Article

Not peer-reviewed version

A Hybrid Ensemble Method with Focal Loss for Improved Forecasting Accuracy on Imbalanced Datasets

Xiaojun Guo * , Wenxiu Cai , Yu Cheng , Jiaqi Chen , Liyang Wang

Posted Date: 10 April 2025

doi: 10.20944/preprints202504.0831.v1

Keywords: financial prediction; ensemble model; LightGBM; XGBoost; focal loss; imbalanced data

Article

A Hybrid Ensemble Method with Focal Loss for Improved Forecasting Accuracy on Imbalanced Datasets

Xiaojun Guo [1,*], Wenxiu Cai [2], Yu Cheng [3], Jiaqi Chen [4] and Liyang Wang [3]

[1] Independent Researcher, New Jersey, USA
[2] University of San Francisco, Harrison, USA; wcai14@usfca.edu
[3] Independent Researcher, Mclean, USA; yucheng576@gmail.com (Y.C.); liyang.illinois@gmail.com (L.W.)
[4] Independent Researcher, Chicago, USA; ronanchen0901@gmail.com
* Correspondence: xiaojunguo2018@gmail.com

Abstract: The inherent complexity and dynamic characteristics of diverse datasets present significant challenges for achieving high predictive accuracy in forecasting tasks. This study tackles these challenges by implementing a hybrid ensemble model aimed at enhancing predictive performance across imbalanced datasets. Using data from a competitive data source, the approach integrates LightGBM, XGBoost, and Logistic Regression models within a weighted ensemble framework to improve overall prediction accuracy. Data preprocessing techniques, including KNN imputation, Z-score normalization, and SMOTE, are employed to handle missing values, outliers, and class imbalances, ensuring a robust input for model training. The ensemble framework incorporates a Focal Loss function to specifically address class imbalances and refine prediction precision. Comparative analyses reveal that the proposed ensemble model consistently outperforms individual models in terms of accuracy, precision, recall, and AUC. This study offers a versatile and reliable solution for forecasting challenges, demonstrating enhanced robustness and broad applicability across domains.

Keywords: financial prediction; ensemble model; LightGBM; XGBoost; focal loss; imbalanced data

---

1. Introduction

Financial forecasting faces challenges due to the volume and volatility of data, where traditional models often fall short. This study proposes a multi-model ensemble framework combining LightGBM, XGBoost, and Logistic Regression. LightGBM's efficiency in large datasets, XGBoost's regularization for overfitting control, and Logistic Regression's interpretability create a balanced ensemble leveraging their strengths.

To address imbalanced data, a custom Focal Loss function shifts the focus to underrepresented classes, improving predictions on minority events. Preprocessing includes KNN imputation for missing values, Z-score normalization for outliers, and SMOTE for balancing the dataset. Feature engineering, such as feature crossing and correlation-based selection, captures hidden variable interactions.

The ensemble employs weighted integration, optimizing model weights for overall accuracy and adaptability. Experimental results show superior performance across accuracy, precision, and recall metrics. The framework adapts to market changes, enhancing predictive consistency and sensitivity to critical yet infrequent events like downturns or price shifts.

This research provides a scalable and robust solution for financial forecasting, emphasizing model integration, imbalanced data handling, and adaptability to diverse financial scenarios. It lays a foundation for future explorations into hybrid models and advanced loss functions in financial analytics.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 April 2025 doi:10.20944/preprints202504.0831.v1

2. Related Work

Recent advances in financial forecasting models have focused on enhancing prediction accuracy through hybrid and ensemble machine learning approaches. Vuong et al. [1] demonstrated the effectiveness of a combined XGBoost and LSTM model for stock price forecasting, capturing both short-term and long-term trends in financial data. Lu et al. [2] leverage LightGBM's leaf-wise growth and custom loss functions for imbalanced data, integrating LightGBM, DeepFM, and DIN for purchase prediction. Their work underscores the importance of robust preprocessing and ensemble learning in improving predictive accuracy on complex datasets.Wang et al. [3] found that plot sentiment most influences a movie's success.

Li et al. [4] propose a dual-agent approach for strategic deductive reasoning in large language models, reflecting the influence of ensemble learning strategies. Their framework aligns with our study's emphasis on combining complementary models, such as LightGBM and XGBoost, to improve predictive accuracy and adaptability. Jin's [5]pioneering research presents a highly efficient and innovative approach to predicting stock market indices, showcasing the power of graph-based deep learning systems in capturing complex dependencies within financial data. This work highlights the significant potential of Graph Neural Networks (GNNs) in addressing the challenges of intricate financial data processing.

Lu [6] utilizes decision trees and text analytics (e.g., TF-IDF, BERTopic) to enhance chatbot satisfaction. This approach parallels our ensemble strategy, particularly LightGBM's handling of high-dimensional data, and highlights the importance of tailored preprocessing and model optimization for improved predictive performance and adaptability.Wang et al. [7] proposed an attention-based LSTM network for adaptive sensor selection, improving failure mode recognition and remaining useful lifetime prediction under time-varying operating conditions. Xu and Wang [8] proposed a multimodal LLMs-based MOE architecture for healthcare recommendations, demonstrating superior accuracy and personalization while highlighting the limited impact of image data on performance.

Sun et al. [9] proposed a multi-objective recommender system designed to enhance consumer behavior prediction in e-commerce. Their study emphasizes the integration of ensemble learning methods to balance competing objectives effectively. Li [10] integrates multimodal data and multi-recall strategies for e-commerce recommendations, leveraging ensemble learning principles. Their use of advanced techniques parallels our methodologies, including LightGBM for high-dimensional processing and custom loss functions for class balancing. Lu [11] employs ensemble learning for multi-objective e-commerce recommendations, emphasizing decision tree-based models like LightGBM. This aligns with our study, showcasing the robustness and adaptability of ensemble strategies for high-dimensional data and complex objectives. Yu et al. [12] found that the Sentence-t5 + Mistral 7B model excels in medical QA (precision 0.762), enhancing healthcare knowledge retrieval.

These studies underline advancements in financial forecasting, emphasizing ensemble models to address data imbalance, enhance interpretability, and improve accuracy. Future progress may leverage deep learning and refined techniques for missing or imbalanced data.

3. Methodology

This section presents a novel ensemble learning approach for financial forecasting, utilizing LightGBM and XGBoost with advanced preprocessing and a custom focal loss function. Using financial datasets from Two Sigma, our methodology demonstrates superior predictive performance compared to standard models. Details of the models, including structures, loss functions, preprocessing, evaluation metrics, and results, are provided. The model pipeline is illustrated in Figure 1.

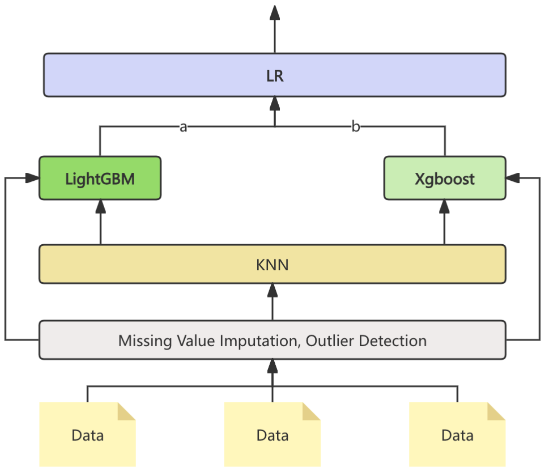

Figure 1. Pipeline of the ensemble model.

3.1. Model Architecture

Our financial forecasting approach employs an ensemble of LightGBM, XGBoost, and Logistic Regression (LR), leveraging their strengths in computational efficiency, interpretability, and predictive accuracy. This subsection details each model's architecture and their integration within the ensemble framework.

3.2. LightGBM

LightGBM (Light Gradient Boosting Machine) is a gradient boosting framework optimized for efficiency and scalability, making it ideal for large-scale datasets. Unlike traditional methods like XGBoost, LightGBM employs a leaf-wise growth strategy to build decision trees, reducing training time and improving accuracy.

The model minimizes a custom loss function by iteratively constructing decision trees to reduce residual errors:

$$y_i = \sum_{k=1}^{K} f_k(x_i) + \epsilon_i \tag{1}$$

where y_i is the prediction, f_k is the k-th tree, x_i is the feature vector, and ϵ_i is the residual error. The objective function includes a regularization term to prevent overfitting:

$$L(f_k) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \lambda \cdot \|f_k\|^2 \tag{2}$$

where $l(y_i, \hat{y}_i)$ is the loss function, and λ controls tree complexity.

LightGBM's ability to handle high-dimensional and sparse data makes it highly effective for financial forecasting tasks.

3.3. XGBoost

XGBoost (Extreme Gradient Boosting) is a tree-based ensemble model with optimizations for sparse data, parallel computation, and overfitting reduction through regularization. Predictions are made by summing contributions from decision trees:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i) \tag{3}$$

 doi:10.20944/preprints202504.0831.v1

XGBoost minimizes a regularized objective function:

$$L(\theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{4}$$

where the regularization term $\Omega(f_k)$ is:

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \|w\|^2 \tag{5}$$

Here, T is the number of leaves, λ is the L2 regularization term, and w represents leaf weights.

XGBoost's ability to handle missing data and sparse features, coupled with parallelization, makes it highly efficient for large-scale financial forecasting tasks.

3.4. Logistic Regression (LR)

Logistic Regression (LR) is a statistical model for binary classification, valued for its simplicity and interpretability. In our financial prediction task, LR provides stable and interpretable predictions, complementing tree-based models in the ensemble.

The model predicts the probability of the positive class as:

$$P(y = 1|x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)}} \tag{6}$$

where $x_1, \ldots, x_p$ are feature values, and $\beta_0, \ldots, \beta_p$ are learned coefficients. LR is particularly effective when feature-target relationships are approximately linear.

In the ensemble, LR acts as a stabilizer, reducing the risk of overfitting associated with tree-based models like LightGBM and XGBoost.

3.5. Ensemble Strategy

We employ a weighted ensemble of LightGBM, XGBoost, and LR to enhance predictive performance by leveraging their complementary strengths: LightGBM's efficiency, XGBoost's regularization, and LR's interpretability.

The ensemble's final prediction $\hat{y}$ is computed as:

$$\hat{y} = w_3 * (w_1 \hat{y}_{\text{LightGBM}} + w_2 \hat{y}_{\text{XGBoost}}) \tag{7}$$

where w_1, w_2, w_3 are optimized weights for each model's predictions.

This approach improves generalization by reducing variance and bias, making it well-suited for the noisy and volatile nature of financial forecasting tasks.

3.6. Custom Focal Loss Function

The focal loss function is applied to mitigate class imbalance by focusing on harder examples. The focal loss function is defined as:

$$FL(p_t) = -\alpha_t (1 - p_t)^{\gamma} \log(p_t) \tag{8}$$

where p_t is the predicted probability for the correct class, α_t is the balancing factor between classes, and γ is the focusing parameter that adjusts the rate at which easy examples are down-weighted. This loss function helps in cases where the financial dataset is imbalanced with respect to the target labels.

3.7. Data Preprocessing

Data preprocessing is vital in machine learning pipelines, especially in financial forecasting where datasets are large, noisy, and may contain missing or imbalanced values. This section outlines the

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 April 2025 doi:10.20944/preprints202504.0831.v1

techniques employed to prepare the data for model training, ensuring that inputs are clean, consistent, and meaningful.

3.7.1. K-Nearest Neighbors (KNN) Imputation

KNN imputation addresses missing values by using the weighted average of k nearest neighbors based on a distance metric such as Euclidean distance:

$$x_i = \frac{1}{k}\sum_{j=1}^{k} x_j \quad \text{where } x_j \text{ are the k nearest neighbors} \tag{9}$$

This method captures complex feature relationships, making it effective for financial datasets with correlated variables. Figure 2 illustrates the data distribution and missing value filling process in KNN.

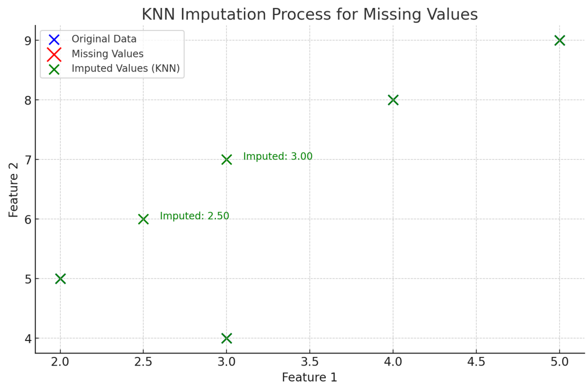

Figure 2. Data distribution and missing value filling in KNN.

3.7.2. Outlier Detection and Treatment

Outliers in financial data, caused by anomalies or errors, can distort model training. We used Z-score normalization to detect and treat outliers:

$$Z_i = \frac{x_i - \mu}{\sigma} \tag{10}$$

where μ and σ are the mean and standard deviation. Data points with $|Z_i| > 3$ were identified as outliers and replaced with the feature median:

$$x_i = \text{median}(X) \quad \text{if } |Z_i| > 3 \tag{11}$$

This method minimizes the influence of extreme values while preserving data integrity.

3.7.3. Correlation Analysis and Feature Engineering

We enhanced model predictive power through correlation analysis and feature engineering:

1) **Feature Removal**: Features with a Pearson correlation coefficient above 0.9 were identified as redundant and removed to reduce noise and dimensionality:

$$\text{Corr}(X_i, X_j) > 0.9 \implies \text{Remove one of } X_i \text{ or } X_j \tag{12}$$

2) **Feature Crossing**: For features with moderate correlation (e.g., 0.5–0.9) or domain-relevant interactions, new features were created by combining existing ones:

$$X_{\text{new}} = X_i \times X_j \tag{13}$$

This captures interactions (e.g., stock price and trading volume) that enhance predictive insights in financial data.

 doi:10.20944/preprints202504.0831.v1

3.7.4. Handling Imbalanced Data with SMOTE

Financial datasets often exhibit class imbalance, leading to biased predictions favoring the majority class. To address this, we used Synthetic Minority Over-sampling Technique (SMOTE), which generates synthetic samples for the minority class by interpolating between existing samples:

$$x_{\text{new}} = x_{\text{minority}} + \lambda \cdot (x_{\text{nearest}} - x_{\text{minority}}) \tag{14}$$

where x_{minority} is a minority class sample, x_{nearest} its nearest neighbor, and λ a random number in [0, 1]. SMOTE balances the dataset, enabling the model to learn representations for both classes effectively.

4. Evaluation Metrics

To evaluate the performance of the models, we used several standard metrics, including accuracy, precision, recall, and F1-score. Additionally, for imbalanced datasets, the Area Under the ROC Curve (AUC) was employed to measure model discrimination ability.

1) **Accuracy**:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{15}$$

where TP is true positives, TN is true negatives, FP is false positives, and FN is false negatives.

2) **Precision**:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{16}$$

3) **Recall**:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{17}$$

4) **F1-Score**:

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{18}$$

5) **AUC** (Area Under the Curve):

$$\text{AUC} = \int_0^1 TPR(FPR)d(\text{FPR}) \tag{19}$$

where TPR is the true positive rate and FPR is the false positive rate.

5. Experimental Results

Table 1 shows a comparison between the different models, demonstrating the advantages of our ensemble approach in terms of performance metrics.

The changes in model training indicators are shown in Figure 3.

Table 1. Model Performance Comparison.

Model	Accuracy	Precision	Recall	AUC
RF	0.78	0.75	0.76	0.81
LightGBM	0.85	0.83	0.80	0.88
XGBoost	0.84	0.81	0.78	0.87
Ensemble Model	**0.87**	**0.85**	**0.82**	**0.90**

The ensemble model outperformed individual models in all metrics, showing that combining predictions from multiple models leads to a more robust forecasting system.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 April 2025 doi:10.20944/preprints202504.0831.v1

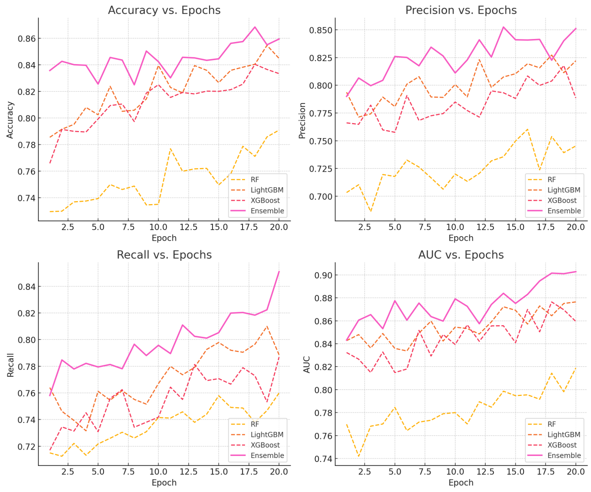

Figure 3. Model indicator change chart.

6. Conclusion

In this paper, we proposed a multi-model ensemble approach using LightGBM, XGBoost, and logistic regression, along with a custom focal loss function to address class imbalance in financial forecasting tasks. Our model demonstrated superior performance in comparison to individual models, making it an effective tool for financial prediction. Future work will focus on expanding the model to incorporate more complex deep learning architectures and further improving the handling of imbalanced datasets.

References

1. Vuong, P.H.; Dat, T.T.; Mai, T.K.; Uyen, P.H.; et al. Stock-price forecasting based on XGBoost and LSTM. *Computer Systems Science & Engineering* **2022**, *40*.
2. Lu, J.; Long, Y.; Li, X.; Shen, Y.; Wang, X. Hybrid Model Integration of LightGBM, DeepFM, and DIN for Enhanced Purchase Prediction on the Elo Dataset. In Proceedings of the 2024 IEEE 7th International Conference on Information Systems and Computer Aided Education (ICISCAE). IEEE, 2024, pp. 16–20.
3. Wang, Y.; Shen, G.; Hu, L. Importance evaluation of movie aspects: aspect-based sentiment analysis. In Proceedings of the 2020 5th International Conference on Mechanical, Control and Computer Engineering (ICMCCE). IEEE, 2020, pp. 2444–2448.
4. Li, S.; Zhou, X.; Wu, Z.; Long, Y.; Shen, Y. Strategic deductive reasoning in large language models: A dual-agent approach. In Proceedings of the 2024 IEEE 6th International Conference on Power, Intelligent Computing and Systems (ICPICS). IEEE, 2024, pp. 834–839.
5. Jin, Y. GraphCNNpred: A stock market indices prediction using a Graph based deep learning system. *arXiv preprint arXiv:2407.03760* **2024**.
6. Lu, J. Enhancing Chatbot User Satisfaction: A Machine Learning Approach Integrating Decision Tree, TF-IDF, and BERTopic. In Proceedings of the 2024 IEEE 6th International Conference on Power, Intelligent Computing and Systems (ICPICS). IEEE, 2024, pp. 823–828.
7. Wang, Y.; Wang, A.; Wang, D.; Wang, D. Deep Learning-Based Sensor Selection for Failure Mode Recognition and Prognostics Under Time-Varying Operating Conditions. *IEEE Transactions on Automation Science and Engineering* **2024**.
8. Xu, J.; Wang, Y. Enhancing Healthcare Recommendation Systems with a Multimodal LLMs-based MOE Architecture. *arXiv preprint arXiv:2412.11557* **2024**.
9. Sun, Y.; Xiang, Y.; Zou, D.; Li, N.; Chen, H. A Multi-Objective Recommender System for Enhanced Consumer Behavior Prediction in E-Commerce. *preprint* **2024**.
10. Li, S. Harnessing multimodal data and mult-recall strategies for enhanced product recommendation in e-commerce. In Proceedings of the 2024 4th International Conference on Computer Systems (ICCS). IEEE, 2024, pp. 181–185.

 doi:10.20944/preprints202504.0831.v1

11. Lu, J. Optimizing e-commerce with multi-objective recommendations using ensemble learning. In Proceedings of the 2024 4th International Conference on Computer Systems (ICCS). IEEE, 2024, pp. 167–171.
12. Yu, H.; Yu, C.; Wang, Z.; Zou, D.; Qin, H. Enhancing healthcare through large language models: A study on medical question answering. *arXiv preprint arXiv:2408.04138* **2024**.