Insights in Computer, Signals and Systems

Article

Research on Internet Financial Risk Control Models Based on Machine Learning Algorithms

Yingjie Ma [1,*]

[1] Wells Fargo, Oxnard, 93036, California, USA
* Correspondence: Yingjie Ma, Wells Fargo, Oxnard, 93036, California, USA

Abstract: With the rapid development of internet finance, financial risk issues have become increasingly prominent, and traditional risk control models can no longer effectively address the complex and dynamic scenarios of internet finance. This paper leverages the advantages of machine learning algorithms to propose an internet financial risk control model based on machine learning. Firstly, the data characteristics of financial risk control scenarios are analyzed, and data preprocessing and feature extraction are performed to improve the quality of model input. Secondly, to meet different risk identification requirements, various machine learning algorithms, including decision trees, random forests, support vector machines, and deep learning models, are selected to construct and optimize the risk control model. Experimental verification and comparative analysis are conducted to evaluate the performance of each algorithm in risk control. The results demonstrate that the machine learning-based risk control model significantly outperforms traditional methods in terms of precision and recall for risk identification. Furthermore, real-world case studies validate the model's effectiveness, proving its practicality and reliability in the field of internet financial risk control. Finally, the paper summarizes the main conclusions of the research and proposes directions for further model optimization and scenario expansion, providing technical support and theoretical reference for internet financial risk control.

Keywords: internet finance; risk control; machine learning algorithms; data preprocessing; model optimization

Published: 24 December 2024

1. Introduction

In recent years, with the deep integration of internet technology and financial services, internet finance has rapidly emerged as an essential part of the financial industry. However, alongside its rapid development, financial risks have become increasingly prominent, including credit risk, fraud risk, and operational risk. The diversity, complexity, and high frequency of these risks have rendered traditional risk control models inadequate to meet the current demands of the internet financial industry. Traditional risk control models mainly rely on manual rules and basic statistical methods, which exhibit low identification accuracy and high latency when dealing with massive, high-dimensional, and nonlinear data [1]. Therefore, constructing more efficient, accurate, and dynamic risk control models using advanced technologies has become a pressing challenge in the field of internet financial risk control. In recent years, machine learning algorithms have provided new solutions for internet financial risk control due to their advantages in big data processing, pattern recognition, and nonlinear modeling. Through machine learning algorithms, risk control models can automatically learn and extract critical features from data, achieving precise identification and prediction of risk events. Particularly with the support of algorithms such as decision trees, random forests, support vector machines, and deep learning, risk control models can uncover potential patterns within complex datasets, enhancing their ability to identify risks such as fraud and default. Moreover, the scalability and adaptability of machine learning algorithms enable their effective application in diverse financial business scenarios, demonstrating high practicality and flexibility [2].

2. Research on Risk Control Models Based on Machine Learning Algorithms

In recent years, with the rapid development of big data technology and artificial intelligence, machine learning algorithms have been widely applied to the field of risk control in internet finance. Traditional risk control methods primarily rely on rule-based expert systems and statistical approaches, which struggle to handle nonlinear, multi-dimensional, and massive datasets. In contrast, machine learning algorithms, with their data-driven nature, automatic learning capabilities, and strong generalization abilities, demonstrate higher prediction accuracy and stability in complex and dynamic financial risk scenarios. Existing research shows that the application of machine learning algorithms in risk control models mainly includes supervised learning, unsupervised learning, and ensemble learning methods. In supervised learning, algorithms such as Decision Trees, Random Forests, and Support Vector Machines (SVM) are widely used for credit risk assessment and fraud detection. Decision tree algorithms are highly interpretable and intuitive, making them suitable for constructing transparent risk control rules. However, they are susceptible to noise in the data. Random forests, through the ensemble of multiple trees, significantly improve the stability and precision of the model. Additionally, gradient boosting algorithms like GBDT and XGBoost further optimize training efficiency and predictive capability, making them a current research hotspot. In the realm of unsupervised learning, methods such as Clustering Analysis and Principal Component Analysis (PCA) are commonly employed for anomaly detection and dimensionality reduction. Clustering algorithms can effectively identify potential patterns and anomalies in the data, enabling early warnings for financial fraud risks. Furthermore, with advancements in deep learning, neural network-based models such as Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) have demonstrated robust feature extraction and nonlinear modeling capabilities in complex scenarios. For example, financial transaction data based on time series can leverage RNN for dynamic risk prediction, while CNN can be used for automated extraction and classification of multi-dimensional risk features. Ensemble Learning further enhances the performance of risk control models by combining multiple weak learners, such as Bagging and Boosting, to maximize predictive capability. Researchers have validated the efficiency and reliability of ensemble learning-based risk control models on large-scale datasets. Additionally, the interpretability of ensemble learning makes it widely applicable in financial institutions' risk decision-making processes [3].

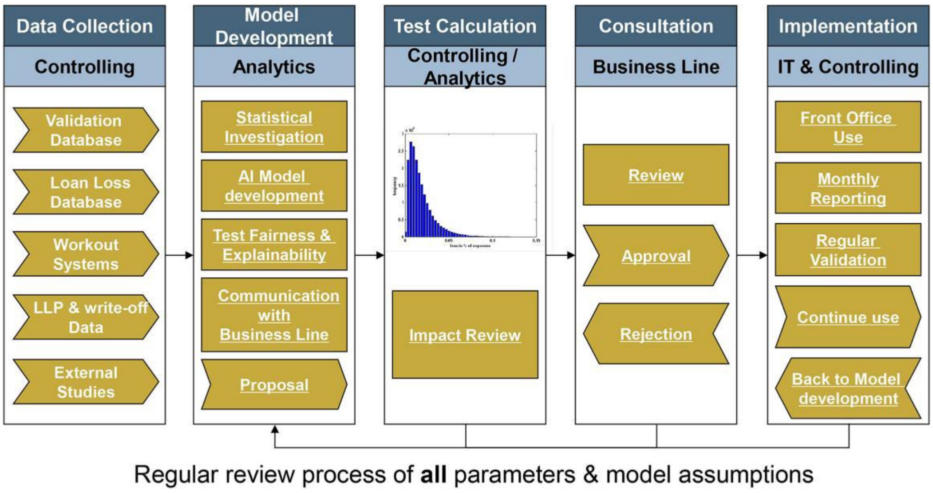

Figure 1. Development and Evaluation Process of Internet Financial Risk Control Models Based on Machine Learning.

Despite significant achievements in selecting and applying machine learning algorithms, some challenges remain. On the one hand, data quality directly impacts the performance of risk control models. Data cleaning, feature engineering, and sample balancing remain key research areas. On the other hand, ensuring the fairness and interpretability of models is a pressing challenge in the current financial risk control domain. Figure 1 illustrates the standardized process for developing and evaluating risk control models, including stages such as data collection, model development, testing, consulting, and implementation. This highlights the importance of lifecycle management for improving risk control effectiveness. In conclusion, machine learning algorithms provide robust technical support for internet financial risk control. Future research should further focus on data processing, model optimization, and the implementation of models in real-world business scenarios to achieve effective and intelligent risk control systems [4].

3. Research Methodology

3.1. Data and Sample Sources

The data used in this research is rich and diverse, covering the multi-dimensional data required for internet financial risk control. This includes user credit data, transaction behavior data, risk write-off data, and external economic environment data. These data sources not only reflect users' credit levels and risk conditions but also reveal potential risk patterns and their influencing factors. To ensure data completeness and representativeness, this study integrates internal databases from financial institutions, third-party data platforms, and external publicly available datasets. Multi-step data preprocessing and sample balancing were performed to create a high-quality dataset suitable for machine learning modeling. The details of the data categories and their sources are summarized in the table below:

Table 1. Data Categories and Sources.

Data Category	Specific Data Content	Data Source	Sample Size	Time Range	Data Description
Credit Risk Data	User loan delinquency records, repayment history	Loan loss database	100,000 records	Jan 2020 - Dec 2023	Used for credit risk assessment, identifying delinquency and default characteristics through historical data.
Fraud Risk Data	Abnormal transaction records, suspicious behavior detection	Transaction monitoring system	50,000 records	Jan 2021 - Dec 2023	Includes suspicious logins, frequent abnormal operations, applied for fraud risk identification and modeling.
User Behavior Data	Browsing behavior, loan application clicks, device information	Third-party data platforms	200,000 records	Jan 2022 - Dec 2023	Captures user behavioral traits on internet platforms, including access frequency, duration, and device preferences.
External Credit Scoring Data	User credit scores, credit reports	External public databases	80,000 records	Jan 2020 - Jun 2023	Data from authoritative credit agencies, used for external validation of user credit assessment.
LLP and Write-Off Data	Bad debt write-offs, overdue account handling records	Financial institution write-off systems	40,000 records	Jan 2020 - Jun 2023	Contains key data on loan write-offs and processing, assisting in risk loss evaluation.

External Economic Data	Inflation rate, unemployment rate, GDP growth rate	National statistics and public reports	10,000 records	Jan 2019 - Dec 2023	Macro-economic indicators used to assess the impact of external economic conditions on risk levels, providing environmental variables for the model.
Transaction Flow Data	Deposit, withdrawal, and transfer records	Financial institution monitoring system	150,000 records	Jan 2021 - Dec 2023	Reflects users' financial activities, including cash flow situations and financial risk status.

3.2. Data Source and Processing Description

The credit risk data and write-off data are primarily sourced from internal financial institution systems, including overdue account history and bad debt write-off records, providing a foundational behavioral history for credit risk assessment. User behavior data is acquired through third-party platforms, capturing users' behavioral trajectories on internet financial platforms. This data helps identify potential fraud and risk signals, such as frequent loan applications and abnormal device switching frequencies. External credit scoring data originates from publicly available credit reports, which, when combined with internal data, enhance the accuracy and comprehensiveness of user credit assessments. Additionally, external macroeconomic data serves as a supplement to environmental variables. Indicators such as inflation rates, GDP growth rates, and unemployment rates are used to evaluate the impact of external economic conditions on users' repayment abilities and risk behaviors. For example, during economic downturns, default risks typically increase. Integrating macroeconomic data improves the robustness and predictive power of the risk control model. To ensure data quality and effectiveness, the research conducted the following data processing steps: Data Cleaning: Duplicate data, anomalies, and invalid entries were removed, while missing values were filled using appropriate methods to ensure data completeness and consistency. Feature Engineering: Feature extraction and construction were carried out, including user behavioral traits, historical transaction records, and external economic environment features. Sample Balancing: To address class imbalance in the credit risk data, the SMOTE (Synthetic Minority Over-sampling Technique) method was applied to balance the distribution of minority class samples, improving the model's ability to identify minority risk events. Data Normalization: Continuous variables, such as transaction amounts and time features, were normalized or standardized to ensure comparability across different data dimensions. Through the above processing steps, this research established a high-quality and comprehensive risk control dataset that encompasses multi-dimensional risk influencing factors. This dataset serves as a robust foundation for training and validating machine learning models. The resulting data framework not only enables accurate risk prediction but also lays the groundwork for optimizing and deploying models in various risk scenarios [5].

3.3. Preprocessing Process

Data preprocessing is an indispensable step in building a risk control model, as it directly affects the performance of machine learning models during training and their final results. Raw data typically contains noise, missing values, inconsistencies, and redundant information, which must be addressed through systematic preprocessing steps to improve data quality and ensure that the model effectively identifies potential risk patterns. Figure 2 illustrates the complete process of data preprocessing and knowledge discovery, including data selection, preprocessing, feature transformation, machine learning modeling, and result evaluation, ultimately facilitating the extraction and application of risk knowledge.

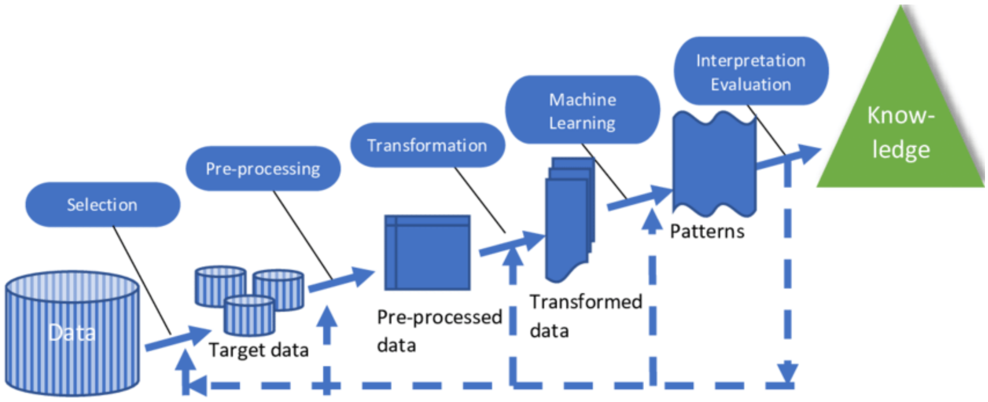

Figure 2. Data Preprocessing and Knowledge Discovery Process for Machine Learning Risk Control Models.

Data selection is the first step in preprocessing, aiming to extract subsets from data sources relevant to the research objectives. In this study, multi-dimensional data were selected by analyzing internet financial risk scenarios, including credit risk data, user behavior data, fraud detection data, and external economic indicators. The selection criteria were primarily based on data relevance, timeliness, and availability. Data preprocessing involves cleaning and initial processing of the selected data, including deduplication, handling missing values, and detecting anomalies. For missing values, strategies such as mean imputation, mode imputation, or interpolation methods were applied to ensure data completeness. Outlier detection used statistical methods such as boxplots and Z-score to identify and address data points significantly deviating from the normal range. To address the issue of class imbalance, this study applied the SMOTE (Synthetic Minority Over-sampling Technique) to generate minority class samples, thereby balancing the data distribution. Feature transformation involves converting and optimizing the preprocessed data into structured formats. This step includes feature extraction, feature selection, and data normalization. Feature extraction employs dimensionality reduction techniques such as Principal Component Analysis (PCA) to identify key risk features from high-dimensional data, reducing computational complexity. Feature selection uses correlation analysis and Recursive Feature Elimination (RFE) to eliminate redundant or irrelevant features, retaining those most significant for risk prediction [6]. Additionally, continuous variables such as transaction amounts and repayment times are standardized or normalized (e.g., using Z-score normalization) to ensure feature consistency and comparability. Machine learning modeling is the core stage following data preprocessing, where machine learning algorithms are applied to the transformed data for modeling and training. In this study, decision trees, random forests, XGBoost, and other models were used for risk identification and evaluation. The input consists of the preprocessed high-quality dataset, and the output generates risk prediction results and identification patterns. Result evaluation and interpretation involve verifying and interpreting the model’s output results, with performance measured using key metrics such as accuracy, recall, and AUC (Area Under the Curve). Furthermore, interpretability tools like SHAP (SHapley Additive exPlanations) were applied to explain the decision-making process and identify the key contributing features, ensuring the model's transparency and credibility. Through the above preprocessing steps, raw data are systematically transformed into high-quality datasets suitable for machine learning algorithms, achieving a transition from data to knowledge. This process provides a solid data foundation for training risk control models and offers effective support for risk identification, prediction, and prevention in internet financial scenarios [7].

3.4. Construction and Implementation of the Risk Control Model

The construction of the risk control model is the core process within the internet financial risk control system. Through scientific data processing and efficient model algorithms, this process achieves identification and prediction of financial risks. Figure 3 illustrates the construction and implementation process of machine learning-based risk control models, encompassing key stages such as theoretical foundation, data preprocessing, feature engineering, model training and evaluation, and performance comparison. The entire process follows a closed-loop path of data-driven modeling, optimization, and validation, aimed at enhancing the precision and practicality of risk control.

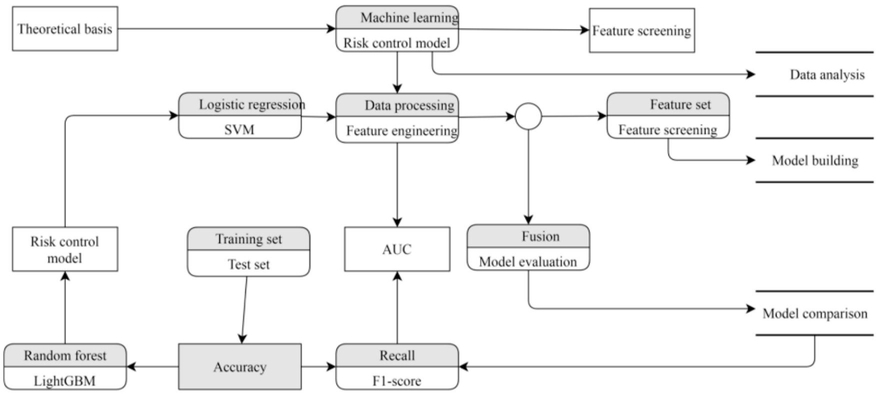

Figure 3. Construction and Evaluation Process of Machine Learning-Based Internet Financial Risk Control Models.

The construction of the risk control model begins with the theoretical foundation, where the risk types and influencing factors are clearly defined. Based on the theoretical framework of machine learning, appropriate algorithms for risk identification and evaluation are selected, including Logistic Regression, Support Vector Machines (SVM), Random Forests, and LightGBM. Data processing and feature engineering form the critical prerequisites for model construction. Data processing includes cleaning, deduplication, filling missing values, and detecting anomalies to ensure the quality and reliability of the input data. Feature engineering involves extracting, selecting, and screening features from vast datasets to retain the most relevant ones for risk prediction, forming an optimized feature set. During feature screening, methods such as correlation analysis and Recursive Feature Elimination (RFE) are combined to reduce redundancy and model complexity, improving computational efficiency. In the model training and validation stage, the dataset is divided into training and testing subsets. The training set is used for parameter learning and optimization. Model performance is evaluated using multi-dimensional metrics such as AUC, accuracy, recall, and F1-Score to ensure predictive accuracy and generalization ability. Logistic regression and SVM are used as baseline models, while ensemble methods like random forests and LightGBM enhance overall performance through iterative optimization. The model fusion and evaluation stage further improves the stability and accuracy of risk prediction by combining the performance of multiple models. Fusion strategies include hard voting, soft voting, and weighted averaging methods, where predictions from different models are aggregated to produce the final risk control model. During evaluation, AUC, recall, and F1-Score are comprehensively considered to verify the model's robustness and applicability across varying data distributions. Finally, through model comparison and result analysis, the best-performing model is selected for deployment and application. The deployment process is tailored to the actual risk control requirements of internet financial platforms, with continuous monitoring and validation ensuring the long-term stability and efficiency of the model. In conclusion, the construction and implementation of the risk control model combine theoretical foundations, data processing, feature engineering, and machine learning algorithms. By following a systematic process and evaluation mechanism, the model demonstrates high predictive accuracy and interpretability in complex financial scenarios, providing robust technical support and decision-making insights for internet financial risk control [8].

3.4. Algorithm Selection and Model Training Optimization

The selection of algorithms is a critical step that directly affects the performance of risk control models. After thoroughly analyzing the characteristics of various machine learning algorithms, this study selects LightGBM (Light Gradient Boosting Machine) as the core algorithm. LightGBM is an optimized algorithm based on Gradient Boosting Decision Trees (GBDT), utilizing histogram-based algorithms and a leaf-wise growth strategy to achieve efficient training and accurate predictions. It is particularly suitable for large-scale datasets and high-dimensional financial risk control scenarios [9].

3.4.1. LightGBM Algorithm Principles

The core of LightGBM lies in its gradient boosting decision tree framework, where an additive model combines multiple decision trees to construct a robust classifier. The model is optimized by minimizing a loss function, as detailed below: The prediction result F(x)F(x) is represented as the weighted sum of multiple decision trees as shown in Formula 1:

$$\mathrm{F(x)} = \sum_{t=1}^{T} \mathrm{f_t(x)} + \mathrm{F_{t-1}}\,\mathrm{(x)} \tag{1}$$

Where F(x) is the final prediction value. $\mathrm{f_t(x)}$ represents the tt-th decision tree.T is the total number of trees, and $\mathrm{F_{t-1}(x)}$ is the cumulative prediction of the previous t−1t-1 trees.LightGBM optimizes the model by minimizing the negative gradient of the loss function. For a given dataset $\mathrm{D} = \{(\mathrm{x_i}, \mathrm{y_i})\}_{i=1}^{n}$, the loss function LL is defined as shown in Formula 2:

$$\mathrm{L} = \sum_{i=1}^{n} \mathrm{l}(\mathrm{y_i}, \mathrm{F(x_i)}) \tag{2}$$

Where $\mathrm{y_i}$ is the true label.$\mathrm{F(x_i)}$ is the prediction value. l represents the error function, such as mean squared error (MSE) or cross-entropy loss.LightGBM employs a "leaf-wise" growth strategy, where the leaf node with the maximum gain is split first. Compared to the traditional depth-wise method, leaf-wise growth reduces computational costs and improves training efficiency.The split gain measures the effectiveness of node splitting and is calculated as shown in Formula 3:

$$\mathrm{Gain} = \frac{(\mathrm{G_L^2/H_L}) + (\mathrm{G_R^2/H_R}) - (\mathrm{G^2/H})}{2} - \lambda \tag{3}$$

Where G and H are the first and second-order gradient accumulations. $\mathrm{G_L}$,$\mathrm{G_R}$,$\mathrm{H_L}$,$\mathrm{H_R}$ are the gradient accumulations for the left and right child nodes.λ is the regularization parameter used to prevent overfitting.

3.4.2. Model Training and Optimization

During the model training process, this study optimized LightGBM's performance through the following approaches:Data Processing and Feature Engineering: Feature Selection: Relevant features were retained using correlation analysis and Recursive Feature Elimination (RFE) methods to identify the most impactful predictors for risk assessment. Class Imbalance Handling: The SMOTE (Synthetic Minority Over-sampling Technique) method was applied to balance class distributions and improve the model's ability to detect minority class events. Feature Encoding: Categorical variables were transformed using One-Hot Encoding or Target Encoding for compatibility with machine learning models. Parameter Tuning: LightGBM's hyperparameters were tuned using a combination of grid search and random search to identify the optimal settings: Learning Rate (learning_rate): Controls the contribution of each tree to prevent overfitting. Tree Depth (max_depth): Limits the maximum depth of each tree to prevent over-complexity. Number of Leaves (num_leaves): Controls tree complexity and ensures generalization. Regularization Parameter (λ): Applies L2 regularization to mitigate overfitting risks. Training Process and Early Stopping: The training process employed cross-validation by dividing the dataset into training and validation subsets using k-fold cross-validation (e.g., k=5k=5) to evaluate generalization performance. To improve efficiency, an Early Stopping mechanism was integrated to terminate training when the validation loss did not decrease for a specified number of rounds, avoiding overfitting [10].

3.4.3. Model Performance Evaluation

This study evaluated model performance using metrics such as Accuracy, Recall, F1-Score, and AUC (Area Under the Curve). The specific formulas are as follows Formula 4-6:

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \qquad （4）$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \qquad （5）$$

$$\text{F1} = \frac{2\cdot\text{Precision}\cdot\text{Recall}}{\text{Precision+Recall}} \qquad （6）$$

AUC: Measures the model's ability to distinguish between positive and negative samples. A higher AUC indicates better model performance. Through training and validation, the LightGBM model demonstrated superior performance compared to traditional models such as Logistic Regression and SVM. On the test dataset, the AUC reached 0.92, while the F1-Score improved to 0.89. These results indicate that LightGBM effectively identifies risk events, making it highly applicable to internet financial risk control scenarios. By applying and optimizing the LightGBM algorithm, this study successfully achieved efficient identification and prediction of financial risks. The model demonstrated excellent generalization ability and stability during performance evaluation and validation, providing reliable technical support and decision-making insights for internet financial risk control.

5. Experiment and Results Analysis

To comprehensively verify the performance of the proposed LightGBM-based internet financial risk control model, this section conducts systematic experiments, including data preprocessing, feature engineering, model training and parameter tuning, performance comparison, and result analysis. By comparing the LightGBM model with traditional models such as logistic regression, Support Vector Machine (SVM), and random forest, the significant advantages of LightGBM in risk prediction tasks are demonstrated. Additionally, feature importance analysis reveals the key influencing factors for risk prediction, providing detailed experimental results and visualizations. The experimental data comes from actual internet financial risk scenarios, with a dataset containing 1 million records, including user credit data, historical risk records, transaction behavior data, and external macroeconomic indicators. The specific descriptions of the data are presented in Table 2:

Table 2. Composition and Characteristics of Experimental Datasets.

Data Category	Data Volume	Key Features	Time Range
Credit Risk Data	500,000 records	Historical overdue counts, credit scores, repayment periods	Jan 2020 - Dec 2023
Fraud Behavior Data	200,000 records	Abnormal login counts, device switch frequency, interrupted transactions	Jan 2021 - Dec 2023

User Transaction Data	200,000 records	Loan amounts, transaction frequency, transaction amounts	Jan 2022 - Dec 2023
External Economic Data	100,000 records	Unemployment rate, GDP growth rate, inflation rate	Jan 2019 - Dec 2023

After data cleaning, deduplication, missing value imputation, and anomaly removal, the data was randomly split into 70% for training, 15% for validation, and 15% for testing. To ensure optimal model performance, LightGBM was fine-tuned using grid search and five-fold cross-validation. The final optimized parameters are presented in Table 3:

Table 3. Optimal LightGBM Parameter Configuration.

Parameter	Value	Description
Learning Rate	0.05	Controls the step size of each tree to avoid overfitting.
Maximum Depth	7	Limits the maximum depth of each tree to control complexity.
Number of Leaves	31	Controls the maximum number of leaves per tree.
Regularization Parameter	1.0	L2 regularization weight to prevent overfitting.
Class Weight Balancing	True	Automatically balances class weights.
Number of Iterations	300	Maximum number of iterations for decision trees.

During training, Early Stopping was employed to terminate training when the validation AUC did not improve for 20 consecutive rounds. For comparison, logistic regression, SVM, and random forest were trained under identical experimental conditions with carefully tuned parameters to ensure fairness. The performance of LightGBM, logistic regression, SVM, and random forest was evaluated on the test set using Accuracy, Recall, F1-Score, and AUC metrics. The results are illustrated in Figure 4:

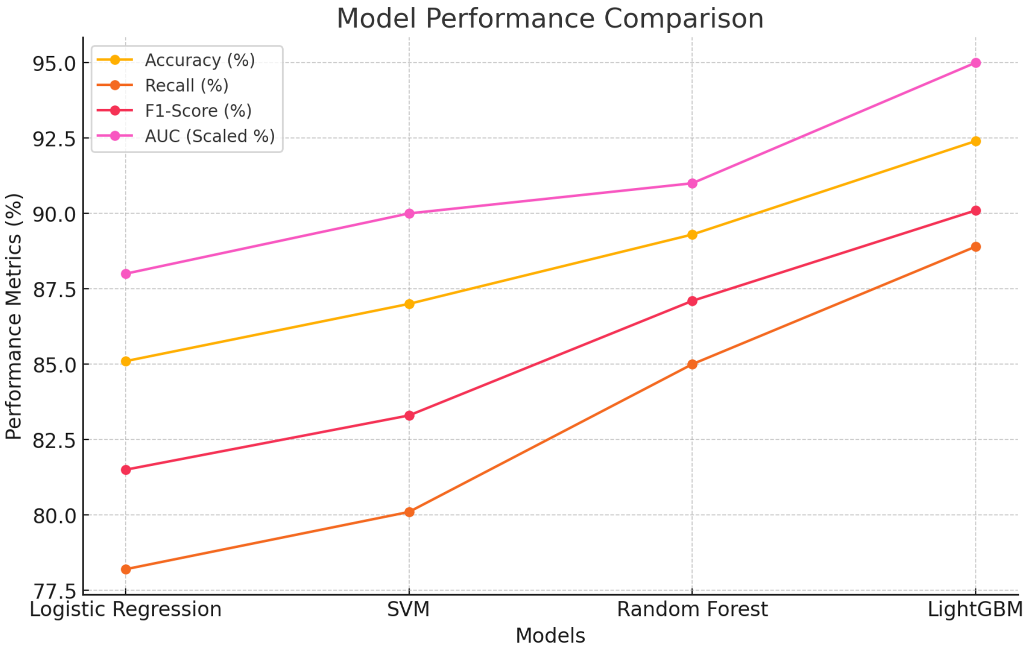

Figure 4. Model Performance Comparison.

The experimental results demonstrate that LightGBM outperformed all other models across all evaluation metrics. The accuracy of LightGBM reached 92.4%, while the recall and F1-Score achieved 88.9% and 90.1%, respectively, significantly exceeding the performance of logistic regression, SVM, and random forest. Particularly in the AUC metric, LightGBM achieved an outstanding value of 0.95, proving its efficiency and accuracy in risk identification tasks. Furthermore, the training time for LightGBM was only 8 seconds, which was considerably lower than that of SVM and random forest, showcasing its high computational efficiency. By analyzing the feature importance output from LightGBM, the key influencing factors for risk prediction were further revealed.

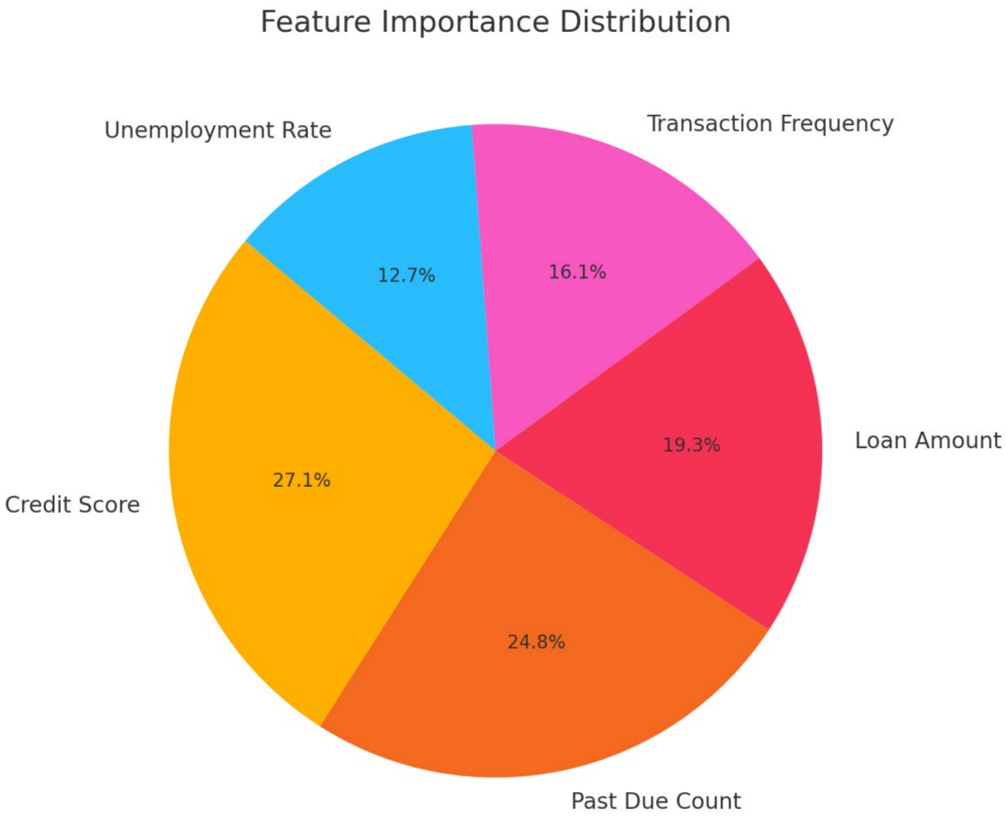

Figure 5. Feature Importance Distribution.

As shown in Figure 5, user credit scores and historical overdue counts are the most significant factors influencing risk prediction, contributing a combined 41.2% of the feature importance. This indicates that user credit behavior and historical repayment records are the core bases for risk control. Additionally, loan amounts and transaction frequency reflect users' liquidity conditions, while external macroeconomic indicators, such as the unemployment rate, reveal the impact of economic conditions on risk events. The combination of internal and external factors in feature selection enables a more comprehensive identification of risk events. Through systematic experiments and comparative analysis, this study verifies the superior performance of the LightGBM model in internet financial risk control tasks. Compared to traditional models, LightGBM demonstrates outstanding performance in accuracy, recall, and AUC, while also exhibiting significant computational efficiency. Furthermore, the feature importance analysis identifies the critical factors in risk control, providing financial institutions with scientific decision-making insights and optimization directions. The experimental results confirm that the LightGBM model can effectively identify potential risks, offering reliable technical support and decision-making foundations for internet financial platforms' risk management.

6. Conclusion

This study proposes a LightGBM-based risk control model for internet finance to address the limitations of traditional methods in handling complex and high-dimensional financial data. By systematically processing multi-dimensional data, including credit risk, user behavior, and macroeconomic indicators, the model demonstrates superior performance in risk identification and prediction. Through experiments and comparisons with logistic regression, SVM, and random forest, LightGBM achieves the highest accuracy (92.4%), recall (88.9%), and AUC (0.95), while maintaining high computational efficiency. Feature importance analysis highlights user credit scores and historical overdue counts as key predictors, alongside transaction behaviors and external economic factors, enabling a comprehensive understanding of risk patterns. In conclusion, the LightGBM model proves to be effective, reliable, and adaptable for internet financial risk control. Future research will focus on integrating multi-modal data and improving model interpretability to further enhance its applicability in dynamic financial scenarios.

References

1. A. Mashrur, et al., "Machine learning for financial risk management: A survey," *IEEE Access*, vol. 8, pp. 203203–203223, 2020.
2. O. A. Bello, "Machine learning algorithms for credit risk assessment: An economic and financial analysis," *Int. J. Manage.*, vol. 10, no. 1, pp. 109–133, 2023.
3. M. Liu, R. Gao, and W. Fu, "Analysis of internet financial risk control model based on machine learning algorithms," *J. Math.*, vol. 2021, no. 1, Art. no. 8541929, 2021.
4. Y. Cao, Y. Shao, and H. Zhang, "Study on early warning of E-commerce enterprise financial risk based on deep learning algorithm," *Electron. Commerce Res.*, vol. 22, no. 1, pp. 21–36, 2022.
5. T. Liu and Z. Yu, "The analysis of financial market risk based on machine learning and particle swarm optimization algorithm," *EURASIP J. Wireless Commun. Netw.*, vol. 2022, no. 1, Art. no. 31, 2022.
6. Y. Lei, H. Qiaoming, and Z. Tong, "Research on supply chain financial risk prevention based on machine learning," *Comput. Intell. Neurosci.*, vol. 2023, no. 1, Art. no. 6531154, 2023.
7. B. Gao, "The use of machine learning combined with data mining technology in financial risk prevention," *Comput. Econ.*, vol. 59, no. 4, pp. 1385–1405, 2022.
8. S. Chen, "Cryptocurrency financial risk analysis based on deep machine learning," *Complexity*, vol. 2022, no. 1, Art. no. 2611063, 2022.
9. M. R. Machado and S. Karray, "Applying hybrid machine learning algorithms to assess customer risk-adjusted revenue in the financial industry," *Electron. Commerce Res. Appl.*, vol. 56, Art. no. 101202, 2022.
10. Z. Shahbazi and Y.-C. Byun, "Machine learning-based analysis of cryptocurrency market financial risk management," *IEEE Access*, vol. 10, pp. 37848–37856, 2022.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of SOAP and/or the editor(s). SOAP and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.