DOI: 10.23977/ferm.2024.070613
ISSN 2523-2576 Vol. 7 Num. 6

# *Performance Evaluation and Interpretation of Non-life Insurance Company Bankruptcy Prediction Model Using LightGBM Algorithm and SHAP Method*

**Peng Dong**

*School of Business, Stevens Institute of Technology, Hoboken, New Jersey, 7030, United States*
*pengdong0128@GMAIL.COM*

***Keywords:*** Non Life Insurance Companies; Feature Engineering; Bankruptcy Prediction; LightGBM; SHAP Explanation

***Abstract:*** This article evaluates and explains the performance of a bankruptcy prediction model for non life insurance companies using the LightGBM algorithm combined with the SHAP method. With the impact of unexpected events such as the epidemic, the market share of the non life insurance industry continues to decline, and companies are facing enormous transformation pressure and even bankruptcy risks. Non life insurance companies not only ensure the property safety of policyholders, but are also closely related to the stability of the economy. Due to the limitations of traditional bankruptcy prediction methods with many assumptions, this paper uses feature engineering to clean, construct, screen, and extract high-dimensional non life insurance enterprise data, and constructs a global bankruptcy prediction model. Research has shown that models processed through feature engineering have better prediction accuracy than untreated models, especially in the F1 value, AUC value, recall rate, and other indicators of the feature construction group and feature selection group, which have significantly improved. The feature extraction group has the largest improvement in accuracy and recall rate. In addition, the interpretation results using the SHAP method show that the feature of "total liabilities and earnings" contributes the most to the prediction of the feature selection model, further verifying the interpretability and accuracy of the model.

## 1. Introduction

With the continuous integration of the global economy and the increasing openness of the insurance industry, people's demand for wealth management continues to grow, and the insurance industry has gradually become an important component of the global economy. In recent years, the sudden COVID-19 epidemic has intensified the challenge of the non life insurance industry, and the global non life insurance market has suffered a heavy blow. Under the impact of the epidemic, the life insurance market has received more attention due to its long-term protection advantages, while the non life insurance market has gradually reduced its market share due to its short protection period and no grace period, and some companies have even fallen into bankruptcy crisis. Traditional bankruptcy prediction methods have many assumptions and significant discrepancies with actual data, making it difficult to provide accurate decision support. Therefore, based on this background, this article applies the LightGBM algorithm to construct a bankruptcy prediction model for non life insurance enterprises, and uses the SHAP method for explanatory analysis, aiming to improve the predictive performance and explanatory ability of the model, and provide effective reference for risk warning and management of non life insurance enterprises. It is a statistical learning method used to handle classification problems, particularly suitable for binary classification problems, where the target variable has only two values. Its basic idea is to convert the output of linear regression into an estimate of class probability through a linear regression model and a transformation called a logistic function, in order to make classification predictions. R Qin designed support vector machines and logistic regression models in his article[1]for detecting accounting fraud. Enhance the learning and generalization ability of unknown situations using support vector machines. The experimental results of RM Kumar and his team members in the article indicate that the multi-layer perceptron neural network (MLPNN), which has been standardized and rescaled by covariates, performs slightly better in prediction than the logistic regression (LR) model. In the testing and validation stage of sample cases, its correct classification rate reached 82.8% [2]. In the article[3], O Takawira et al. focused on using logistic regression (LR) to model sovereign credit rating (SCR), aiming to identify its determining factors and predict future rating changes. L Xiaojie et al [4] used three data mining algorithms, including support vector machine, decision tree, and logistic regression, to build a stock classification prediction model in the article. It is a machine learning algorithm based on gradient boosting framework, developed by Microsoft, aimed at solving machine learning and data mining problems on large-scale datasets. Y Meng et al. proposed an arc fault detection algorithm based on a fully trained LightGBM model in the article [5], which can efficiently and accurately detect multiple arc faults. In their article [6], Lin et al. optimized the parameters of the LightGBM algorithm using genetic algorithm, and LightGBM based on GA optimization achieved significant performance improvement. The optimized model has improved average accuracy and average running time by 0.5% and 72.12%, respectively.AH Carlson The regression model was [7] using the two-step consistency estimator of Heckman. It is similar to a two-step consistent Heckman estimator but allows heteroscedasticity in the first step and a more general control function specification. AC Jaures In their study [8], used the travel cost method, descriptive statistics, and two-step Hekenman method to analyze the use and economic value of indigenous seasonal climate forecast (ISCF) in Benin. A Krishnan et al [9] used the Heckman two-step program (1979), and the results show that the ECB has a positive impact on the OFDI of enterprises. The results show that companies using more leverage and ECB are those with higher OFDI intensity. MAJi-Liang et al [10] estimated the determinants of family behavior in commercial legume planting through the Heckman two-step model, and used the endogenous treatment regression (ETR) method to examine the impact of commercial legume planting on family economic welfare.

## 2. Predictive Model Construction

Gradient boosting is a boosting method that uses serial integration to increase the weight of misclassified samples while reducing the weight of correctly classified samples, in order to improve the classifier's prediction accuracy for erroneous samples. Its core idea originates from gradient descent, with the goal of forming a powerful learning model through the combination of multiple weak learners. This method optimizes the loss function through greedy algorithm in each iteration, continuously approaching the optimal solution. Gradient Boosting Decision Tree (GBDT) is a typical implementation method that can effectively improve prediction performance by integrating multiple shallow tree models. In classification tasks, GBDT typically uses CART trees, which select the best splitting point by calculating the Gini coefficient of each feature. Through the forward distribution algorithm, GBDT continuously improves the performance of each tree and ultimately integrates the prediction results of all trees to achieve more accurate predictions. This model can adapt to different types of data (such as discrete and continuous) and has the ability to prevent overfitting, making it highly practical and accurate in the field of machine learning.

LightGBM and XGBoost are two famous GBDT implementations, although they are similar in basic principles and both fit the negative gradient of the loss function as the residual of the decision tree, LightGBM performs better in multiple aspects. This includes faster training speed, lower memory usage, higher accuracy, and better ability to handle large-scale data. In addition, LightGBM supports parallel learning and can directly process category features, thereby improving model performance. Some researchers have established an equation relationship between LightGBM and XGBoost, covering histogram based decision tree algorithm (Histogram), one-sided gradient sampling (GOSS), and mutually exclusive feature bundling (EFB). The developers of LightGBM focus on addressing some of the shortcomings of XGBoost, optimizing computational complexity, memory consumption, and efficiency, while reducing the risk of overfitting. Compared with XGBoost, LightGBM performs well in processing features and samples, and its memory consumption is particularly prominent when dealing with massive amounts of data. This model adopts a depth constrained leaf first strategy, effectively reducing computational and storage costs while avoiding the possibility of overfitting. LightGBM also implements a fast, distributed, and efficient gradient boosting decision tree framework, suitable for various machine learning tasks such as classification, regression, and sorting. Among them, the histogram based decision tree algorithm reduces the number of candidate splitting points by discretizing continuous features and generates histograms to help determine the optimal splitting point.

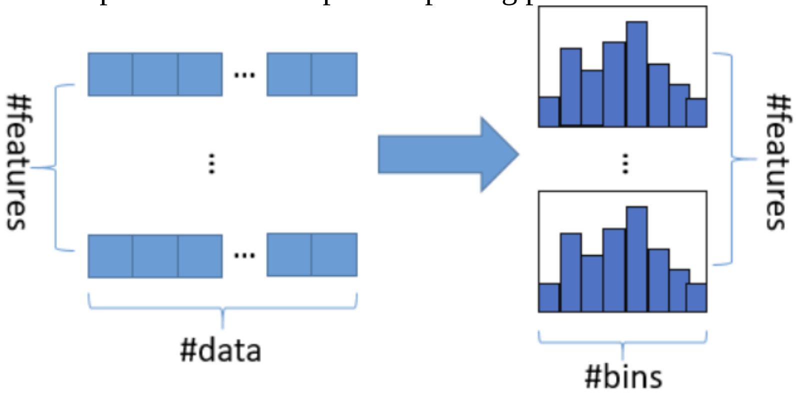

Figure 1: Schematic diagram of histogram algorithm principle

The principle of histogram algorithm is shown in Figure 1. Simply put, the number of boxes (# bins) required for each feature is determined based on the feature values, and a unique integer encoding is assigned to each box. Next, divide all feature values into several intervals, with the number of intervals equal to the number of boxes. Then, the floating-point eigenvalues in the sample will be updated to the corresponding box encoding, and finally expressed through histograms (# bins). The essence of this algorithm lies in aggregating massive data through histogram statistical methods. This processing method has multiple advantages: on the one hand, the data storage of discretized features is simpler, the calculation speed is faster, and it has strong robustness, thereby improving the stability of the model. On the other hand, the histogram method does not require additional storage of pre sorted results, only retaining the discretized feature data, which significantly reduces memory consumption. In addition, XGBoost repeats the gain calculation when traversing each feature value, while LightGBM's histogram algorithm only needs to calculate k times (k is considered a constant), reducing the time complexity from O (data x feature) to O (k x feature). Obviously, when the sample size (# data) is much larger than k, the computational cost will be significantly reduced.

Traditional statistical models such as linear models have good explanatory power and can evaluate feature importance through sklearn's importance interface. However, for data scientists who have a deep understanding of the internal mechanisms of the model, simply focusing on accuracy is not enough. In recent years, machine learning models have continuously evolved, and researchers have begun to explore the reasons behind the effectiveness of models, giving rise to interpretable methods. These methods are becoming increasingly important and can help decision-makers accurately understand the application of models and improve service quality. Interpretability refers to the reader's understanding of the algorithm's operation and prediction results, especially in complex machine learning models. High interpretability helps to reveal the relationship between prediction results and features, and evaluate the credibility of the model. The Shapley value was proposed by Professor Lloyd Shapley of the University of California, Los Angeles. The Shapley Additive Explanations (SHAP) method is used to solve the problem of profit distribution in cooperative games by ensuring that participants' profits match their contributions. On this basis, SHAP values provide an independent method for model interpretation, ensuring consistency in feature effects and consistency between local samples.

When calculating the SHAP value of the LightGBM model in this article, the structure of the tree model needs to be utilized. Assuming M represents the set of all features and S is a permutation subset of M. The SHAP value of the jth feature can be considered as its contribution. For sample i, its jth eigenvalue is $x_{ij}$, and $f_{x_i}(S)$ represents the average predicted value of the model for sample $x_i$ among all samples when only using the feature set S. When S is empty, the initial value of the model is $\phi_{ij}$. After adding feature values to the feature set S, the marginal contribution value of the model can be represented as $f_{x_i}(S)$. Therefore, the SHAP value calculation formula for the jth feature of the i-th sample is:

$$\phi_{ij} = \sum_{S \subseteq M \setminus \{x_{ij}\}} \frac{|S|!(m-|S|-1)!}{m!} \left[ f_{x_i}\left(S \cup \{x_{ij}\}\right) - f_{x_i}(S) \right], j \geq 1 \tag{1}$$

## 3. Experimental Results and Analysis

This article studies the binary classification problem of treating a company's operating status as bankruptcy and normal operation. Based on the ISIS database provided by Bureau van Dijk (BvD), the bankruptcy of non life insurance companies worldwide is predicted. The ISIS database contains detailed information on 16404 insurance companies, and the company status is divided into three categories: "active", "inactive", and "unknown". "Active" refers to normal operation, while "inactive" refers to bankrupt companies. After data processing, 6619 well managed enterprises and 617 bankrupt enterprises were finally selected. After imbalanced data processing, a total of 12230 non life insurance enterprises were included for model analysis.

In the lower right corner of Figure 2, 0 represents the control group, 1 represents the feature construction group, 2 represents the feature selection group, and 3 represents the feature extraction group in the legend. By comparing the ROC curves, it can be seen that the AUC value of the control group is 0.70, indicating relatively low classification ability. After feature construction, the AUC value of the sample group significantly increased, reaching 0.97. However, after feature selection, the AUC value decreased to 0.88, and then in the feature extraction stage, due to slight information loss, the AUC value further decreased to 0.85. These values are still higher than the initial level of the control group, indicating that all aspects of feature engineering have effectively enhanced the classification prediction ability of the LightGBM model and improved overall performance.

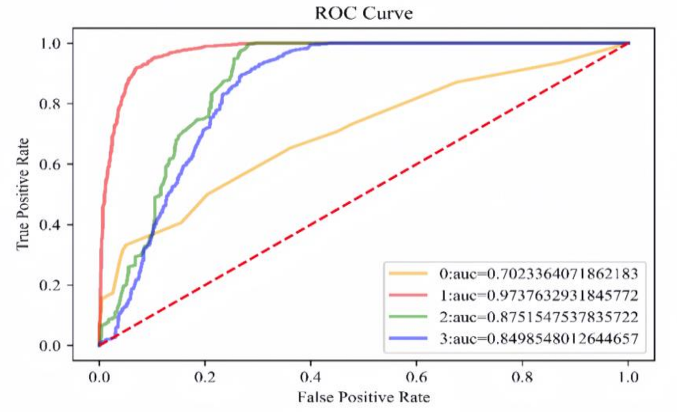

Figure 2: Comparative analysis of ROC curves

As shown in Figure 3, each row in the chart represents a characteristic variable, and the horizontal axis represents the SHAP value. Each point corresponds to a sample, and the color depth of the point reflects the value of the feature variable: the redder the color, the greater the value of the feature; The bluer the color, the smaller the feature value. In order to make the visualization of SHAP values more differentiated, the input samples were binned. From the previous analysis, it can be seen that the characteristic variables "total liabilities and earnings" play a key role in prediction. Taking this feature as an example, further analyze its performance in the entire sample. Observing the variable x40, samples with smaller box numbers have positive SHAP values and larger absolute values, mainly distributed between 1.5 and 2.5, corresponding to the blue area on the right side of the first row in the figure; Samples with larger numbers have negative SHAP values, mainly concentrated between -4 and -0.5, corresponding to the red area on the left side of the graph; The SHAP values of the centrally numbered samples are mostly between 0 and 1.5, as shown in the purple area in the figure.

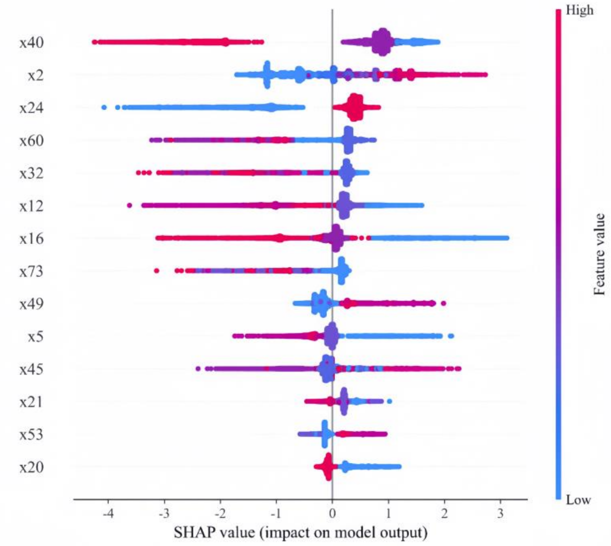

Figure 3: Visualization of SHAP values for the overall sample

## 4. Conclusion

This study utilizes the ISIS database on the Bureau van Dijk platform to classify and predict the activity and bankruptcy status of non life insurance companies worldwide. By constructing the control group, feature construction group, feature selection group, and feature extraction group of the LightGBM model, the impact of each step of feature engineering on the model's prediction performance was analyzed. LightGBM, as an efficient implementation of GBDT, adopts a leaf growth strategy that can effectively process large-scale data and reduce the risk of overfitting. Although the model has been widely recognized in theory, its interpretability still needs further verification. Therefore, this article conducted an in-depth analysis of the prediction results using the SHAP method. The main conclusion is that after feature engineering processing, the accuracy of each group of samples exceeded 76.45% of the control group, indicating that feature engineering significantly improved the accuracy of the model. The performance of feature construction and feature selection groups was particularly outstanding. In the comparison of evaluation indicators for different feature engineering methods, the feature construction group performed the best in F1 and AUC values. The AUC, recall, and accuracy of the feature selection group were all improved by more than 10%, while the accuracy and recall of the feature extraction group were improved by about 15%. The SHAP interpretation method comprehensively demonstrates the impact of feature variables on prediction results. Research has found that the feature of "total liabilities and earnings" contributes the most to classification results, with higher values indicating smaller SHAP values and a greater tendency to classify samples as active.

The innovation of this study is mainly reflected in two aspects: firstly, by gradually improving sample processing through feature engineering, the problem of data imbalance has been effectively solved, significantly improving the predictive performance of the model; Secondly, the SHAP interpretability method was used to rationalize the prediction results of the machine learning model, enhancing the understanding of the model's operation process. Looking ahead to future research, we can explore in depth the expansion of data sources to include more non life insurance companies that have not yet been uniformly counted, thereby enhancing the applicability of the model; Meanwhile, attention should also be paid to the pattern of missing values, treating them as feature inputs to avoid model underfitting caused by data loss.

## References

*[1] Qin L, Wang X, Yin L, et al. A distributed evolutionary based instance selection algorithm for big data using Apache Spark[J]. Applied Soft Computing, 2024, 159. DOI:10.1016/j.asoc.2024.111638.*

*[2] Singh T, Khanna R, Satakshi, et al. Improved multi-class classification approach for imbalanced big data on spark [J]. Journal of supercomputing, 2023.*

*[3] Nithya T M, Umanesan R, Kalavathidevi T, et al. Deep LearningModel for Big Data Classification in Apache Spark Environment[J]. Intelligent Automation and Soft Computing, 2023, 37(9):2537-2547. DOI:10.32604/iasc.2022.028804.*

*[4] Yadav M L, Hawamdeh S. Query Execution Time Analysis Using Apache Spark Framework for Big Data: A CRM Approach[J]. Journal of information & knowledge management, 2022. DOI:10.1142/S0219649222500502.*

*[5] Chen R, Yang B. Construction of an Intelligent Analysis Model for Website Information Based on Big Data and Cloud Computing Technology[J]. Discrete Dynamics in Nature and Society, 2022, 2. DOI:10.1155/2022/7876119.*

*[6] Rehman A, Saba T, Haseeb K, et al. IoT-Edge technology based cloud optimization using artificial neural networks [J]. Microprocessors and Microsystems, 2024, 106. DOI:10.1016/j.micpro.2024.105049.*

*[7] Carlson A H. Gtsheckman: Stata module to compute a generalized two-step Heckman selection model[J]. Statistical Software Components, 2022.*

*[8] Amegnaglo Cocou Jaurès, Akwasi M B, Asomanin A K. Use and economic benefits of indigenous seasonal climate forecasts: evidence from Benin, West Africa[J]. Climate and Development, 2022.*

*[9] Krishnan A, Padmaja M. External Commercial Borrowings and Outward Foreign Direct Investment: Evidence From Indian Manufacturing Firms[J]. Asian Economics Letters, 2023. DOI:10.46557/001c.74858.*

*[10] Jiliang Ma, Fan Li, Huijie Z, et al. Commercial cash crop production and households' economic welfare: Evidence from the pulse farmers in rural China[J]. Journal of Agricultural Sciences: English edition, 2022, 21(11):3395-3407.*