Research and Analysis on XGBoost, Random Forest, and Their Hybrid Models in Retail Sales Forecasting

Zedong Wang

Department of Mathematical Sciences, Xi'an Jiaotong-Liverpool University, Suzhou, 215123, China
sgzwa133@liverpool.ac.uk

Abstract. Retail sales forecasting plays a crucial role in inventory management, marketing planning, and supply-chain optimization, and the choice of model directly influences how well retailers respond to demand volatility, promotions, and seasonal patterns. With the growing complexity of retail sales forecasting, the use of machine learning models for data analysis has become increasingly important. This study reviews relevant literature to summarize the performance of XGBoost and Random Forest in retail sales forecasting, and analyzes the effectiveness of hybrid models combining these two algorithms. XGBoost features high prediction accuracy, effective regularization to mitigate overfitting, and strong capability in modeling nonlinear relationships and imbalanced data. Random Forest demonstrates high stability, robustness to noise, and excellence in feature importance analysis. Hybrid models combining XGBoost and Random Forest achieve more accurate and reliable forecasting results than single models. This study provides practical guidance for model selection in retail sales forecasting. In future work, more competitive hybrid models will be further explored and evaluated.

Keywords: Retail Sales Forecasting, XGBoost, Random Forest, Hybrid Model, Machine Learning.

1 Introduction

Retail sales forecasting is one of the most essential tasks in the retail industry. By predicting future sales, retailers get the opportunity to adjust and optimize inventory management according to the prediction result, ensure marketing resources allocation, improve staffing planning, and achieve sustainable business development. With the rapid development of machine learning, sales forecasting has evolved from simple methods in statistical terms to advanced predictive models. As the need of prediction result becomes complex and multi-dimensional, it should construct a model that combines dynamic daily factors and static store characteristics.

Among lots of prediction models, XGBoost has been widely recognized as a powerful model for retail sales forecasting. Sun states that XGBoost incorporates regularization terms in its objective function to reduce overfitting, enhance

© The Author(s) 2026
L. C. Yoong (ed.), *Proceedings of the 2026 2nd International Conference on Financial Innovation and Marketing Management (FIMM 2026)*, Advances in Economics, Business and Management Research 408,
https://doi.org/10.2991/978-94-6239-801-6_138

generalization, improve training robustness, and handle data noise effectively in regression prediction tasks [1]. Agarwal demonstrates that XGBoost outperforms other common models (including Random Forest) in retail sales forecasting, excelling at modeling complex nonlinear relationships and achieving higher prediction accuracy and generalization ability [2]. Fatima et al. compared Random Forest and XGBoost across various tasks, finding that XGBoost generally outperforms Random Forest in handling complex nonlinear relationships, missing values, and imbalance data, making it a more accurate choice for retail sales forecasting [3].

While XGBoost model is widely regarded as state-of-the-art for structured data prediction tasks, Random Forest also has a wide-spread impact. The analysis of Random Forest and XGBoost for retail sales prediction by Andrabi et al. shows that both models achieve high accuracy, with only a small performance gap between them, confirming Random Forest as a competitive choice [4]. Rohaan et al.'s B2B sales forecasting study found Random Forest significantly outperformed other models, confirming its high effectiveness for retail sales forecasting applications [5]. Klionskiy et al. demonstrated that Random Forest can effectively estimate feature importance, eliminate redundancy, and improve the accuracy of feature selection and analysis [6]. Tabak and Alp found that Random Forest achieved the highest accuracy across all data periods and performs well with limited or imbalanced data [7].

Although XGBoost and Random Forest have shown competitive performance respectively, relying solely on one model cannot achieve optimal prediction results. For this reason, recent studies have focused on hybrid approaches that fuse the two models to further improve forecasting accuracy. Ezeilo et al. stated that hybrid machine learning models integrate the advantages of different algorithms to deliver more stable and accurate forecasts in complex retail environments [8]. Islam et al. proposed a hybrid RF-XGBoost-LR model that combines the strengths of individual algorithms and outperforms single models in sales prediction [9]. Li et al. proposed a two-layer RF-XGBoost hybrid model that outperforms single models and effectively handles complex, high-dimensional, and heterogeneous sales data [10]. Ismail et al. verified that a stacking ensemble combining Random Forest, XGBoost, and SVM yields higher prediction accuracy than single models in sales forecasting [11].

This study reviews relevant studies to demonstrate the impact of XGBoost and Random Forest in retail sales forecasting, and then introduces the advantages of hybrid models combining XGBoost and Random Forest.

2 Comparison of XGBoost and Random Forest Applications in Retail Sales Forecasting

2.1 XGBoost in Retail Sales Forecasting

Gradient Boosting Decision Tree (GBDT) is a traditional ensemble learning algorithm based on the boosting framework, which constructs predictive models by iteratively training weak decision trees. As an optimized and improved version of GBDT, XGBoost has become one of the most widely used tree-based models in recent years.

Furthermore, Sun claimed that XGBoost introduces additional regularization terms in the objective function to control model complexity, which can effectively reduce overfitting and improve generalization performance [1]. Meanwhile, it can increase the robustness of model training, which allows it to adjust to a variety of challenging regression prediction tasks, and reduce the interference caused by data noise.

In a study of retail sales forecasting, the XGBoost algorithm demonstrated significant improvements over traditional forecasting methods across multiple product categories [12]. Compared with the original method which achieved an average MAE of 1.30 and RMSE of 1.94, the XGBoost model reduced the average MAE to 0.92 and RMSE to 1.27, representing an accuracy improvement of 29.23% and 34.54%. The results showed strong alignment between predicted and actual sales values, confirming the efficacy of XGBoost in capturing complex sales patterns and enabling daily sales forecasts instead of only monthly predictions.

As Agarwal reported, XGBoost showed its outstanding capability in modeling complex nonlinear relationships and feature interactions within retail sales data [2]. Due to features' meticulous design, precision preprocessing techniques, thorough model evaluation and hyperparameter optimization, XGBoost achieved a test accuracy of 97.20%, which was higher than Random Forest, decision trees and linear regression models. It also obtained lower RMSE and MAE, showing better prediction accuracy and generalization ability.

Fatima et al.'s study has compared the performance of Random Forest and XGBoost across various machine learning tasks [3]. It showed that XGBoost outperforms Random Forest, especially in handling imbalanced data, missing values, and complex non-linear relationships in most scenarios. Despite being tree-based methods, they differ widely in how they address important issues with real-world data. With its gradient boosting framework, XGBoost performs better when handling complex non-linear relationships, missing values, and class imbalance. In both regression and classification tasks, it outperforms Random Forest in terms of prediction errors and accuracy by using regularization and iterative error correction to minimize overfitting. Random Forest, on the other hand, is less effective in situations with imbalanced data or complex patterns since it depends on bagging and tree diversity, which makes it easier to deploy. These findings demonstrate that XGBoost is a more accurate choice for forecasting retail sales, where data frequently includes missing values, seasonal variations, and a variety of influencing factors.

2.2 Random Forest in Retail Sales Forecasting

While XGBoost model is widely regarded as state-of-the-art for structured data prediction tasks, Random Forest also has a wide-spread impact. Consistent with the mainstream view that gradient boosting models often outperform bagging methods on structured data, XGBoost generally achieves higher accuracy than Random Forest in retail sales forecasting tasks. In a comparative analysis of machine learning algorithms for retail sales prediction by Andrabi et al., both models were evaluated using a real-world dataset, and the results showed that while XGBoost achieved a slightly higher test accuracy of 0.9834, Random Forest still obtained a competitive accuracy of 0.9732

[4]. The small performance gap of only 0.0102 between the two models indicates that Random Forest can deliver strong and reliable results, even when compared to the state-of-the-art boosting model.

Rohaan et al. conducted a comparative study of machine learning algorithms for B2B sales forecasting and found that Random Forest achieved the highest precision compared with Gradient Boosting Classifier and Logistic Regression [5]. At the 99% confidence level, the study confirmed that Random Forest performed significantly better than Gradient Boosting Classifier, with the confidence interval of their performance difference not containing zero. Random Forest, even with a minor accuracy disadvantage in some cases, remains a highly effective choice for retail sales forecasting applications.

According to the study by Klionskiy et al., the random forest classifier is a meta-estimator that can refine the results of correlation analysis and improve the accuracy of the final feature set [6]. The authors used ExtraTreesClassifier, ExtraTreesRegressor, and RandomForestClassifier to estimate feature importance for twelve features. Among these methods, RandomForestClassifier can clearly identify which features are important and which are redundant. The researchers noted that by recognizing feature importance, removing redundant features, evaluating dependencies between features, and handling nonlinear relationships, Random Forest can effectively enhance analytical accuracy.

This phenomenon can be explained by the inherent characteristics of Random Forest. As discussed in the comparative analysis by Jin, Random Forest is the least likely model to overfit to noisy data and outliers, which are prevalent in retail sales data due to promotional events and seasonal fluctuations [13]. In contrast, XGBoost's iterative boosting mechanism is more prone to overfitting when hyperparameters are not carefully tuned.

In the study by Tabak and Alp, they used sales data from 284 stores of a major retail company in Turkey and evaluated Random Forest, XGBoost, and Gradient Boosting with four macro-averaged performance metrics [7]. They divided the data into three groups: pre-pandemic, pandemic, and the combined period to examine whether changes in consumer behavior caused by the pandemic affected model performance. The final results showed that Random Forest achieved the highest accuracy in all three groups, at 97.67%, 97.66%, and 98.91% respectively. Tabak and Alp noted that RF not only performs well with limited or imbalanced data but also benefits from flexible tuning strategies [7].

2.3 Hybrid Models Combining XGBoost and Random Forest

As mentioned earlier, both XGBoost and Random Forest have respective advantages in retail sales forecasting, and each performs differently under various conditions. Therefore, some researchers have proposed hybrid models that combine the two algorithms. As stated by Ezeilo et al., hybrid machine learning models are developed to integrate complementary strengths of different algorithms [8]. Such models can effectively capture complex patterns in consumer behavior, promotional impacts, seasonal changes, and multi-channel characteristics in dynamic retail environments. By

combining different ensemble methods, hybrid approaches overcome the shortages of single models and achieve more stable and accurate predictions. They exhibit better performance in common evaluation metrics, making them more suitable for practical and complicated retail forecasting scenarios.

Islam et al. introduced a hybrid model called RF-XGBoost-LR [9]. In this framework, Random Forest is used to reduce variance and prevent overfitting, while XGBoost improves prediction accuracy and reduces training error. Using data from 45 stores and 99 departments across various regions in the United States over a three-year period, they compared the error metrics and goodness-of-fit between single models and the hybrid model. Experimental results show that the RF-XGBoost-LR model achieved lower MSE and MAE values as well as the highest R^2 score. According to Islam et al., this hybrid model not only overcomes the inherent limitations of individual models but also achieves superior performance across multiple evaluation metrics [9].

Li et al. demonstrated the superior performance of the hybrid Random Forest and XGBoost model in sales forecasting [10]. The team conducted experiments using agricultural product sales data from a supermarket in China from 1 July 2020 to 30 June 2023. In the first layer, Random Forest was adopted to generate initial predictions and calculate the corresponding residuals. Then, hierarchical clustering was applied to classify the characteristics of the residuals and generate residual clustering features. In the second layer, XGBoost was employed to learn from these residual clustering features and output correction terms. The final forecasting result was obtained by summing the initial prediction of RF and the correction prediction of XGBoost. The experimental results showed that the proposed RF-XGBoost model outperformed standalone models in all evaluation metrics, including MAE, MSE, RMSE, MAPE, and R^2, and also achieved competitive performance compared with other RF-based hybrid models. Li et al. argued that the RF-XGBoost hybrid framework can not only be good at processing high-dimensional features with complex interaction relationships and handling nonlinear and heterogeneous data, but also compensate for the shortcomings of individual models in automatically capturing temporal trends, seasonality, and time-series dependencies [10].

This is also supported by the study of Ismail et al. [11]. They combined Random Forest, XGBoost, and SVM and constructed a stacking ensemble model. By using the stacked generalization ensemble technique, several models were trained on supermarket sales records, and their predictions are combined with the help of another model. The experimental results show that the stacking ensemble model achieved the best R^2 value of 0.9990, as well as lower MAE and MSE metrics compared with standalone models. Ismail et al. claimed that ensemble methods, especially Stacking, significantly improve prediction accuracy compared with single models [11].

3 Conclusion

In summary, XGBoost and Random Forest have their respective advantages in retail sales forecasting, and their hybrid models integrate complementary strengths while addressing individual limitations. XGBoost offers strong generalization ability,

effective regularization to reduce overfitting, and superior performance in modeling complex nonlinear relationships, imbalanced data, and noisy retail datasets. In contrast, Random Forest provides high stability, robustness to outliers, reliable feature importance analysis, and consistent performance across diverse data scenarios. Hybrid models combining XGBoost and Random Forest overcome the inherent limitations of single models. Whether constructing a stacking ensemble model or using hierarchical prediction with residual correction and summation, these hybrid approaches significantly improve forecasting accuracy. These findings confirm the great value of prediction models and their hybrid structures in supporting data-driven decision-making for retail operations. This study only analyzes several hybrid models, which is limited in quantity. In future study, more competitive hybrid models need to be explored and evaluated.

References

1. Sun, Y.: Sales prediction based on machine learning approach. In: Proceedings of the 2024 9th International Conference on Social Sciences and Economic Development (ICSSED 2024). Advances in Economics, Business and Management Research (2024). https://doi.org/10.2991/978-94-6463-459-4_113
2. Agarwal, P.: Enhancing financial forecasting in erp systems using xgboost: a robust sales prediction model. In: 2nd International Conference on Emerging Technologies and Sustainable Business Practices-2024 (ICETSBP 2024). Atlantis Press (2024). https://doi.org/10.2991/978-94-6463-544-7_26
3. Fatima, S., Hussain, A., Amir, S.B., Awan, M.G.Z., Ahmed, S.H., Aslam, S.M.H.: XGBoost and Random Forest algorithms: An in depth analysis. Pakistan Journal of Scientific Research 3(1), 26–31 (2023). https://doi.org/10.57041/vol3iss1pp26-31
4. Andrabi, S.H.H., Alice, K., Srivastava, S.A.: Sales Forecasting using XGBoost. TechRxiv, 07 November 2022 (2022). https://doi.org/10.36227/techrxiv.21444129.v1
5. Rohaan, D., Topan, E., Groothuis-Oudshoorn, C.G.M.: Using supervised machine learning for B2B sales forecasting: A case study of spare parts sales forecasting at an after-sales service provider. Expert Systems with Applications 188, 115925 (2021). https://doi.org/10.1016/j.eswa.2021.115925
6. Klionskiy, D.M., Razmochaeva, N.V., Geppener, V.V., Popov, N.V.: Study of Data Mining Algorithms for Retail Sales Management. In: Conf. paper, Jun (2024)
7. Tabak, K., Alp, S.: Hyperparameter Optimization in Machine Learning Classification Models: A Case Study in the Retail Industry. Journal of Artificial Intelligence and Technology (2026). https://doi.org/10.37965/jait.2026.0945
8. Ezeilo, O.J., Ikponmwoba, S.O., Chima, O.K., Ojonugwa, B.M., Adesuyi, M.O.: Hybrid machine learning models for retail sales forecasting across omnichannel platforms. Shodhshauryam, International Scientific Refereed Research Journal 5(2), 175–190 (2022)
9. Islam, M.T., Ayon, E.H., Ghosh, B.P., Chowdhury, M.S., Shahid, R., Roy puja, A., Rahman, S., Bhuiyan, M.S., Nguyen, T.N.: Revolutionizing Retail: A Hybrid Machine Learning Approach for Precision Demand Forecasting and Strategic Decision-Making in Global Commerce. Journal of Computer Science and Technology Studies 6(1), 33–39 (2024). https://doi.org/10.32996/jcsts.2024.6.1.4

10. Li, J., Lin, B., Wang, P., Chen, Y., Zeng, X., Liu, X., Chen, R.: A hierarchical RF-XGBoost model for short-cycle agricultural product sales forecasting. Foods 13(18), 2936 (2024). https://doi.org/10.3390/foods13182936
11. Ismail, M., Tukur, H.M., Friday, M.: Sales prediction using ensemble machine learning model. International Journal of Scientific Research and Modern Technology (IJSRMT) 4(3), 24–35 (2025). https://doi.org/10.38124/ijsrmt.v4i3.350
12. Dankorpho, P.: Sales Forecasting for Retail Business using XGBoost Algorithm. Journal of Computer Science and Technology Studies 6(2), 136–141 (2024). https://doi.org/10.32996/jcsts.2024.6.2.15
13. Jin, Y.: Sales Prediction Based on Machine Learning Approach: Support Vector Machine, Xgboost and Random Forest. Highlights in Science, Engineering and Technology 88, 104–109 (2024). https://doi.org/10.54097/6dxq4p06

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (http://creativecommons.org/licenses/by-nc/4.0/), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

