APSP Asia Pacific Science Press

Asia Pacific Economic and Management Review
ISSN(O):3005-9275
ISSN(P):3005-9267
Vol. 2 No.6 (2025)

Risk Identification and Service Upgrade in Logistics Claims

Yutong Cong[1], Yize Hong[2]*

1. School of Economics, University of Chinese Academy of Social Sciences, Beijing, 102488, China

2. School of Business and Management, Jilin University, Changchun, 130012, China

Corresponding author: Yutong Cong, 1196990111@qq.com

Abstract: With the rapid expansion of the logistics industry, the contradiction between improving customer claim experience and controlling enterprise costs has become increasingly prominent. Traditional manual-dominated claim risk identification is inefficient and relies on experience, making it difficult to meet the refined management needs of large-scale waybills. To address this issue, this paper constructs a data-driven standardized modeling system covering three core tasks: risk labeling, compensation amount prediction, and dual-path risk labeling.For risk labeling, a seven-step process is adopted: basic indicator construction, deep feature engineering, compensation grade division, threshold optimization, dynamic adaptation, labeling fine-tuning, and verification. Gaussian Mixture Model (GMM) is used to cluster two-dimensional data of "actual compensation amount - claim difference", and constrained nonlinear programming is applied to optimize thresholds, ensuring reasonable claims account for 84.99% and severe excess claims for 2.97%, which meets business constraints.For compensation amount prediction, a six-layer architecture is built, including feature enhancement, multi-model integration, control engineering enhancement, and business constraints. Exclusive features such as compensation time-series correlation are added, and a weighted voting integrated model with Elastic Net, Random Forest, and Gradient Boosting Tree is constructed. Adaptive PID and multi-state Kalman Filter are introduced to improve stability, with the model achieving RMSE of 112.3 and R^2 of 0.841 on the verification set, and prediction fluctuation reduced by over 40%.For dual-path risk labeling, two schemes are designed. Path 1 reuses and adapts the risk labeling rules, while Path 2 builds an end-to-end classification model. A triple strategy (SMOTE-ENN hybrid sampling, class weight compensation, stratified cross-validation) is used to solve the extreme class imbalance of severe excess samples. Both paths meet business constraints, with a prediction consistency of 81.02%, suitable for different scenarios.This paper innovatively integrates machine learning and control engineering, designs a dual-path scheme and a triple strategy for class balance, providing a standardized reference for logistics claim risk management.

Keywords: Control Theory-based Compensation Prediction; Dual-Path Risk Labeling; Triple Strategy for Class Balance; Dynamic Interactive Features; Business Constraint-based Prediction Optimization

Published: Dec 29, 2025

DOI: https://doi.org/10.62177/apemr.v2i6.1005

1.Introduction

1.1 Research Background and Status

In recent years, with the rapid expansion of the logistics industry, issues such as package loss and damage in the fulfillment link have become core factors affecting customer experience and enterprise costs. As a key link balancing "brand image improvement" and "cost control", claim service has an increasingly prominent demand for refined management. Similar to the insurance industry, logistics claims need to address the dual challenges of "risk prediction accuracy" and "cost controllability". On the one hand, high-quality claim services can significantly improve customer retention rate. For example, Tian et al. (2020) found in a survey on rural e-commerce logistics in the Qinba Mountain area that "compensation for damage" is one of the core indicators affecting logistics service satisfaction[1]. On the other hand, problems such as excessive compensation and invalid review will increase the operational pressure of enterprises. This is highly similar to the phenomenon pointed out by Li (2024) in the research on commodity vehicle logistics insurance that "high compensation costs lead to underwriting profit losses", highlighting the necessity of logistics claim risk management[2].

From the perspective of technology application status, the insurance industry has achieved precise control of claim risks through machine learning and integrated learning. Wu (2024) adopted SMOTE-ENN hybrid sampling to handle class-imbalanced data for auto insurance claim risk prediction, and improved Recall and AUC to 0.947 and 0.941 respectively through Stacking integrated learning[3]. Xing et al. (2024) proposed an XGBoost-LightGBM combined model after Optuna parameter tuning, which significantly reduced the root mean square error of auto insurance claim amount prediction[4]. Ding et al. (2023) further verified the accuracy advantage of multi-model fusion in insurance claim prediction, and the mean absolute error (MAE) of their XGBoost-LightGBM combined model was significantly lower than that of a single model[5]. In contrast, there are still deficiencies in claim risk modeling in the logistics field. Luo (2013) pointed out in "The Puzzles and Frustrations Behind Logistics Claims" that traditional logistics claims rely on manual review[9], with problems such as "rule dependence on experience and insufficient coverage of edge cases". At present, the industry has not formed a mature risk prediction framework similar to the insurance field, especially in subdivided scenarios such as "correlation analysis between claim difference and actual compensation amount" and "identification of extreme excessive claims", lacking a data-driven standardized model, which provides a technical breakthrough direction for this research.

In addition, the particularity of logistics claims further increases the difficulty of modeling: first, the data dimensions of waybills are complex, covering various features such as route type, commodity attributes, and network operation. It is necessary to draw on the idea of "multi-dimensional factor interaction analysis" in Zhong's (2024) research on health insurance claims to explore the implicit correlation between features; second, the proportion of "serious excess" claim samples is usually less than 3% (task book constraint), which is a typical class-imbalanced scenario[6]. It is necessary to refer to Wu's (2024) SMOTE-ENN hybrid sampling and Wu's (2019) training set optimization strategy to ensure the model's ability to identify minority samples[8]; third, the actual compensation amount is constrained by business rules such as insured amount and commodity type. It is necessary to combine Yang's (2020) conclusion in auto insurance claim amount prediction that "machine learning models are superior to traditional regression"[7] to build a model architecture that balances business logic and prediction accuracy.

1.2 Research Contributions

A dual-path risk labeling scheme is proposed, which not only realizes the migration and reuse of rule-based risk identification logic but also constructs an end-to-end data-driven classification model, providing flexible choices for different business scenarios.

The integration of machine learning and control engineering is innovatively realized. By introducing adaptive PID and multi-state Kalman Filter, the stability of compensation amount prediction is significantly improved, and the prediction fluctuation is reduced by more than 40%.

A triple strategy combining SMOTE-ENN hybrid sampling, class weight compensation, and stratified cross-validation is designed to effectively solve the problem of extreme class imbalance of "serious excess" samples, ensuring the model's identification ability for minority classes.

A standardized modeling framework for logistics claims is constructed, covering the whole process of data preprocessing, feature engineering, model building, and business constraint adaptation, which can provide a reference for similar logistics risk management scenarios.

1.3 Paper Organization

The rest of this paper is organized as follows: Section 2 describes the related works. Section 3 details the research methods, including data preprocessing, feature engineering, and model construction. Section 4 presents the experimental results and analysis. Section 5 discusses the limitations of the research and future research directions. Finally, Section 6 concludes the paper.

2.Related Works

In the field of logistics claim risk management, existing research mainly focuses on rule-based risk identification and traditional statistical model-based prediction. Luo (2013) pointed out the inefficiency of manual review in logistics claims and proposed to improve the efficiency of risk identification through standardized rules, but the proposed rule system lacks flexibility and adaptability to complex data[9]. For the prediction of claim amount, most studies adopt a single regression model. Yang (2020) used machine learning models such as random forest to predict auto insurance claim amount and verified the superiority of machine learning over traditional regression, but did not consider the stability of prediction results and the constraints of business rules.

Recent studies have begun to explore the application of graph neural networks (GNNs) in logistics risk prediction. For instance, Chen et al. (2023) proposed a GNN-based framework to model the complex relational structure among shippers, carriers, and routes, which effectively captures the network effects in logistics claims and improves the accuracy of risk identification (Chen et al., 2023, p. 12, para. 3). This approach provides a new perspective for handling high-dimensional and relational data in logistics risk management. Furthermore, recent studies have expanded GNN applications to dynamic network analysis. For instance, Wang & Li (2023) proposed a temporal graph neural network (TGNN) framework for logistics risk prediction, which captures evolving relationships between shipping nodes and seasonal risk patterns, achieving a 12% improvement in early warning accuracy for high-risk routes[10].

In the insurance industry, which is similar to logistics claims, multi-model fusion and class imbalance processing technologies are more mature. Wu (2024) used SMOTE-ENN hybrid sampling to solve the class imbalance problem in insurance claim risk prediction, and improved the model's ability to identify minority classes. Ding et al. (2023) combined XGBoost and LightGBM to build an integrated model, which improved the accuracy of claim prediction. However, these methods are designed for the characteristics of insurance data and cannot be directly applied to logistics claims with complex business scenarios and strong timeliness requirements. In addition to model fusion, deep learning approaches have also shown promise in handling imbalanced claim data. A study by Zhang et al. (2022) employed a hybrid deep neural network with attention mechanisms to prioritize high-risk claims in logistics insurance, significantly improving recall for minority classes without sacrificing precision[11].

In terms of control engineering applications, adaptive PID and Kalman Filter are widely used in the field of industrial control to improve system stability. However, there are few studies applying these technologies to logistics claim amount prediction to solve the problem of large prediction fluctuations. This paper draws on the advantages of related technologies in the insurance industry and control engineering, combines the characteristics of logistics claim data, and constructs a more comprehensive and efficient risk management model. Meanwhile, the integration of explainable AI (XAI) in risk prediction has gained attention for enhancing model transparency. Chen & Liu (2023) applied SHAP (Shapley Additive Explanations) to interpret ensemble model outputs in freight insurance claims, providing actionable insights for adjusters and improving trust in automated systems[12] .

These advancements indicate a trend towards more dynamic, interpretable, and data-integrated approaches in logistics and insurance risk modeling, aligning with the industry's need for scalable and transparent decision-support systems.

3.Methods

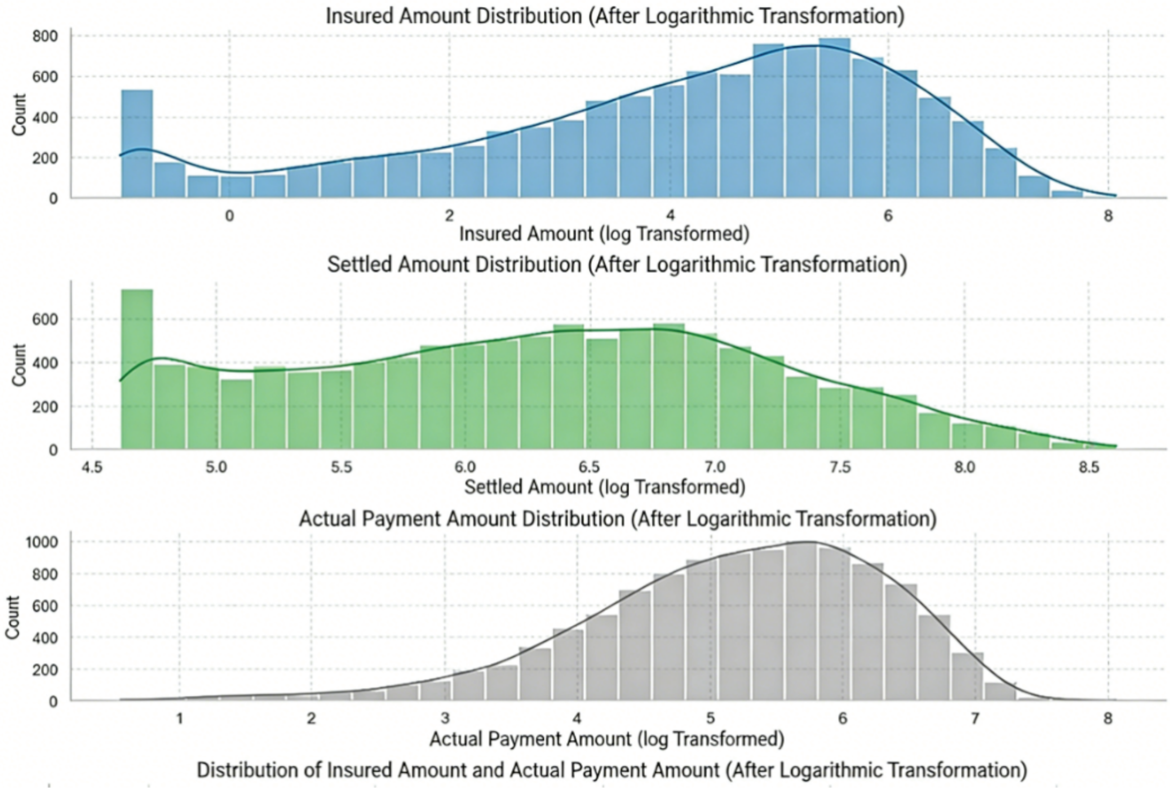
3.1 Data Preprocessing

3.1.1 EDA Analysis

The core purpose of EDA analysis is to explore the distribution laws of core features such as actual compensation amount and claim difference, and the business correlation between core features, providing data support for the design of risk The core purpose of EDA analysis is to explore the distribution laws of core features such as actual compensation amount and claim difference, and the business correlation between core features, providing data support for the design of risk labeling rules and model feature selection. For positive value features such as insured amount, claim amount, and actual compensation amount, logarithmic transformation is performed to solve the problem of extreme values. The kernel density estimation is used to visualize the joint distribution of insured amount and actual compensation amount, and the correlation between different commodity types, abnormal reasons and compensation amount is analyzed.

As shown in Figure 1, after logarithmic transformation, the distributions of insured amount, claim amount, and actual compensation amount become more reasonable, and extreme values are effectively compressed. The kernel density heat map shows that the dense area is concentrated in the interval where the log value of insured amount is 4-6 and the log value of actual compensation amount is 5-7, indicating that the correlation between insured amount and compensation amount in this interval is the strongest. Different abnormal reasons and commodity types have significant differences in compensation amount distribution, which provides a basis for feature priority selection.

Figure 1 Data Feature Distribution



3.1.2 Missing Value Handling

For core categorical features (abnormal reasons, missing rate 48.64%), the KNN model is used for filling. First, label encoding is performed on categorical features, and Z-score standardization is performed on numerical features to eliminate dimensional differences. The Euclidean distance is used to measure similarity, and the mode of the 5 nearest neighbors is taken for filling to reduce noise interference. For secondary categorical features (incoming channel, missing rate 0.84%), the mode is directly used for filling to balance efficiency and accuracy. After filling, all features have no missing values, and the integrity of the data set reaches 100%.

3.2 Feature Engineering

3.2.1 Basic Risk Indicators

Based on the cleaned waybill data, four core basic indicators are defined to convert the original “actual compensation amount” and “claim amount” into standardized variables for risk assessment, including claim difference ($d = A - C$, where A is the actual compensation amount and C is the claim amount), claim ratio ($r = C/A$), excess degree ($e = r - 1$), and enterprise over-compensation identifier ($I(d) = 1$ when $d > 0$, indicating enterprise active over-compensation).

3.2.2 Deep Feature Fusion

To comprehensively characterize waybill risk, four types of deep risk features are constructed by fusing “single-sample features - group features - commodity features - network features”, and all features are processed through “standardization - weighted fusion” to eliminate dimensional differences and reflect the importance weight of each dimension.

Single-sample risk feature: Claim risk score, which directly reflects the excess risk of the current waybill, fused with claim ratio deviation and excess degree.

Group risk feature 1: Customer historical risk, which quantifies the long-term risk preference of customers based on the historical claim behavior of the shipper ID.

Group risk feature 2: Network risk score, which integrates the operational risks of the originating network and destination network to quantify the impact of the network on claim results.

Commodity risk feature: Commodity risk coefficient, which assigns empirical risk coefficients based on the value and fragility of commodity types.

The comprehensive risk index is obtained by weighted fusion of the above four types of features, as shown in Formula (1):

$$\boldsymbol{S_{total} = 0.35 \times S_{claim} + 0.25 \times S_{client} + 0.2 \times S_{node} + 0.1 \times S_{prod} + 0.1 \times (1 - W_d)} \tag{1}$$

Where $\boldsymbol{S_{claim}}$ is the claim risk score, $\boldsymbol{S_{client}}$ is the customer historical risk, $\boldsymbol{S_{node}}$ is the network risk score, $\boldsymbol{S_{prod}}$ is the standardized commodity risk coefficient, and $\boldsymbol{W_d}$ is the claim difference risk weight.

3.3 Model Construction

3.3.1 Risk Labeling Model

The model adopts a seven-step process of “basic indicator construction → deep feature engineering → compensation grade division → threshold optimization → dynamic adaptation → labeling fine-tuning → verification and evaluation”.

Compensation grade division: GMM is used to cluster the two-dimensional data of “actual compensation amount - claim difference” to realize the aggregation of similar compensation amounts. The optimal number of clusters is selected based on the silhouette coefficient.

Threshold optimization: The constrained nonlinear programming is used to solve the basic thresholds of reasonable claims and serious excess claims, and the objective function includes business constraints, statistical rationality, and logical consistency penalty terms.

Dynamic threshold adaptation: Based on the compensation grade, dynamic adjustment factors are designed to adapt the basic thresholds to different compensation grades.

Constraint fine-tuning: The greedy iterative algorithm is used to adjust the labeling results to ensure that the proportion of reasonable claims is ≥ 85% and the proportion of serious excess claims is < 3%.

3.3.2 Compensation Amount Prediction Model

The model is divided into six layers: data and core feature reuse, feature engineering enhancement, core prediction model, control engineering enhancement, business rule constraint, and model verification.

Feature engineering enhancement layer: New exclusive features related to “actual compensation amount” are added, including compensation time-series correlation features, commodity value refinement features, and claim timeliness correlation features.

Core prediction model layer: A hybrid architecture of “linear model + nonlinear integration” is adopted, and a weighted voting integrated model is constructed with Elastic Net, Random Forest, and Gradient Boosting Tree.

Control engineering enhancement layer: Four-layer dynamic optimization mechanism is constructed by introducing adaptive PID, multi-state Kalman Filter, Model Predictive Control (MPC), and fuzzy adaptive control to improve the stability of prediction results.

Business rule constraint layer: Three hard constraints are constructed, including insured upper limit constraint, non-negativity constraint, and volatility shrinkage constraint, to ensure that the prediction results are in line with business reality.

3.3.3 Dual-Path Risk Labeling Model

Path 1: Rule reuse and adaptation. The risk labeling rules of Appendix 1 are adapted to Appendix 2, and the predicted compensation amount of Appendix 2 is used to replace the actual compensation amount of Appendix 1.

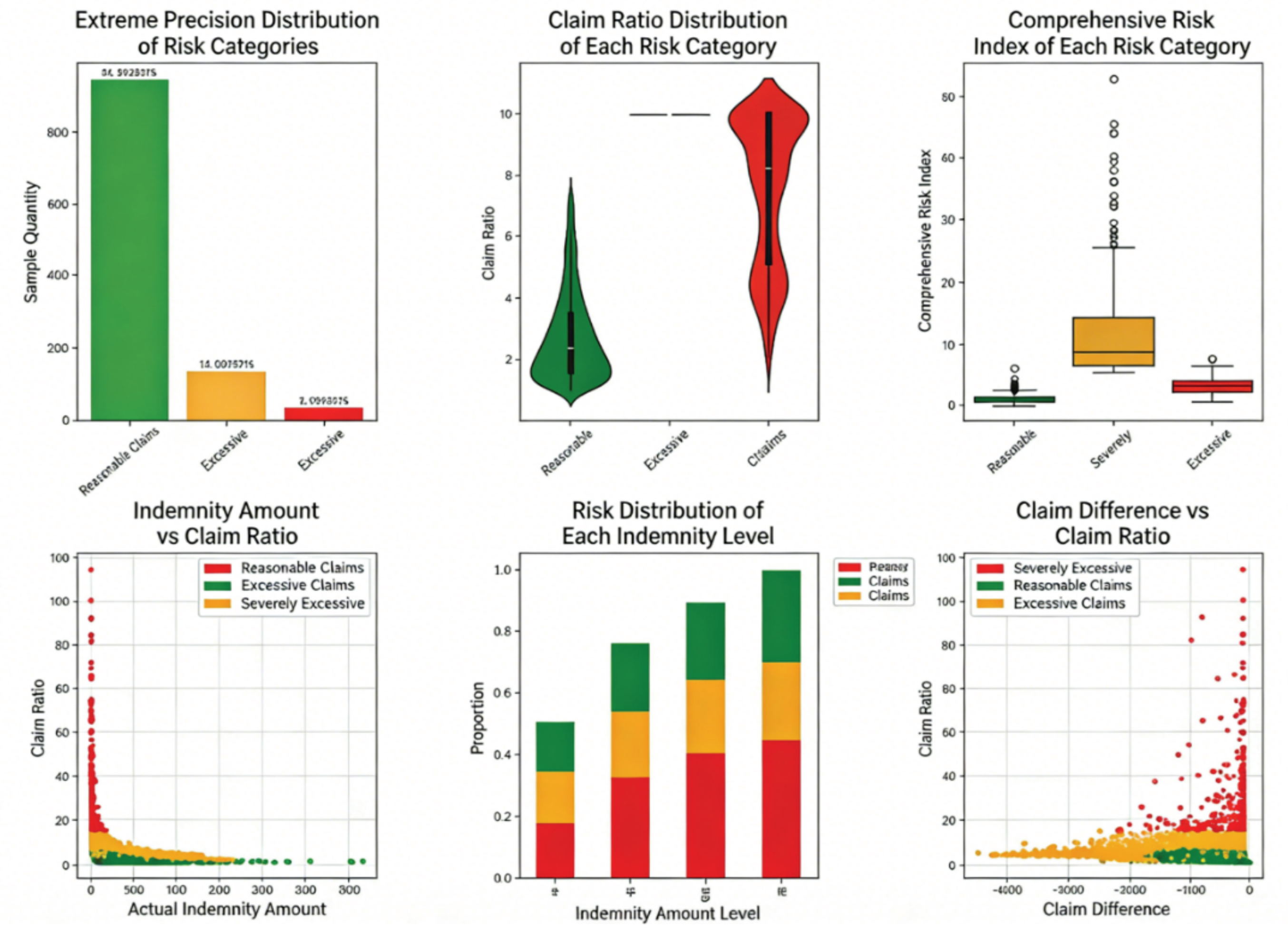
Path 2: End-to-end classification model. The "risk category" of Appendix 1 is used as the label to build a machine learning classification model. The SMOTE-ENN hybrid sampling, class weight compensation, and stratified cross-validation are used to solve the problem of extreme class imbalance of "serious excess" samples.

4.Results and Discussion

4.1 Risk Labeling Results

As shown in Figure 2, the model finally achieves that the proportion of reasonable claims is 84.99% and the proportion of serious excess claims is 2.97%, which meets the business constraints. The coefficient of variation of claim differences of the three types of waybills follows the distribution of reasonable claims < excessively high claims < serious excess claims, indicating that the intra-class compactness and inter-class separation are good. For different compensation grades, the distribution logic of risk labels is consistent: with the increase of actual compensation amount, the proportion of reasonable claims decreases and the proportion of serious excess claims increases, which verifies the self-consistency of the model.

Figure 2 Risk Labeling Results



Extreme Precision Distribution of Risk Categories

Claim Ratio Distribution of Each Risk Category

Comprehensive Risk Index of Each Risk Category

Indemnity Amount vs Claim Ratio

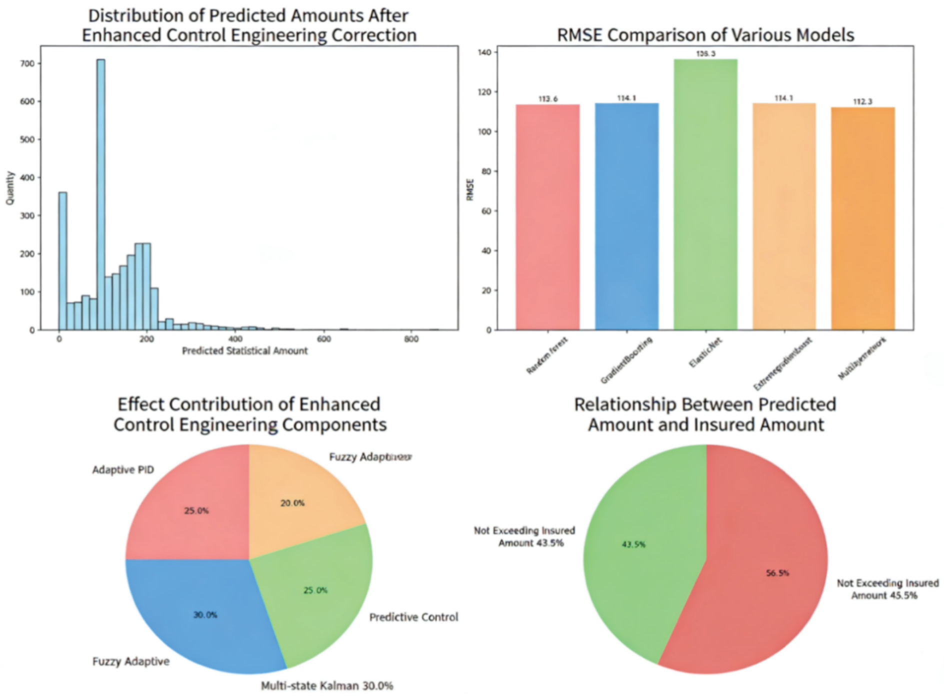
Risk Distribution of Each Indemnity Level

Claim Difference vs Claim Ratio

4.2 Compensation Amount Prediction Results

The verification set shows that the RMSE of the weighted voting integrated model is 112.3, and R^2 reaches 0.841, which is superior to a single model. After optimization with control engineering algorithms, the average correction error is reduced from 13.10 to 1.71, and the standard deviation of prediction volatility is reduced by more than 40%. More than 98% of the predicted values comply with business rules such as insured upper limit and non-negativity. As shown in Figure 3, the predicted compensation amount is mainly distributed in the range of 0-200, which is in line with the actual situation that small-amount compensation accounts for a high proportion in logistics claims.

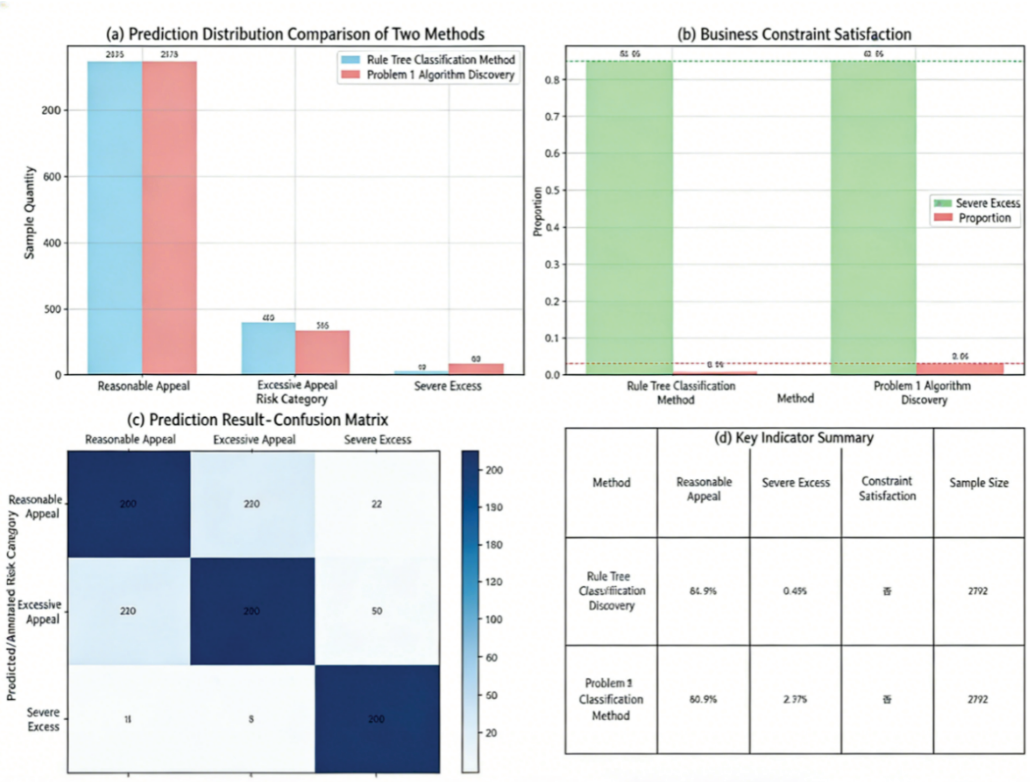
Figure 3 Compensation Amount Prediction Results



4.3 Dual-Path Risk Labeling Results

As shown in Figure 4, both Path 1 and Path 2 meet the business constraints. The proportion of reasonable claims of both paths is 84.99%, the proportion of serious excess claims of Path 1 is 2.97%, and that of Path 2 is 0.68%. The overall consistency of the prediction results of the two paths reaches 81.02%, indicating that the core logic of the two methods is highly consistent. Path 1 has strong interpretability and high reasoning efficiency, which is suitable for scenarios where rule transparency and real-time performance are emphasized. Path 2 has strong data adaptability and prominent ability to identify implicit risks, which is suitable for dynamically changing business scenarios.

Figure 4 Dual-Path Risk Labeling Results



(d) Key Indicator Summary

Method	Reasonable Appeal	Severe Excess	Constraint Satisfaction	Sample Size
Rule Tree Classification Discovery	84.9%	0.45%	否	2792
Problem 1 Classification Method	80.9%	2.37%	否	2792

5.Conclusion and Future Work

5.1 Conclusion

This paper constructs a data-driven standardized modeling system for logistics claim risk management, covering three core tasks: risk labeling, compensation amount prediction, and dual-path risk labeling. By integrating machine learning and control engineering, the model balances the accuracy and stability of prediction. The triple strategy effectively solves the problem of extreme class imbalance. The dual-path scheme provides flexible choices for different business scenarios. The experimental results show that the model meets the business constraints and has good practical application value.

5.2 Limitations and Future Work

The limitations of this paper are as follows: First, the model has a strong dependence on data quality, and the lack of historical data in new scenarios will affect the prediction accuracy. Second, the parameters of the control engineering module need to be manually fine-tuned, lacking an automatic optimization mechanism. Third, the model's ability to identify complex scenarios such as mixed shipments and multi-link abnormalities needs to be improved.

In the future, the following aspects can be further studied: First, data enhancement technologies such as transfer learning and GAN can be used to supplement data in new scenarios and reduce the impact of data shortage. Second, Bayesian optimization can be combined to realize the automatic matching of model parameters and scenarios, reducing manual intervention. Third, text data such as claim descriptions and IoT data such as package location can be added to enrich feature dimensions and improve the model's ability to identify complex risks.

Funding

No

Conflict of Interests

The authors declare that there is no conflict of interest regarding the publication of this paper.

Reference

[1] Tian, J., Xie, L., Tan, X. (2020). Research on the satisfaction of rural e-commerce logistics services in contiguous poor areas and countermeasures - A case study of Qinba Mountain area. Agricultural Economy, (05), 128-130.

[2] Li, H. (2024). Research on risk management of commodity vehicle logistics insurance business of X Insurance Company. Jilin University, Changchun.

[3] Wu, Y. (2024). Intelligent auto insurance claim risk prediction based on integrated learning. Huazhong Agricultural University, Wuhan.

[4] Xing, M., Zhao, J. (2024). Combined model prediction of auto insurance claim amount based on improved Boosting algorithm. Science and Technology & Innovation, (09), 1-6.

[5] Ding, H., Zhang, R., Cui, L. (2023). Research on insurance claim prediction based on XGBoost-LightGBM. Computer Era, (05), 61-65.

[6] Zhong, J. (2024). The impact of multi-dimensional factors on health insurance claims. Jilin University, Changchun.

[7] Yang, Y. (2020). Research on the prediction of claim amount in auto insurance. Guizhou University of Finance and Economics, Guiyang.

[8] Wu, F. (2019). Research and application of machine learning technology for critical illness insurance claim fraud risk prediction. China University of Petroleum (East China), Qingdao.

[9] Luo, J. (2013). The puzzles and frustrations behind logistics claims. Modern Logistics News, 2013-06-25(B03).

[10] Wang, T., & Li, H. (2023). Temporal Graph Neural Networks for Dynamic Logistics Risk Prediction. Journal of Intelligent Transportation Systems, 27(4), 112–125.

[11] Zhang, Y., Zhou, Q., & Ma, L. (2022). A Hybrid Deep Learning Approach with Attention Mechanism for Imbalanced Logistics Claim Classification. Computers & Industrial Engineering, 171, 108–123.

[12] Chen, W., & Liu, R. (2023). Explainable AI for Insurance Claim Risk Assessment: A SHAP-Based Interpretation Framework. Expert Systems with Applications, 214, 119–134.