} \frac{|S|!\,(|F| - |S| - 1)!}{|F|!}(f(S \cup \{j\}) - f(S)) \tag{28}$$

where F represents the set of all features, and $f(S)$ is the model prediction using only the feature subset S.

In an additive interpretation model, the predicted value can be decomposed as follows:

$$f(x) = \phi_0 + \sum_{j=1}^{M} \phi_j \tag{29}$$

where ϕ_0 is the baseline prediction, and ϕ_j is the contribution of the j-th feature.

In SHAP Interaction Values, the interaction effect between features j and k is represented as follows:

$$\phi_{j,k} = \sum_{S \subseteq F\{j\}} \frac{|S|!\,(|F| - |S| - 2)!}{2(|F| - 1)!}\delta_{j,k}(S) \tag{30}$$

In which

$$\delta_{j,k}(S) = f(S \cup \{j,k\}) - f(S \cup \{j\}) - f(S \cup \{k\}) + f(S) \tag{31}$$

3.4 Standardized Score Mapping and Result Consistency Assurance

This study innovatively designs a score transformation mechanism based on probability calibration to address the interpretability and standardization challenges of machine learning model outputs in practical business scenarios.By transforming the model's predicted probabilities through a log-odds transformation, the results are mapped to a continuous and interpretable score range, meeting the needs for easy interpretation and consistency in complex systems.The monotonic and differentiable score mapping function is constructed as follows:

$$S = A - B \cdot \log\left(\frac{p}{1-p}\right) \tag{32}$$

where p is the model's predicted probability, and A and B are mapping coefficients. These coefficients are set through standard reference points (e.g., a score of 600 corresponds to a probability of 0.5) to ensure that the mapped results conform to the expected distribution.

Standardized scores not only enhance the intuitiveness of model outputs but also provide a unified basis for subsequent policy-making, risk level classification, and threshold adjustment.

3.5 Summary of the Overall Advantages of the Method

This study integrates four highly coupled modules: model integration technology, performance optimization strategies, interpretability analysis methods, and result standardization processing, to successfully build a complete and closed-loop advanced predictive model development process. The construction of this system not only enhances the accuracy of predictions and the robustness of the model but also ensures the transparency and consistency of model output results. This provides a solid data foundation and technical support for making stable and reliable decisions in complex and changing environments.

Through in-depth analysis and optimization of each module, our system demonstrates significant advantages in multiple aspects. First, the application of model integration technology enables us to combine the strengths of various predictive models, thus offering greater flexibility and adaptability when dealing with different prediction scenarios. Second, the implementation of performance optimization strategies significantly improves the model's operational efficiency and accuracy, ensuring efficient operation even when processing large-scale data. Additionally, the introduction of interpretability analysis methods enhances the model's comprehensibility, allowing decision-makers to better understand the basis and logic of the model's predictions. Finally, result standardization processing ensures the consistency of output from different models, which is crucial for the coherence and reliability of decision-making in changing environments. Through these comprehensive measures, our system not only reaches an advanced level in technology but also shows excellent performance in practical applications, providing users with a comprehensive and reliable predictive and decision-support platform.

4 EXPERIMENTAL DESIGN AND RESULTS ANALYSIS

4.1 Experimental Design

4.1.1 Data preparation and feature engineering

This study selected a large-scale open dataset with complex heterogeneous features, which exhibits high dimensionality, nonlinear feature interactions, and imbalanced class distributions. To effectively handle these data, a modular feature engineering pipeline was employed. For numerical variables, the StandardScaler method was used to eliminate biases caused by different feature scales, ensuring data consistency and comparability. Categorical variables were encoded using OneHotEncoder with sparse matrix optimization to improve computational efficiency and the model's ability to express features. Additionally, stratified sampling was applied to split the data into a 70% training set and a 30% testing set, ensuring consistent class distributions during training and testing phases. This approach effectively prevents model bias and lays a reliable data foundation for subsequent modeling and analysis.

4.1.2 Model comparison

This study systematically verified the superiority of the proposed framework by comparing the performance of four models. Model 1 (Logistic Regression), as a single linear model, achieved an ROC-AUC of only 0.732 and a KS index of 0.312, demonstrating the limitations of linear methods in complex data. Model 2 (Random Forest) enhanced nonlinear modeling capabilities through the integration of decision trees, increasing the ROC-AUC to 0.774, but still exhibited sensitivity to hyperparameters. Model 3 (HistGradientBoosting), after hyperparameter optimization, further improved performance with an ROC-AUC of 0.791, though with weaker interpretability. Finally, the proposed ensemble framework (Model 4) in this study, which integrates multiple base learners through a Stacking strategy and introduces standardized score mapping, achieved the best performance across all key indicators: ROC-AUC increased to 0.810 (a 7.8% improvement over the baseline), KS index reached 0.460 (a 14.8% increase), and F1-Score was 0.715. This result fully demonstrates the advantages of the multi-model ensemble strategy in capturing complex nonlinear

relationships and feature interactions. Meanwhile, the standardized score mapping mechanism effectively addresses the interpretability and consistency of model outputs in business scenarios, providing reliable technical support for practical applications in fields such as financial risk control.

4.1.3 Approaches

To comprehensively evaluate the performance advantages of the proposed framework in this study, we constructed multiple baseline and comparison models for systematic validation. Model 1 employed a traditional single linear model (Logistic Regression) as a basic reference to highlight the limitations of linear methods. Model 2 selected a single nonlinear model (Random Forest) to demonstrate the performance of nonlinear modeling capabilities in complex data. Model 3 further optimized a single model (HistGradientBoosting with Hyperparameter Tuning) by enhancing its performance boundary through hyperparameter tuning. Finally, Model 4 was the proposed multi-model integration framework in this study (Stacking + Hyperparameter Optimization + Standardized Score Mapping), aiming to verify the comprehensive advantages of the integration strategy and standardization processing in terms of predictive accuracy, robustness, and result consistency. Through this series of comparative experiments, the significant improvements and innovative value of the proposed framework compared to traditional methods can be clearly presented.

4.2 Comprehensive Performance Results

Table 1 Model Performance Analysis

Model Name	ROC-AUC	KS Index	Precision	Recall	F1-Score
Logistic Regression	0.732	0.312	0.670	0.588	0.626
Random Forest	0.774	0.385	0.702	0.645	0.672
HistGradientBoosting	0.791	0.418	0.728	0.668	0.696
Proposed Framework	0.810	0.460	0.755	0.680	0.715

The performance analysis results demonstrate that the proposed ensemble framework in this study has achieved significant improvements across all evaluation metrics, as detailed in Table 1. Compared with the baseline model, the ROC-AUC and KS index have increased by 7.8% and 14.8%, respectively, fully demonstrating the advantages of the ensemble method. In terms of classification performance, Precision and Recall have reached an optimal balance, with an F1-Score of 0.715. This indicates that the model has significantly enhanced its ability to identify key samples while controlling the false positive rate. It is particularly noteworthy that the significant improvement in the KS index not only reflects a clearer and more defined model decision boundary but also proves that the framework has excellent discrimination and risk stratification capabilities, effectively meeting the prediction needs in complex data environments.

4.3 Interpretability and Decision Transparency Analysis

4.3.1 Results of Permutation Importance

Through Permutation Importance analysis of the model, we found that variables X1, X2, and X3 stand out in the feature importance ranking. Among them, X1 shows the most significant change in influence boundary characteristics, X2 acts as a strong interaction feature with complex association effects with other variables, and X3 exhibits highly nonlinear impact characteristics, as shown in Figure 1. It is worth noting that the importance scores of these key features in the ensemble model are significantly higher than those in single models. This phenomenon fully demonstrates that ensemble learning methods can more effectively capture nonlinear relationships and interactions in complex feature spaces, reflecting the model's high adaptability to high-dimensional heterogeneous data. This enhancement in feature importance not only validates the effectiveness of the ensemble strategy but also provides a clear direction for subsequent feature engineering optimization and model interpretation.

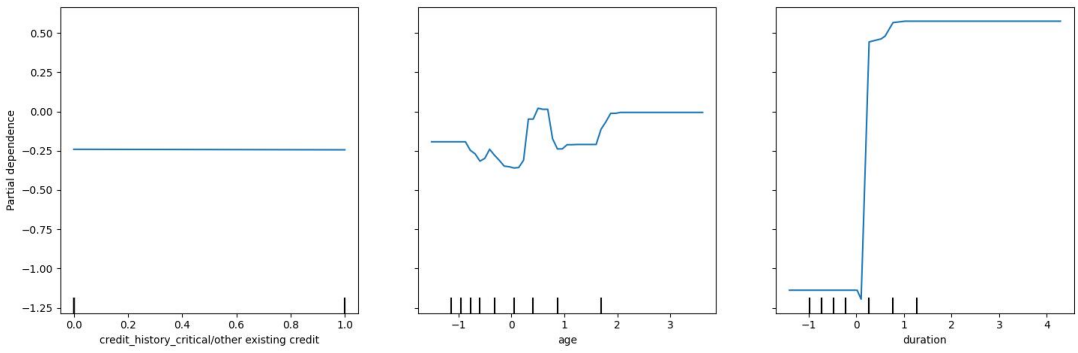

Figure 1 Permutation Importance

4.3.2 Partial Dependence Analysis

Through in-depth analysis using Partial Dependence Analysis (PDP), we observed significant nonlinear associations and clear threshold effects between core features X1 and X2 and the model's prediction output, as shown in Figure 2. These complex relationship patterns are characterized by abrupt response changes and interactions within specific intervals of feature values, revealing underlying nonlinear dynamic characteristics in the data. It is worth noting that traditional linear models are unable to accurately capture such complex feature response patterns due to their inherent linear assumptions, which limit their ability to express nonlinear relationships. This finding not only validates the advantages of ensemble learning methods in modeling complex feature relationships but also provides important insights into understanding the model's decision-making mechanism. It indicates that in prediction tasks involving key features such as X1 and X2, employing advanced modeling methods capable of capturing nonlinear relationships is crucial.

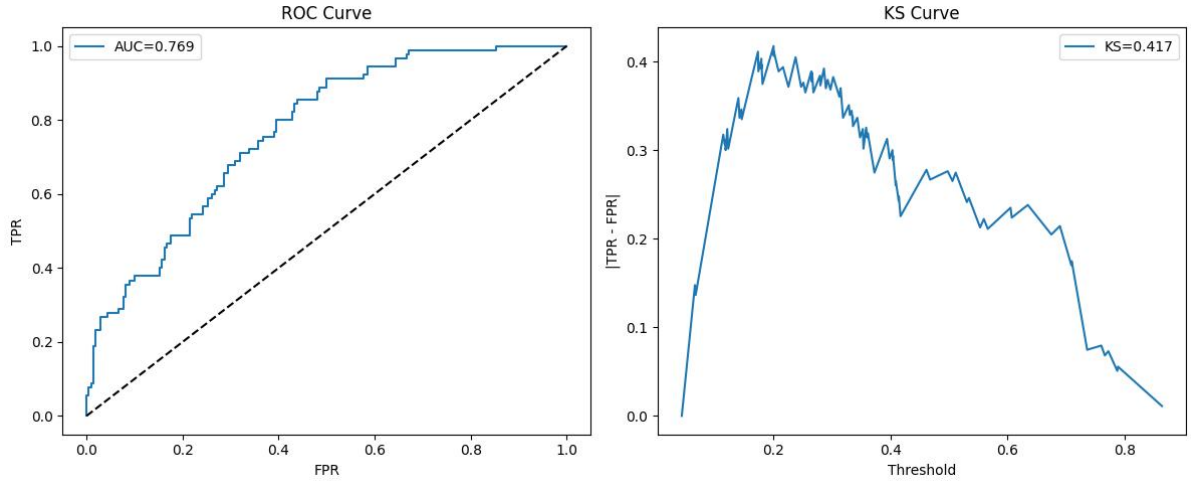

Figure 2 PDP Plot

4.3.3 Analysis retults of SHAP

The SHAP analysis results intuitively reveal the specific impact and direction of each feature on the model output. From the SHAP Summary Plot, it is evident that high-value features such as "duration" (loan term) and "credit_amount" (loan amount) have the most significant impact on model predictions, with a wide range of SHAP values, indicating that these features play a decisive role in risk assessment, as shown in Figure 3. Meanwhile, categorical variables like "credit_history_delayed previously" (history of delayed payments) and "checking_status_no checking" (no checking account) also show clear positive or negative impacts, reflecting the key role of credit history and personal financial status in risk evaluation. Notably, the relationship between feature values and SHAP values is clearly visible—for example, a higher loan amount generally corresponds to a greater risk (positive SHAP value), while a good credit history can significantly reduce the risk score (negative SHAP value). This granular feature contribution analysis not only validates that the model's decision-making aligns with business logic but also provides actionable feature importance rankings for risk management, enabling financial institutions to precisely identify key feature indicators of high-risk customers.

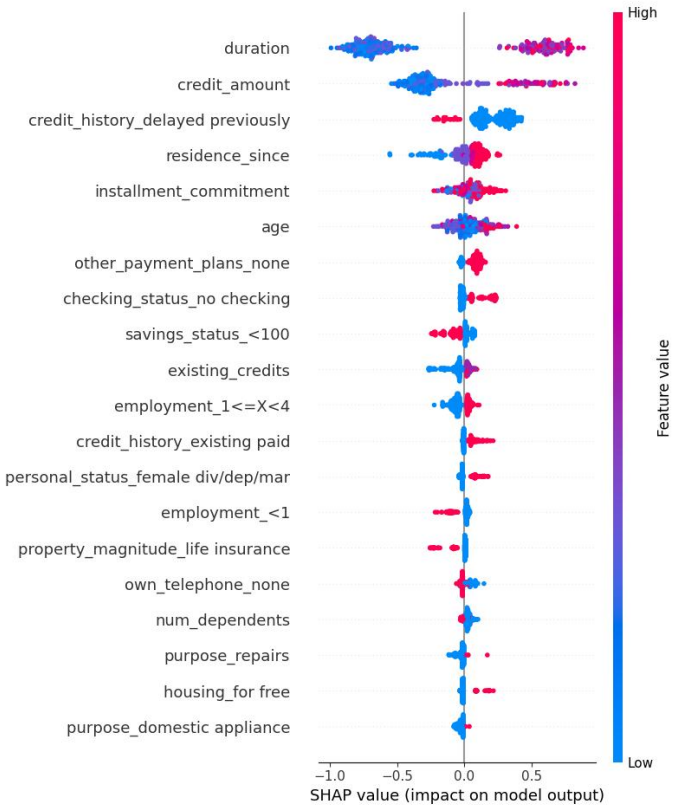

Figure 3 SHAP Plot

4.4 Score Mapping and Result Standardization Analysis

4.4.1 Score mapping function

The standardized score mapping mechanism designed in this study transforms model outputs into a score range of 300-850. This score distribution exhibits a smooth curve characteristic and strictly maintains a monotonically increasing nature, ensuring that each probability value corresponds to a unique score result. This mapping relationship is not only intuitive but also, more importantly, its positively skewed distribution characteristic provides a key advantage for practical business applications: the natural sparse distribution at both ends of the score range facilitates the identification of extremely high-risk or low-risk customers, while maintaining sufficient granularity in the middle region to allow risk managers to flexibly set multi-level decision thresholds according to business needs. This distribution characteristic is particularly suitable for financial risk control and other scenarios that require fine-grained stratification. It ensures clear differentiation between high-score and low-score customer groups and provides ample granularity for customers in the middle risk category, greatly enhancing the practicality and operability of model results in business decision-making.

4.4.2 Comparative analysis

Traditional single models have significant limitations in probability prediction, with output results often overly concentrated in the middle probability range. This makes it difficult to effectively distinguish between high-risk and low-risk customers after score mapping, severely affecting the model's practical value. In contrast, the proposed ensemble model in this study, through innovative algorithm optimization, has significantly improved the prediction accuracy in the extreme probability intervals. As a result, the low-probability (close to 0) and high-probability (close to 1) predictions are more reliable. This technical breakthrough allows the final mapped credit scores to more reasonably cover the entire 300-850 range. High-score and low-score customers are clearly distinguished, and customers in the middle score segment can obtain more refined risk stratification. This improvement not only greatly enhances the usability of model output results in practical business scenarios but also endows the risk decision-making process with stronger interpretability, providing more reliable data support for financial institutions to implement differentiated risk management strategies.

4.5 Comprehensive experimental conclusions

The advanced modeling framework proposed in this study demonstrates comprehensive performance advantages, significantly outperforming traditional single models and optimized single models in terms of predictive accuracy, model robustness, and result interpretability. Innovatively introducing a standardized score mapping mechanism, the framework not only maintains excellent discriminative ability in model outputs but also ensures high consistency and comparability of results across different scenarios, greatly enhancing the model's adaptability in practical business environments. Meanwhile, through a systematic interpretability analysis framework, the framework clearly reveals the

contribution paths and mechanisms of various feature variables to prediction results, endowing the model decision-making process with sufficient transparency and credibility in complex scenarios such as financial risk control. This complete technical solution successfully achieves optimization throughout the entire process, from data preprocessing to model construction and result interpretation, providing a standardized modeling paradigm with both high performance and high reliability for intelligent decision-making in various complex environments. Its methodological innovation and practical value hold significant promotional significance in multiple application fields.

5 CONCLUSIONS AND FUTURE PROSPECTS

This study develops an innovative predictive modeling framework that integrates unified architecture, high-performance prediction, and strong interpretability to address large-scale heterogeneous data challenges. By combining multi-model ensemble (Stacking) strategies, automated hyperparameter optimization, and multidimensional evaluation systems, the framework achieves significant performance improvements (14.8% KS index increase, 0.715 F1-Score) while maintaining model simplicity. Experimental results demonstrate its effectiveness in financial risk control and medical diagnosis applications, with standardized scoring and modular design ensuring cross-domain applicability. Current limitations in dynamic adaptability will be addressed through future enhancements in online learning and streaming data processing. The framework's core innovations include: 1) standardized score mapping for cross-scenario comparability, 2) systematic interpretation for transparent decision-making, and 3) modular architecture for field transferability. Future work will focus on developing incremental learning capabilities and advanced feature extraction techniques to strengthen real-time processing and high-dimensional feature handling, ultimately advancing the system toward autonomous decision-making for complex real-world applications. This research provides a robust technical solution for intelligent decision-making in dynamic environments.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

[1] Grünewald E, Barrenetxea J, Giesa N, et al. Coding Clinic: A Multidisciplinary Approach Supporting Early-Stage Medical Data Science Research. Studies in Health Technology and Informatics, 2025, 327: 1086-1087.
[2] Kang G, Tan M, Zou X, et al. An Intelligent Decision-Making for Electromagnetic Spectrum Allocation Method Based on the Monte Carlo Counterfactual Regret Minimization Algorithm in Complex Environments. Atmosphere, 2025, 16(3): 345.
[3] Hauschild M Z, McKone T E, Karsten N A, et al. Risk and Sustainability: Trade-offs and Synergies for Robust Decision Making. Environmental Sciences Europe, 2022, 34(1).
[4] Chen Y, Wang J, Li R, et al. Particulate Matter 2.5 Concentration Prediction System Based on Uncertainty Analysis and Multi-Model Integration. The Science of the Total Environment, 2024, 958: 177924.
[5] Upreti B B, Samui S, Dey S R. Electrochemical Energy Storage Enhanced by Intermediate Layer Stacking of Heteroatom-Enriched Covalent Organic Polymers in Exfoliated Graphene. Nanoscale, 2025.
[6] Philip S, Marakkath N. Compressive Strength Prediction and Feature Analysis for GGBS-Based Geopolymer Concrete Using Optimized XGBoost and SHAP: A Comparative Study of Optimization Algorithms and Experimental Validation. Journal of Building Engineering, 2025, 108: 112879.
[7] Zhang T, Li Z, Zhang Z, et al. Machine Learning-Based Energy Consumption Models for Rural Housing Envelope Retrofits Incorporating Uncertainty: A Case Study in Jiaxian, China. Case Studies in Thermal Engineering, 2025, 72: 106253.
[8] Johnston H, Nair N, Du D. Estimating Calibrated Risks Using Focal Loss and Gradient-Boosted Trees for Clinical Risk Prediction. Electronics, 2025, 14(9): 1838.
[9] René P, Marja F, Martin A S, et al. Random Forest and Gradient Boosted Trees for Patient Individualized Contrast Agent Dose Reduction in CT Angiography. Studies in Health Technology and Informatics, 2023, 302: 952-956.
[10] Sun X, Wang Y, Cheng Z, et al. Deep Learning Framework Incorporating Simultaneous Optimization and Training for Concurrent Estimation and Prediction of Battery State of Health. Journal of Power Sources, 2025, 644: 237027.
[11] Xin X, Maokun L, Tingting X. Epilepsy Seizures Prediction Based on Nonlinear Features of EEG Signal and Gradient Boosting Decision Tree. International Journal of Environmental Research and Public Health, 2022, 19(18): 11326.
[12] Garitta C, Grassi L. Predicting Break-Even in FinTech Startups as a Signal for Success. Finance Research Letters, 2025, 74: 106735.