Textile & Leather Review

ISSN 2623-6281 | www.tlr-journal.com | 10.31881/TLR

Risk Labeling and Classification Prediction of Logistics Claims Based on Machine Learning

Yu Chen, Cheng Liang, Biaoyong Liang

How to cite: Chen Y, Liang C, Liang B. Risk Labeling and Classification Prediction of Logistics Claims Based on Machine Learning. Textile & Leather Review. 2026; 9:3073-3095. https://doi.org/10.31881/TLR.2026.3073

How to link https://doi.org/10.31881/TLR.2026.3073

Published:25 April 2026

Risk Labeling and Classification Prediction of Logistics Claims Based on Machine Learning

Yu Chen*, Cheng Liang, Biaoyong Liang

School of Mathematics and Information Science, Guangzhou University, Guangzhou 510006, Guangdong, China
*18148959528@163.com

Article
https://doi.org/10.31881/TLR.2026.3073
Published 25 April 2026

ABSTRACT

Targeting the scenario of logistics claim governance, this paper proposes a risk labeling and prediction framework that incorporates business proportion constraints. First, missing values are processed, and two core features—"claim gap" and "claim amount ratio"—are constructed and standardized. Subsequently, a capacitated clustering method is employed to generate three risk labels. Under the hard constraints of "reasonable demands proportion ≥ 85%" and "serious excess proportion ≤ 3%", the sample proportions of the three categories in the final clustering solution are 0.850004, 0.119996, and 0.029999, respectively. Secondly, with the "actual compensation amount" as the regression target, models including Linear Regression, Decision Tree, XGBoost, LightGBM, CatBoost, and Random Forest are compared. The Random Forest model achieves R^2=0.7706, RMSE=135.5171, and MAE=92.1084 on the test set, and is utilized for the first-stage prediction in indirect classification. Finally, two risk classification routes are compared: direct classification (training classifiers using clustering labels as supervision signals) and indirect classification (predicting the compensation amount first, reconstructing the two-dimensional features, and outputting risk categories based on the nearest-centroid rule). The results indicate that indirect classification maintains consistency with the labeling rules while exhibiting more stable comprehensive performance and better business interpretability.

KEYWORDS

k-means clustering, linear regression, random forest, decision tree, XGBoost

INTRODUCTION

After experiencing rapid expansion in the past few years, improving customer service and experience has become a core issue for logistics companies. In the process of logistics transportation, there are often inevitable problems such as package loss and damage, and the corresponding logistics companies not only

need to make corresponding compensation; Efforts should also be made to control the cost of claims and ensure the sustainable and healthy development of the enterprise.

Claims services are usually divided into two parts: pre-sales and after-sales. The main goal of the pre-sales stage is to identify risks in advance and provide customers with more suitable physical services to improve user experience; The main goal of the after-sales process is to control the cost of claims, which can not only ensure reasonable claims, but also reduce the occurrence of excessive claims.

Current research mainly focuses on three directions: first, forecasting logistics volume and demand; second, clustering analysis based on customer information and waybill types; and third, supervised classification for risk identification. Although these methods provide a certain methodological foundation for logistics scenarios, they remain insufficient for the specific context of claim governance. This is because these approaches primarily focus on optimizing statistical metrics and generally do not directly incorporate business proportion constraints during the labeling stage. As a result, while the final models may exhibit good fitting capabilities, they fail to meet the actual needs of enterprise governance. Furthermore, existing work often discusses compensation regression and risk classification separately, lacking a systematic research framework that integrates compensation estimation, risk labeling, and decision interpretation into a unified pipeline.

In actual operations, to meet these specific governance requirements, after the customer files a claim, the enterprise needs to first judge its reasonableness, mainly based on the difference in the claim, and label it as “reasonable demand”, “high demand”, and “serious excess”, and take different measures to deal with them. And it is required that the proportion of “reasonable demands” waybills should not be less than 85%, and the proportion of “seriously excessive” waybills should be less than 3%. In general, the claim differences for “reasonable demands” waybills are relatively dense, while the claim differences for “seriously excessive” waybills are relatively sparse, and there are significant differences in the claim differences between different types of waybills.

It is worth noting that the three risk labels in this study are not ground-truth labels manually verified for each waybill, but rather governance labels generated by a capacitated clustering model combined with business proportion constraints and historical data distribution. Therefore, the subsequent modeling and evaluation of direct and indirect classification approach aim to compare the stability and interpretability of the two prediction routes in reproducing this labeling rule, rather than claiming to predict absolute true manual labels.

To address the aforementioned limitations and business-related constraints, this paper constructs an integrated modeling framework tailored for logistics claim governance. Specifically, it explicitly introduces proportion constraints during the risk labeling stage. Moreover, under unified data preprocessing and feature engineering conditions, this study jointly compares the direct and indirect prediction routes. The application of this framework successfully balances consistency with governance rules, model interpretability, and feasibility for practical deployment.

After risk labeling, the results obtained from the labeling can also be used to analyze future claims for waybills, automatically classify them, and control the risks and costs of claims; It can also be used for reverse optimization of the pre-sales logistics process. Before the customer places an order, the logistics service provider can analyze the waybill in advance based on the user's historical shipping and claims behavior, and provide corresponding services. For a large number of waybills, this model can provide reasonable protection for important goods of high-quality customers.

THE DIFFERENCE AND PROPORTION OF WAYBILL CLAIMS BASED ON CLUSTERING METHOD

Coordination and standardization of feature scales

The dimensions of the claim difference and the claim amount ratio are different, and in general, there is a significant difference in the magnitude of the two values. If the Euclidean distance between the two is directly calculated in the original space, it is easy to lead to a single dimension dominating, causing the update of the cluster center to be mainly affected by a single condition, which in turn leads to incomplete model consideration. This contradicts the description that 'the larger the claim difference and the lower the claim amount ratio, the more likely it is to be marked as high or severely exceeded'. Therefore, it is necessary to improve the comparability of the two conditions through standardization in the early stages of model establishment, so that the distance can be significantly affected by the joint influence of the two indicators.

Define standardized vectors [1-2], the application of normalization methods can be referred to References [1–2]:

$$\bar{z}_i = \left(\frac{x_i - \mu_x}{\sigma_x}, \frac{y_i - \mu_y}{\sigma_y} \right)^T \tag{1}$$

Among them, μ_x and σ_x represent the sample mean and standard deviation of the claim difference; μ_y , σ_y represent the sample mean and standard deviation of the claim amount ratio.

After this transformation, the scale before and after can be maintained uniformly, and the two components after the standard have zero mean and unit variance, so that the square clustering in the clustering objective function can reflect the meaning of "how many times is the distance from the standard deviation", and the statistical explanation is reasonable; At the same time, it can also maintain a linear structure, as standardization does not change the relative order and linear relationship of samples in two-dimensional space, thus not affecting the capture of trends by clustering; And it conforms to linear reversibility, because the standard process is a linear transformation, the problem may need to be solved at the scale in the future, and the inverse transformation can be completed by converting formula (2). Overall, this model can ensure that the clustering objectives truly reflect the combined effect of claim differences and claim amount ratios.

Data feature analysis

From Figure 1, it can be seen that the vast majority of the sample set is around 0 and slightly biased towards the negative side, showing a long tail feature extending to the left, indicating that there are generally cases where the compensation is lower than the claim, and there are a small number of extreme negative balance cases in the waybills. The scatter plot in Figure 2 depicts the relationship between the actual compensation amount and the claim difference, with the red segmented mean line showing a monotonic downward trend; As the compensation amount increases, the average claim amount continues to decline, indicating that there are also significantly reduced waybills hidden in the high compensation range, reflecting the strong differentiation of "samples with low claim amount ratios and large absolute differences".

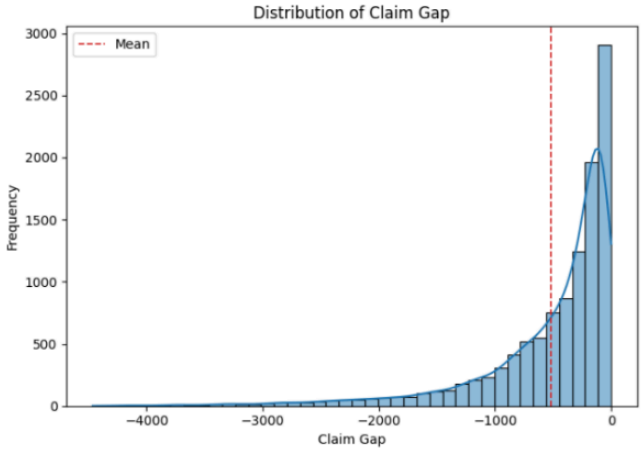

Figure 1. Distribution of Claim Gap

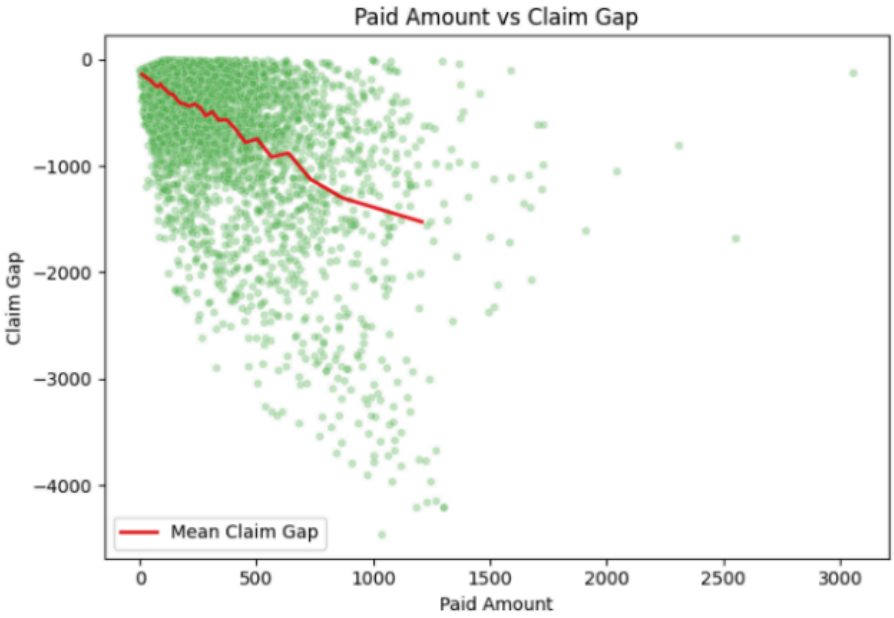

Figure 2. Paid Amount vs Claim Gap

Using mathematical methods to characterize clustering objectives

The basic goal of the model is to minimize the sum of squared distances within the class, that is:

$$J = \sum_{k=1}^{3} \sum_{\bar{z}_i \in C_k} \| \bar{z}_i - c_k \|_2^2 \tag{2}$$

among them, $\{c_k\}_{k=1}^{3}$ represents the cluster centers in the standardized space, and C_k corresponds to the three risk categories: C_1 (reasonable demands), C_2 (high demand), and C_3 (serious excess). By minimizing the objective function J under capacity constraints, the optimal partition with the minimum within-cluster sum of squares can be obtained.

Mapping constraints to model parameters

To strictly enforce hard constraints, they can be transformed into upper and lower bounds on category capacity. Assuming N is the total sample size, the minimum capacity of reasonable demand C_1 is $\underline{n}_1 = [0.85N]$; The maximum capacity of severe excess C_3 is $\overline{n_3} = [0.03N]$; To ensure the feasibility of integer programming, a technical lower bound $\underline{n}_3 = 1$ is also set. The capacity range corresponding to the high demand C_2 is determined by the remaining sample size, denoted as $[\underline{n_2}, \overline{n_2}]$. among which,

$$\underline{n}_2 = \max\{0, N - \overline{n}_3 - \overline{n}_1\} \tag{3}$$

$$\overline{n}_2 = N - \underline{n}_1 - \underline{n}_3 \tag{4}$$

To ensure feasible solutions, $\underline{n}_2$ and $\overline{n}_2$ can be moderately fine tuned in practical use to satisfy the overall condition $\sum_k \underline{n}_k \le N \le \sum_k \overline{n}_k$.

After setting the capacity, the hard constraints are transformed into mathematical constraints, and the output of the model will automatically meet the proportion requirements, and a bounded feasible region will be constructed to avoid wireless expansion of a certain category during the iteration process.

Decision variables for model allocation

To formalize the classification, define the indicator variable u_{ik} : $u_{ik} = 1$ when sample i is assigned to category k, otherwise it is 0. The variable system needs to satisfy the constraint of unique allocation $\sum_{k=1}^{3} u_{ik} = 1$, ensuring that each sample belongs to only one category; At the same time, it must meet its corresponding capacity limit, namely $\underline{n}_k \le \sum_{i=1}^{N} u_{ik} \le \overline{n}_k$.

After introducing decision variables, the objective function can be transformed into a discrete optimization form with binary variables:

$$\min_{\{u_{ik}\},\{c_k\}} \sum_{i=1}^{N} \sum_{k=1}^{3} u_{ik} d_{ik} \tag{5}$$

Among which, $d_{ik} = \| \bar{z}_i - c_k \|_2^2$.

This expression reveals the essence of the model: selecting the optimal allocation under capacity constraints to minimize the overall distance cost.

Constrained clustering algorithm and pseudocode

In this study, the capacitated K-means algorithm is employed for risk labeling. The core mechanism lies in computing the optimal allocation that satisfies the upper and lower capacity bounds via 0-1 integer programming while adjusting assignments with fixed cluster centers. Subsequently, the cluster centers are updated based on the newly obtained assignment, and this process iterates until convergence. In practical implementation, the integer programming is resolved utilizing the CBC solver provided by OR-Tools. The pseudocode is presented as follows:

Algorithm 1 Capacitated K-means with hard size bounds

Input: standardized samples {z_i}_{i=1..N}, K=3, bounds [n_k^min, n_k^max], max_iter, tol

Initialize centers {c_k}_{k=1..K} by random sampling

for t = 1.max_iter do

Solve assignment (0-1 integer program):

minimize Σ_i Σ_k u_{ik} ||z_i - c_k||_2^2

subject to Σ_k u_{ik} = 1, ∀i

n_k^min ≤ Σ_i u_{ik} ≤ n_k^max, ∀k

u_{ik} ∈ {0,1}

Update centers: c_k ← mean({z_i | u_{ik}=1})

if ||C_new - C_old||_F ≤ tol then break

end for

Output: labels, centers

Model solving and result analysis

After convergence, the results obtained are shown in Table 1, and it can be observed that each category precisely satisfies the set capacity interval. And the iteration stops in the 8th round, and the corresponding sum of squares within the class, that is, the overall error value, is 12276.74. When converted to a single sample value, it is about 1.10, which is within a reasonable range; By observing a set of coordinate values obtained by performing anti standardization on the three cluster centers that have converged, as shown in Figure 3, it can be intuitively seen that the claim difference significantly decreases with the risk level.

Table 1. Risk Label Sample Size and Proportion Constraints

risk label	Minimum sample size	Maximum sample size	Minimum proportion	largest proportion
Reasonable demands	9492	10608	0.850004	0.949942
High demand	558	1340	0.049969	0.119996
Serious excess	1	335	0.000090	0.029999

Total number of samples: 11167, sum of constraints: minimum 10051≤N, maximum 12283

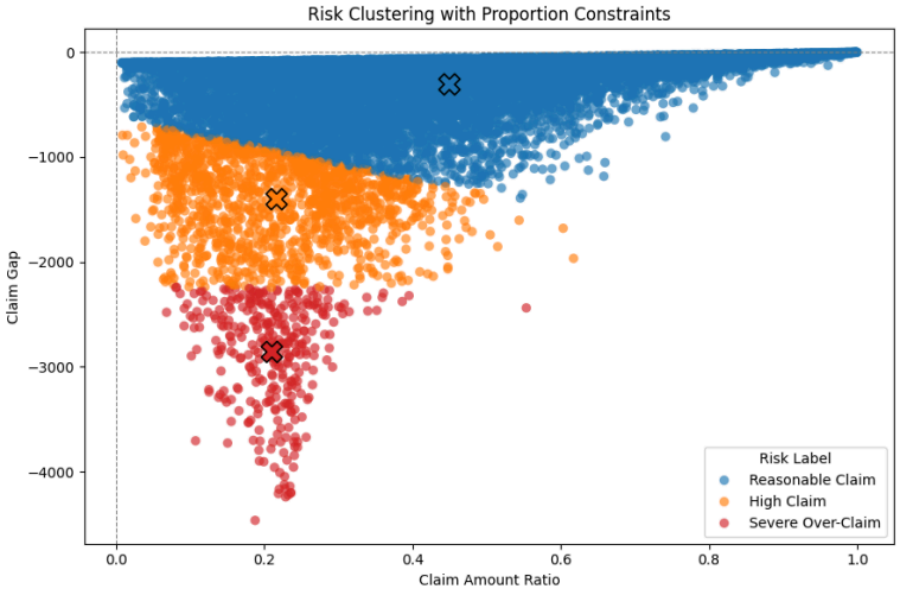

Figure 3. Risk Clustering with Proportion Constraints

Based on the running results, explain how to divide the waybill according to the annotation rules: first, calculate the ratio of claim difference and claim amount in each waybill according to the definition in data preprocessing, then standardize it, convert the standardized parameters into zero mean and unit variance space, and finally substitute them into the convergence to obtain the cluster center and corresponding upper and lower bounds of capacity. Directly map the assigned results to three risk labels: “reasonable demand”, “high demand”, and “serious excess”.

After inspection, the results are shown in Tables 2 and 3. Observation shows that the overall silhouette coefficient [3-4] is 0.3883(The application of the K-Means clustering algorithm can be found in References [3–4]), and the average silhouette coefficients for the three categories are 0.5728, 0.476, and 0.3694, respectively, indicating that high-risk groups have more compact clusters. After further BF and FK tests, the corresponding p-values were 3.22×10^{-119} and 2.53×10^{-114} , respectively, which rejected the null hypothesis of homogeneity of variance. Finally, the standard deviation, median absolute deviation, and interquartile range calculated based on risk labels were observed to further quantify the dispersion differences in various types of claims. Therefore, it can be explained that the higher the risk level, the more uncontrollable the corresponding compensation results.

Table 2. Silhouette Coefficient Statistics by Risk Label

Risk label	Average silhouette coefficient	Median	Minimum	Maximum	Sample size
Reasonable demands	0.5728	0.6207	0.2181	0.7074	335
High demand	0.3694	0.4632	−0.4445	0.5864	9492
Serious excess	0.4762	0.5244	−0.2727	0.6881	1340

Overall silhouette coefficient: 0.3883

Table 3. Dispersion Statistics by Risk Label

Risk label	standard deviation	Median absolute deviation	interquartile range	sample size
Reasonable demands	504.0214	398.3759	741.0750	335
High demand	267.3087	206.6610	369.7200	9492
Serious excess	375.5755	309.1316	547.3375	1340

Brown Forsyth * 1: F=279.6154, p=3.2224e-119 Flegner Kielen x ²=523.1352, p=2.5271e-114

Finally, a structural hypothesis test can be conducted on the capacity setting, which requires the sum of upper bounds to cover all samples while ensuring that the sum of lower bounds does not exceed the total sample size, and that the capacity interval for each category is not empty. If there is a violation, the model has no feasible solution under the current setting. Combining the data in Table 1 and substituting it into the hypothesis testing process, it can be proven that the hypothesis holds and there is no contradiction in the capacity setting. Combining the contour coefficients, variance tests, hypothesis tests, and mutual confirmation of decision variables obtained from the solving process, it is shown that the model not only meets the proportion requirements of the problem, but also retains significant statistical differences between risk levels.

PREDICTING ACTUAL CLAIMS BASED ON HISTORICAL WAYBILLS

After data preprocessing, the high cardinality fields such as the origin city, destination city, and reason for anomalies are first subjected to a single hot transformation, that is, uniformly transformed into a sparse indicator matrix; Then, by truncating singular value decomposition, the sparse matrix is reduced to lower dimensions, compressing redundancy while preserving the main information. In implementation, TruncatedSVD from scikit-learn is utilized to directly reduce the dimensionality of the sparse matrix (without the need for densification), and a heuristic upper bound is adopted to select the embedding dimension: n_components

is set to not exceed 64 and not exceed the one-hot dimension. For low cardinality categories, their exclusive hot representation is directly retained. After obtaining the embedded features, merge them with the original numerical features to form a unified modeling data input. And in the subsequent model building, variables with high contribution can be selected to reduce feature dimensions, and computing resources can be focused on truly effective information.

Multi model construction

To explore the relationship between the input features and the target variable in this article, multiple model algorithms were introduced for comparison, and the optimal model was selected to solve the problem. The following six methods were specifically chosen:

linear regression model

Linear regression [5-6] assumes a linear relationship between the independent variable and the objective(The application of multiple linear regression can be referred to References [5–6]), and estimates the parameters by minimizing the mean square error, i.e

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ...\beta_n x_n \tag{6}$$

Among them, $\hat{y}$ represents the predicted compensation amount, x_i represents the i-th feature, and β_i represents the regression coefficient. This model has a simple and easy to explain structure, which can serve as a cardinality reference for subsequent nonlinear models.

Decision Tree Regression Model

Decision tree regression [7-8] achieves the minimum intra node variance by recursively partitioning the feature space (The application of decision trees can be referred to References [7–8])

$$\min_{s,t_L,t_R} \left[\sum_{x_i \in t_L} (y_i - \bar{t}_L)^2 + \sum_{x_i \in t_R} (y_i - \bar{t}_R)^2 \right] \tag{7}$$

where s represents the optimal partition threshold, t_L and t_R respectively represent the sample sets of the left and right child nodes.

In this model, a single tree can capture non-linear and feature interactions, but overfitting is prone to occur in complex noise situations.

XGBoost Regression Model

XGBoost regression [9-10] uses the gradient boosting framework and combines second-order Taylor expansion to optimize the objective function(The application of XGBoost can be referred to References [9–10]), that is

$$L^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) \tag{8}$$

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{9}$$

Among them, $l()$ represents the loss function, f_t represents the t-th regression tree, $\hat{y}_i^{(t-1)}$ represents the prediction of the previous iteration, w_j represents the leaf node weight, T represents the number of leaf nodes, and γ and λ are used to control the regularization of the structure and weight.

This model can accurately control the model complexity in each iteration, making XGBoost more robust to noise.

LightGBM regression model

The LightGBM model [11-12] is based on a leaf node first growth strategy and uses gradient histograms to accelerate splitting (The application of LightGBM can be referred to References [11–12]). A typical split gain calculation is as follows:

$$extGain = \frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda} - \gamma \tag{10}$$

Among them, G_L and G_R respectively represent the gradient sum of left and right nodes, H_L and H_R represent the corresponding Hessian sum, λ controls the weight regularization of leaf nodes, and γ is the splitting penalty.

This model reduces the cost of performing high-dimensional sparse features while maintaining accuracy.

CatBoost regression model

The CatBoost model [13-14] is designed to predict categorical variables using ordered boosting and symmetric decision tree structures (The application of CatBoost can be referred to References [13–14]), which can be expressed as:

$$\hat{y} = \sum_{t=1}^{M} \eta_t b_t(x) \tag{11}$$

Among them, $b_t(x)$ represents the output of the t-th symmetric tree on sample x, η_t is the learning rate coefficient, and M is the total number of trees.

This model effectively alleviates the bias problem caused by category encoding by sorting target statistics and ordered samples.

Random Forest Regression Model

Random forest regression [15-16] is the process of taking the average of the results from multiple decision trees (The application of the random forest algorithm can be referred to References [15–16]), i.e.:

$$\hat{y} = \frac{1}{M} \sum_{m=1}^{M} T_m(x) \tag{12}$$

Among them, M represents the number of trees in the forest, and $T_m(x)$ represents the predicted value of the mth tree for sample x.

This model has the advantages of suppressing overfitting and measuring feature importance by utilizing self sampling and random sub feature mechanisms [17-18] (The application of machine learning classification can be referred to References [17–18]).

Build model evaluation indicators

To unify the measurement of predictive performance and business interpretability, three commonly used regression indicators [19-20] are selected for evaluation (The application of these three evaluation models can be referred to References [19–20]):

Determination coefficient R^2

The coefficient of determination R^2 is used to evaluate the explanatory power of the model on the target variance, and its formula is as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{13}$$

Among them, $\bar{y}$ represents the average compensation amount of the sample, y_i represents the true value, and $\hat{y}_i$ represents the predicted value. The closer the R^2 is to 1, the more fully the model explains the difference in compensation. The determination coefficients R^2 corresponding to the six models are shown in Figure 4.

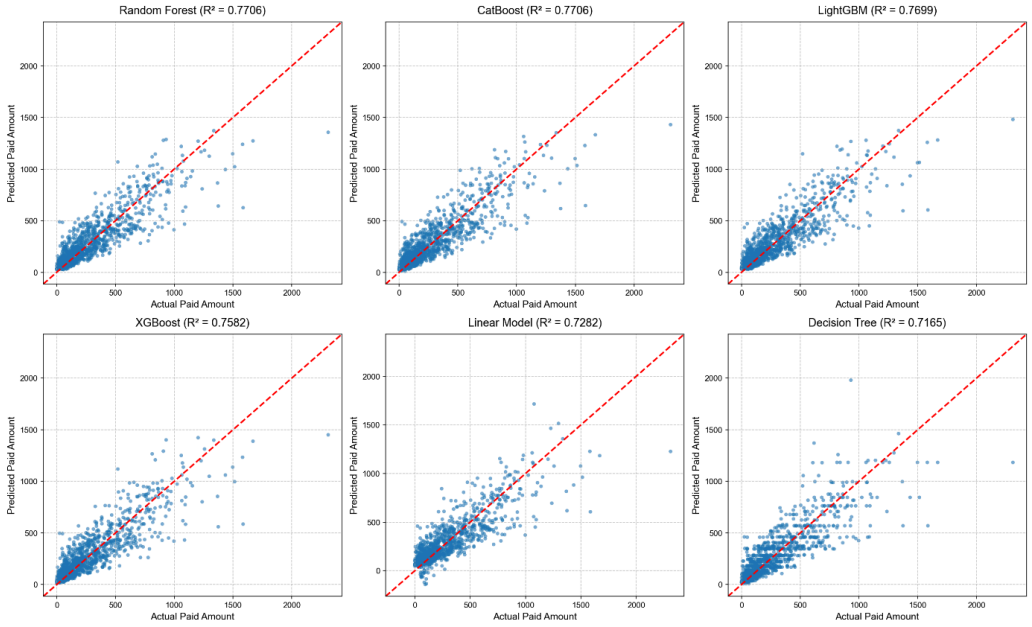

Figure 4. R^2 Fit

Root Mean Square Error (RMSE)

RMSE is used to measure the square average magnitude of errors, and the corresponding formula is as follows:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{14}$$

Among them, n is the sample size. RMSE is sensitive to large deviations and can help identify waybills with excess payout risks. The RMSE corresponding to the six models is shown in Figure 5.

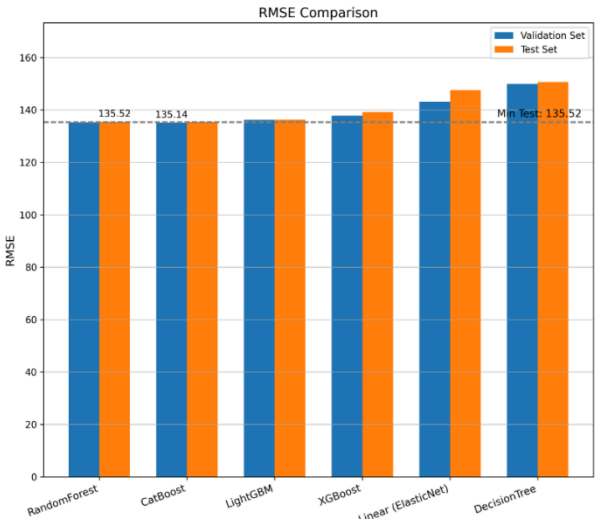

Figure 5. RMSE Comparison

Mean Absolute Error (MAE)

MAE is used to measure the average absolute value of errors, and its corresponding formula is as follows:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}| y_i - \hat{y}_i | \tag{15}$$

Among them, the absolute value term is used to reflect the degree of deviation of a single sample, making it easier to understand the performance of the model on average deviation. The MSE corresponding to the six models is shown in Figure 6.

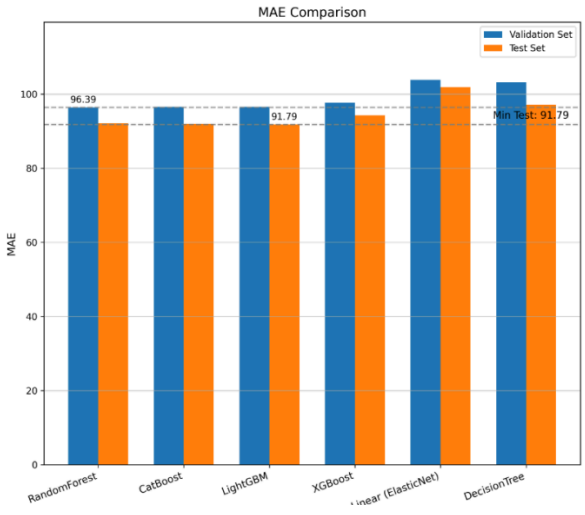

Figure 6. MAE Comparison

Based on the evaluation results of the comprehensive validation set and test set, the random forest achieved the highest decision coefficient, and the corresponding RMSE and MAE both corresponded to the lowest values, indicating its stronger modeling ability for high-dimensional features after embedding processing. So, random forest is selected as the main regression model to answer question two.

Result Analysis

From the perspective of feature contribution, it can be seen from Figure 7 that the feature importance of the random forest indicates that the importance score corresponding to the claim amount is much higher than other indicators, indicating that the prediction of the compensation amount should first consider the user's original demands; The 'delivery time' and 'insured amount' are located in the 2nd and 3rd positions respectively, indicating that the service response time and insured strategy also have a certain impact on the compensation amount; And the corresponding numbers 4, 5, and 6 are "destination order volume", "delivery timeout duration", and "compensation ratio at both ends of the network", indicating that the quality of operation at the corresponding network may have a certain regulatory effect on the compensation amount. Some data have a relatively small contribution and only serve as supplementary information, but can be used in practice to segment customer types.

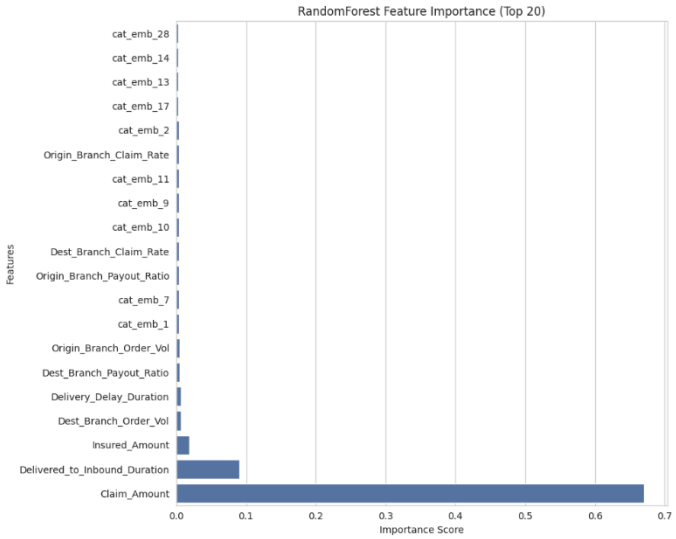

Figure 7. Random Forest Feature Importance (Top 20)

From the corresponding test set analysis, the R^2 =0.7706, RMSE=135.5171, and MSE=92.1084 of the deep forest model were calculated, as shown in Table 4. This level indicates that deep forests have a good expla-

nation of the model, with an average error of less than 100 yuan, which can basically meet the business needs of claims cost assessment, the results are shown in Table 4.

Table 4. Deep Forest Model Test Results Table

split	rmse	mae	R2
validation	135.230224	96.386484	0.768688
test	135.517070	92.108405	0.770615

INDIRECT CLASSIFICATION BASED ON REGRESSION AND NEAREST-CENTROID MAPPING

To map the one-dimensional prediction results of the regression model to the three risk labels, a "two-stage" indirect classification mechanism is adopted in this study. In the first stage, a regression model is used to predict the actual compensation amount for each waybill. In the second stage, the predicted compensation amount is combined with the claim amount to reconstruct two-dimensional input features consistent with the clustering annotation—specifically, the claim gap is defined as the predicted compensation amount minus the claim amount, and the claim amount ratio is the ratio of the predicted compensation to the claim amount. These features are then standardized using the mean and standard deviation saved from the clustering phase. Subsequently, the Euclidean distance from this standardized point to the three converged cluster centers is calculated, and the risk labels—"reasonable demands", "high demand", and "serious excess"—are assigned based on the nearest-centroid rule. This mechanism ensures that the indirect classification outputs remain consistent with the clustering annotation rules, and allows regression errors to be intuitively interpreted through distance variations.

THE RELATIONSHIP BETWEEN INPUT FEATURES AND TARGET VARIABLES

To explore the relationship between input features and target variables in this article and improve the prediction accuracy of the results, multiple model algorithms were introduced for comparison, and the optimal model was selected to solve the problem. Specifically, XGBoost, CatBoost, LightGBM, and Random Forest were chosen.

Note: The supervised labels used for "direct classification" in this section are the risk labels (silver labels) generated by the aforementioned capacitated clustering. The evaluation objective is to compare the reproduction capability and stability of different modeling routes towards this annotation rule.

Model training

After performing feature engineering on the four models, each model is trained to complete benchmark fitting. Then, combined with strategies such as class weights, SMOTE oversampling, and threshold adjustment, the problem of imbalanced class distribution is initially alleviated, allowing the model to balance accuracy and robustness in complex situations. After training, the model is retained along with its corresponding encoder, SVD transformer, and feature list, and can be directly called in the future.

Build model evaluation indicators

To measure the overall hit level and the ability to identify minority classes in extremely imbalanced data distributions, two evaluation metrics, confusion matrix and F1, are introduced:

The confusion matrix helps to break down the hit and miss structures of various categories, effectively focusing on whether severely overstocked samples have been missed. It specifically includes three indicators: accuracy, recall, and overall accuracy. The formula for the corresponding precision Pk and for any category k is as follows:

$$P_k = \frac{TP_k}{TP_k + FP_k} \tag{16}$$

The comparison chart of each model after different processing and adding indirect classification is shown in Figure 8.

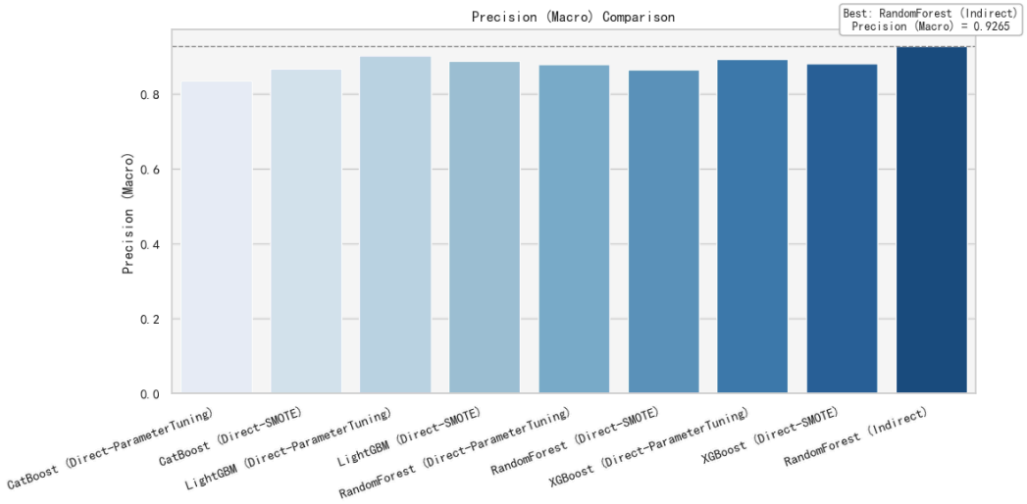

Figure 8. Precision (Macro) Comparison

The formula for recall rate R_k is as follows:

$$R_k = \frac{TP_k}{TP_k + FN_k} \tag{17}$$

The corresponding comparison chart is shown in Figure 9.

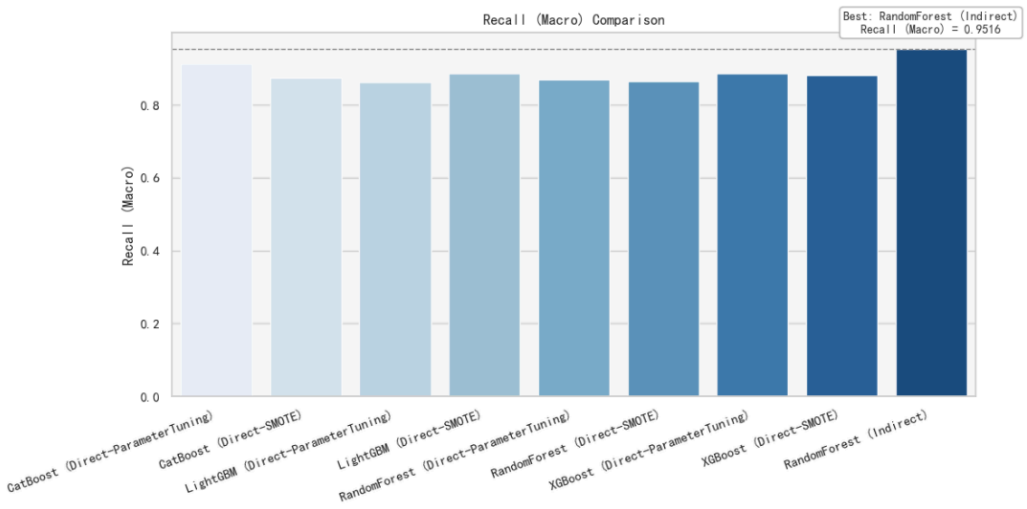

Figure 9. Recall (Macro) Comparison

The formula for overall accuracy ACC is as follows:

$$ACC = \frac{\sum_{k=1}^{K} TP_k}{N} \tag{18}$$

The corresponding comparison chart is shown in Figure 10.

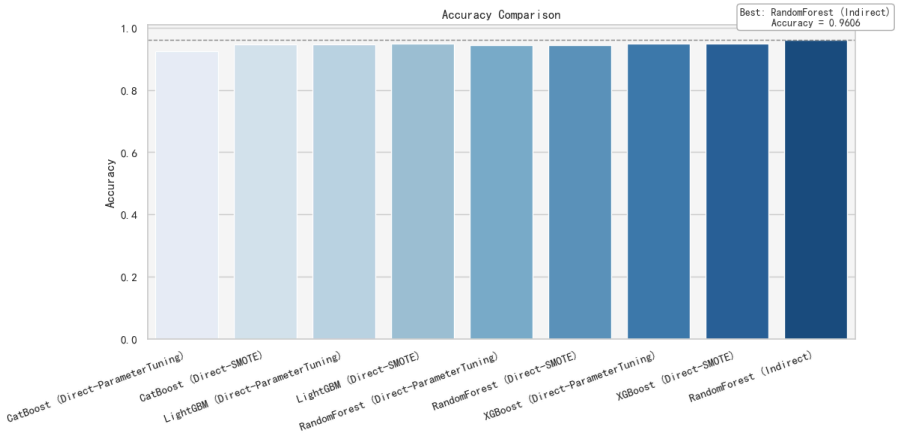

Figure 10. Accuracy Comparison

Among them, TP_k , FP_k , FN_k respectively represent the number of true positives, true negatives, and false negatives for category k, K is the total number of risk categories, and N is the total number of samples. F1 is used to balance the relationship between precision and recall, as follows:

$$F1_{macro} = \frac{1}{K}\sum_{k=1}^{K}\frac{2P_kR_k}{P_k + R_k} \tag{19}$$

The corresponding comparison chart is shown in Figure 11.

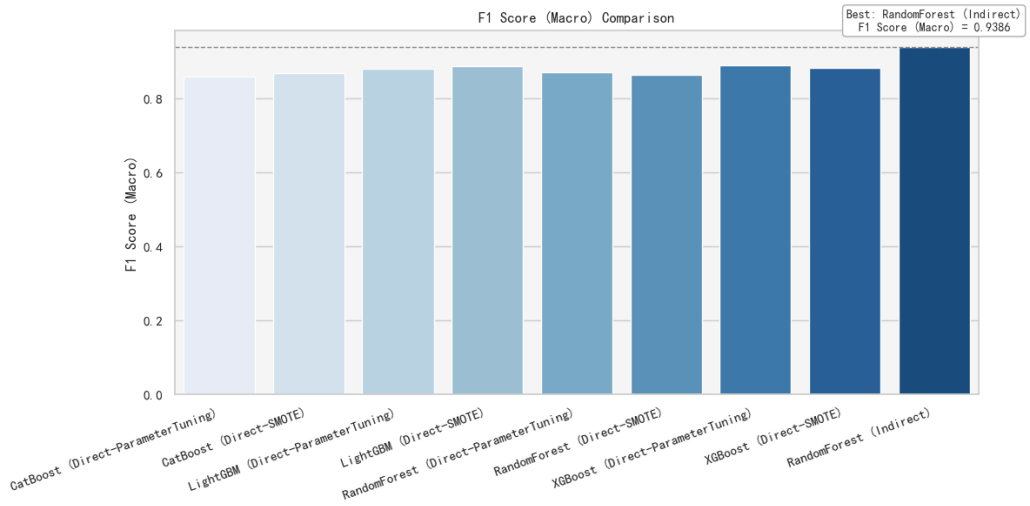

Figure 11. F1 Score (Macro) Comparison

Based on these statistical measures, select the model with the best performance through different processing methods as the direct classification method for solving the problem, and add indirect classification for comparison. From the results in the picture, it can be seen that the XGBoost model with parameter tuning performs the best in direct classification, but the metric data in indirect classification is better than that in direct classification. To better solve this problem, the indirect classification method will be used as the core method for the subsequent solving process.

Result Analysis

From the evaluation results, the evaluation data corresponding to indirect classification are shown in Table 5. The overall accuracy of the validation set is 0.9606, the macro average recall rate is 0.9516, the macro average accuracy rate is 0.9265, and F1 is 0.9386. Can accurately and directly output three types of labels: “reasonable demands”, “high demands”, and “serious exceedance”.

Table 5. Classification Model Performance Evaluation Results (Macro-Averaged)

accuracy	0.9606
Precision-macro	0.9265
recall-macro	0.9516
f1-macro	0.9386
Name:score,dtype	Float64

The final classification result is shown in Figure 12, which indicates that the prediction is basically correct, indicating a high feasibility of the proposed solution.

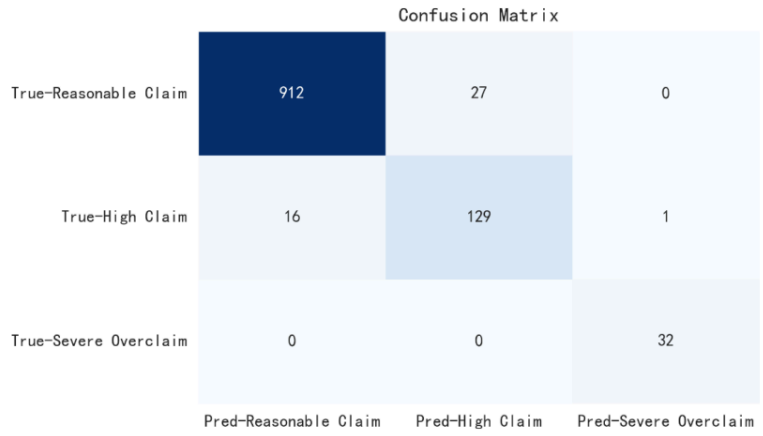

Figure 12. Confusion Matrix

To better understand the two comparison methods, compare the advantages and disadvantages of direct classification and indirect classification:

The advantage of direct classification method is that it is a single model, easy to deploy and coordinate; You can directly participate in parameter tuning and threshold optimization by using evaluation indicators as optimization objectives; Its probability output is beneficial for threshold, cost sensitivity, and business rule fusion. The disadvantage is that it requires learning artificial boundaries generated from data, which are greatly affected by label noise; Moreover, under imbalanced categories, it is easy to lean towards the majority class. Without optimization, the recall of minority classes is easily limited.

The advantage of indirect classification is that it maintains strong consistency with label definitions; And its decision coefficient is relatively high. When projected onto two dimensions, the “nearest center” is used for decision-making, which is relatively stable, interpretable, and has clear boundaries; For minority optimization, the recall rate is usually higher. The disadvantage is that it cannot achieve end-to-end optimization directly

based on classification indicators; And it may be due to errors in the two stages that errors accumulate, resulting in misclassification.

CONCLUSIONS

This article transforms complex practical problems of claim governance into simple mathematical models. By using evaluation indicators such as silhouette coefficient, determination coefficient, and F1 score, the established model yields convincing results in both statistical performance and business applications. The selected random forest regression model has an excellent goodness of fit. Even under high randomness and complex related factors, its R^2 reaches 0.77, while the RMSE and MAE remain at a low level, indicating that the model has high accuracy and robustness in predicting the actual compensation amount. Furthermore, by establishing an indirect classification mechanism, it provides a more business-interpretable output compared to direct classification models. However, this study still has certain limitations: the model relies mainly on historical data, and the threshold is set to a fixed value. Future research could consider incorporating unstructured textual data, such as customer communication records, and designing dynamic proportion constraints that can adaptively adjust according to real-time logistics business volumes, thereby enhancing the practical application value of the model.

Author Contributions

Conceptualization – Yu Chen; methodology – Yu Chen; formal analysis – Cheng Liang; investigation – Yu Chen; resources – Cheng Liang; writing-original draft preparation – Yu Chen; writing-review and editing – Cheng Liang; visualization – Biaoyong Liang; supervision – Biaoyong Liang. All authors have read and agreed to the published version of the manuscript.

Conflicts of Interest

The authors declare no conflict of interest.

Funding

This research received no external funding.

Acknowledgements

Not applicable.

REFERENCES

[1] Mand M, Singh B, et al. A novel data normalization technique based on a piecewise continuous, symmetric function with tanh-transformation. Soft Computing. 2026:1-43. doi: 10.1007/s00500-026-11232-Y

[2] Wang X, Shi D, Xue F, et al. Boosting the detection of enhancer-promoter loops via normalization methods for chromatin interaction data. Nature Communications. 2026; 17(1):2299. doi: 10.1038/s41467-026-69082-Z

[3] Koca G, Egilmez O, Turan B. Intergenerational attitudes towards migration: K-means clustering analysis of European countries. Quality & Quantity. 2026:1-25. doi: 10.1007/s11135-026-02767-1

[4] Wang Z, Mi J, Li Z. Multi-granularity rough k-means clustering in the optimization of collaborative filtering algorithms. Applied Intelligence. 2026; 56(6):201. doi: 10.1007/s10489-026-07224-Y

[5] Min X, Wang C. Optimization of multivariate linear regression model for ionospheric disturbed index. Advances in Space Research. 2026; 77(9):9675-9689. doi: 10.1016/j.asr.2026.03.041

[6] LeSueur A, Bianchi P, Gallarati S, et al. Best Practices and Considerations for Applying Multiple Linear Regression in Organic Chemistry Research. The Journal of Organic Chemistry. 2026. doi: 10.1021/acs.joc.5c03206

[7] Sowmya C S, Prathap J A. A new fusion modified decision tree algorithm and local binary histogram pattern-based improved KNN algorithm for fault investigation in power inverter. Computers and Electrical Engineering. 2026; 135:111171. doi: 10.1016/j.compeleceng.2026.111171

[8] Peng Z, Wang Y, et al. Joint domain knowledge graph and attention-based decision trees for coke prediction in ethylene cracking furnace tube. Measurement. 2026; 276:121467. doi: 10.1016/j.measurement.2026.121467

[9] Peng Y, Zhang S, et al. Risk Prediction of Water Inrush in Diversion Tunnel Crossing Water-Rich Fault Based on NRBO-XGBoost Algorithm. Applied Sciences. 2026; 16(8):3831. doi: 10.3390/app16083831

[10] Liu X, Wang X, et al. Explainable machine learning for predicting mechanical ventilation in Alzheimer's patients with pneumonia: a SHAP-guided XGBoost nomogram based on MIMIC-IV. Scientific Reports. 2026. doi: 10.1038/s41598-026-48214-x

[11] Hou C, Liu W, et al. A Quality-Control Fusion Algorithm for Cloud-Radar Data in Complex Weather Scenarios Integrating LightGBM and Neighborhood Filtering. Remote Sensing. 2026; 18(5):691. doi: 10.3390/rs18050691

[12] Dong T, Huang J, Wang C. Multi-objective optimization of high-speed permanent magnet machine based on fast hierarchical surrogate model integrating PDP-LightGBM. COMPEL: The International Journal for Computation and Mathematics in Electrical and Electronic Engineering. 2026; 45(1):18-35. doi: 10.1108/COMPEL-05-2025-0233

[13] Li Z, Ji Y, et al. Nondestructive Detection of Eggshell Thickness Using Near-Infrared Spectroscopy Based on GBDT Feature Selection and an Improved CatBoost Algorithm. Foods. 2026; 15(8):1286. doi: 10.3390/foods15081286

[14] Zhao Q, He Y, et al. Predicting perovskite solar cell stability using CatBoost with Bayesian optimization and SHAP interpretability analysis. Journal of Alloys and Compounds. 2026; 1057:186879. doi: 10.1016/j.jallcom.2026.186879

[15] Shangguan Y, Gao C, et al. Research on Control Factors and Parameter Optimization of Surfactant Flooding in Low-Permeability Reservoirs Using Random Forest Algorithm. Processes. 2026; 14(7):1108. doi: 10.3390/pr14071108

[16] Ding S, Gong Y, et al. Factors associated with the prognosis of diabetic foot ulcers treated with silver ion hydrogel dressings combined with negative pressure wound therapy and the construction of a risk prediction model: a study of two machine learning methods, LASSO and random forest algorithms. International Journal of Diabetes in Developing Countries. 2026:1-15. doi: 10.1007/s13410-026-01636-9

[17] Asadi V, Sheikhi N. COSMOS2025: Machine Learning Classification of Early- and Late-type Galaxies at $0 < z < 3$. The Astrophysical Journal. 2026; 1002(1):26. doi: 10.3847/1538-4357/ae4eca

[18] Vilimkova Kahankova R, Barnova K, et al. Machine learning based classification in obstetrics: evaluating models, partitioning strategies, and key predictors in cardiotocography. doi: 10.1186/s12884-026-09130-0

[19] Li J, Hua K, et al. Development and Deployment of an Explainable Machine Learning Model for Preoperative Prediction of Thigh Liposuction Volume in Female Patients. Annals of Plastic Surgery. 2026. doi: 10.1097/sap.0000000000004758

[20] Tribhuvan M, Phatke S N. Comment on “Exploration of a Multimodal Machine Learning Model Integrating Ultrasound and Clinical Indicators for the Diagnosis of Diabetic Peripheral Neuropathy”. Journal of Ultrasound in Medicine. 2026. doi: 10.1002/jum.70275