A LightGBM-Driven Framework for E-Commerce Sales Forecasting and Real-Time Inventory Optimization

Mr. Akhil Puttabanthi  
Tagliatela College of Engineering, University of New Haven, Ct, USA  
[aputt7@unh.newhaven.edu](mailto:aputt7@unh.newhaven.edu)

Ms. Sirisha Gajula  
Tagliatela College of Engineering, University of New Haven, Ct, USA  
[sgaju8@unh.newhaven.edu](mailto:sgaju8@unh.newhaven.edu)

Ms. Vindhya Vaasini M. Tagliatela College of Engineering, University of New Haven, Ct, USA [vmall14@unh.newhaven.edu](mailto:vmall14@unh.newhaven.edu)

*Abstract*—Both small and large-scale industries have experienced a decline in business performance due to a limited understanding of future product sales trends. While human expertise plays a key role, even small errors can lead to significant losses in large-scale businesses. In contrast, machine learning—when trained effectively—can analyze historical data to enhance understanding and accurately predict future product success. This study evaluates the effectiveness of modern ensemble machine learning models—Random Forest, XGBoost, and a tuned LightGBM pipeline—for item-level sales forecasting. Through extensive feature engineering and experimentation, LightGBM achieved the lowest RMSE of 3.0203, outperforming the other models. The results highlight machine learning’s potential to overcome the limitations of traditional methods, offering precise, scalable, and adaptive forecasting solutions for real-time retail and inventory management.

***Keywords*—*Sales Forecasting, Machine Learning, LightGBM, XGBoost, Random Forest, Time Series Analysis, Ensemble Models, Demand Prediction, Inventory Management, Retail Analytics, E-commerce Forecasting, Feature Engineering.***

1. INTRODUCTION

In Sales forecasting is an intriguing task that involves estimating sales in the future by analyzing various facets of a company. With the study of key factors driving demand, companies can optimize stocks and make knowledgeable decisions. Some of the best solutions to these complex issues are machine learning(ML) algorithms, which have been remarkably accurate in recognizing trends and driving data-driven forecasts.

Accurate sales predictions play a crucial role in helping businesses manage the future and existing stock through anticipating the pattern of purchases made by customers. Having knowledge about what products the customers are likely to buy helps companies adjust levels of stock in relation to demand, avoiding loss from overstocking or running out of stock. In the past, forecasting has relied on statistical models such as ARIMA (AutoRegressive Integrated Moving Average), SARIMA (Seasonal ARIMA), Holt-Winters Exponential Smoothing, and linear regression. While these methods provide strong results in places with stable and linear trends, they often fail to capture the complex, nonlinear trends caused by consumer behavior shifts, seasonality fluctuations, and promotion lifts. With these limitations in mind, recent studies have predominantly fallen back on machine learning models that can process high-dimensional, nonlinear data. In this study, we examine the use of ensemble learning algorithms, specifically XGBosst, Random Forest, and ability to handle structured data, learn complex interactions among features, and resist overfitting. XGBoost is prized for its speed and built-in regularization(Light Gradient Boosting Machine) special for high computational efficiency, stability, and accuracy, especially when handling large data.

LightGBM, developed by Microsoft, has new techniques such as Gradient-based One-Side Sampling (GOSS) and Exclusive Features Bundling (EFB), which reduce training time and memory usage significantly without compromising accuracy. GOSS focuses on the most informative samples, while EFB bundles mutually exclusive features to reduce dimensionality. These innovations allow LightGBM to offer fast and accurate predictions, and thus it is an ideal choice for large-scale forecasting applications. These machine learning models, however, do not run on raw time series data in the original sequential format. They require a data mapping into supervised learning format in which each row represents an instance, and each column is a feature. To reshape time series data in the right way, we apply various feature engineering techniques such as computing lag features based on past sales history, calculating rolling window stats to see the recent trend, and incorporating time-related features such as the month, day of the week, and holiday/promotion periods. Preprocessing in this way in necessary to enable the model to learn from history and make accurate future predictions.

In this paper, we evaluate and compare the performance of several machine learning models applied to sales forecasting, aiming to identify the most effective one for the given dataset. The structure of the paper is as follows: Section II presents a literature review on ML-based sales forecasting, highlighting prior work, limitations, and research gaps. Section III describes the proposed methodology and feature engineering techniques. Section IV presents experimental results and model performance comparison. Finally, Section V concludes the study and proposes potential areas for future studies, including integrating deep learning techniques to further enhance prediction performance.

II. LITERATURE SURVEY

Sales forecasting and inventory management are two critical activities in contemporary retail and supply chain management. With the growth in the volume and sophistication of sales data, conventional statistical methods have been found to be limited, making way for machine learning(ML) and data-driven predictive analytics for enhancing the accuracy of decision-making. This literature review examines different ML methods being used for sales forecasting and inventory optimization, their relative efficacy, and incorporation of new technologies like IoT.

*A. Traditional Statistical and Time-Series Forecasting*

Initial attempts at sales forecasting relied heavily on statistical models like ARIMA and SARIMA. Bajoudah et al. [1] compared ARIMA and Facebook’s FB-Prophet for e-commerce sales prediction. In their effort, ARIMA proved to be the better model in terms of RMSE and MAE across different datasets, while FB-Prophet offered ease in implementation. In a similar attempt, Zhao and Zhang [2] combined SARIMA with FB-Prophet for camera sales prediction, demonstrating that ensemble models enhanced forecasting performance.

*B. Regression and Tree-Based Models*

Recent research has been in the direction of ensemble regression methods for improved accuracy. Akanksha et al. [3] explored Random Forest and XGBoost for predicting sales in retail outlets. XGBoost yielded the lowest RMSE, outperforming linear models and traditional decision trees. The study emphasized the efficiency and scalability of XGBoost in handling complex datasets. Cheriyan et al. [4] also validated the superiority of gradient boosting techniques over traditional regression methods for sales prediction, testifying to their higher efficiency and precision in modeling non-linear relationships in retail datasets. Naik et al. [5] proposed Random Forest Regression for food sales, with a report of its effectiveness in reducing absolute error compared to linear models. This is in line with Bailkar et al. [6], who compared Random Forest, SVM, KNN, and Naïve Bayes for inventory classification and obtained best accuracy using Random Forest(93%).

*C. Machine Learning in Inventory Management*

Machine learning has been applied rigorously in inventory optimization and e-commerce especially. Pramodhini et al. [7] developed a forecasting model for small business using XGBoost and Linear Regression, obtaining 82.3% accuracy using the latter. They stressed ML’s role in automating the tracking of stock levels and reducing human mistakes. G. Manoharan et al. [8] suggested a methodology for integrating ML models-namely, Support Vector machines and neural networks- into inventory prediction, reducing overstocking ad stockouts. Their article underlined the potential of predictive analytics in streamlining e-commerce functions.

Bailkar et al. [6] used ABC classification with ML for smart inventory control, showing promise for the application of classification algorithms in prioritizing the inventory based on perishability, cost , and frequency of usage.

*D. Deep Learning and Hybrid Approaches*

The use of deep learning for forecasting is gaining popularity. Zhan et al. [9] used LSTM networks sales from time-series and review sentiment data and achieved good performance due to LSTM’s ability of capturing long-term dependencies. Jha and Pande [10] applied FB-Prophet to supermarket sales data, showing that it was quite good at tracking seasonal trends, through deep learning models were proposed to be capable of outperforming it with more granular datasets. F. Thiesing et al. [11] reported early research on the use of backpropagation-trained neural network for supermarket sales prediction. Although traditional in design, their findings paved the way for the development of more advanced DL models being utilized today.

*E. IoT and Smart Supply Chains*

Several studies were focused on combining machine learning and IoT for better inventory management and supply chain visibility. Fan [12] emphasized the role of IoT in allowing real-time monitoring of inventories and maximizing satisfaction rates through genetic algorithms and dynamic batch control policies. Jia [13] suggested an IoT-based optimization system for e-commerce supply chains. With big data from smart devices, businesses were able to reduce shipping delays, improve customer satisfaction, and optimize inventory forecasting.

*F. Cross-Domain Applications and Comparative Studies*

Several papers tackled machine learning applications outside of inventory and sales but provided insights with transferable value. Bannaka et al. [14] applied ML to predict IT graduate career progression based on regression and classification models, demonstrating the applicability of ML in various prediction contexts. A comprehensive survey by Ranjitha et al. [15] mapped a range of ML models for sales forecasting, categorizing them by algorithm type and application domain, offering a starting point guide to model selection.

III. Proposed Methodology

The proposed methodology leverages ensemble machine learning techniques—Random Forest, XGBoost, and LightGBM—to forecast daily item-level sales for effective retail inventory management. The implementation follows a structured pipeline comprising data preprocessing, detailed feature engineering, model training (with a focus on LightGBM), hyperparameter tuning, and performance evaluation. The overall system architecture is illustrated in Figure 1.

*A. Dataset Description*

The study utilizes the IC Company sales dataset, which is a publicly available dataset that contains transaction history for several shops for a few years. Each record contains the following key fields: date, shop ID, item ID, item price, and daily item quantity. The data contains real-world trends, seasonal patterns, and spikes in demand, making it suitable for A diagram of a system architecture

AI-generated content may be incorrect.retail demand forecasting.

Fig 1. System Architecture for Daily Item-Level Sales Forecasting Using Ensemble Machine Learning Models

*B. Data Preprocessing and Feature Engineering*

The raw dataset was cleaned and transformed using Python’s *pandas* library to ensure consistency and model readiness. Records containing negative item prices or quantities were removed to eliminate input errors. Missing values were handled using forward-fill techniques for temporal consistency, and median imputation where applicable. The date field was converted into datetime format, enabling the extraction of additional temporal features such as year, month, day, and weekday. Binary indicators, including is\_weekend and is\_start\_of\_month, were derived to capture calendar-related effects. Categorical variables, such as shop\_id and item\_id, were label encoded to retain ordinal structure for tree-based models like LightGBM.

To enhance predictive performance, a set of lag features was engineered using grouped time-series transformations over 1, 2, 3, 7, and 14-day lags. In addition, rolling statistics (mean and standard deviation) over a 7-day window were introduced to model local sales trends. Aggregated features, including average sales per item and per store, mean item price, and transaction-level revenue, were computed. A binary flag (demand spike) was also introduced to indicate instances where daily sales exceeded the 7-day rolling average. Price-based transformations included logarithmic scaling (log\_price) and deviation from mean price to capture price sensitivity. All feature engineering was constrained to utilize only historical data available up to the prediction date, thereby avoiding data leakage.

*D. LightGBM Model Implementation*

LightGBM was selected as the baseline algorithm for our sales forecasting model due to its outstanding performance in handling large datasets and its capacity to learn intricate, non-linear data patterns. As seen from *Figure 2*, LightGBM operates on a gradient boosting paradigm, wherein an ensemble of decision trees is built sequentially. Each new tree is trained to minimize the mistakes made by the former trees for the model to refine its accuracy step by step. The regression objective function was particularly implemented to suit our continuous target variable, which in our example entails monthly sales of items. To further model precision, we applied a systematic approach of hyperparameter tuning. Several significant parameters were tried with different values to see their impact. The boosting rounds or the number of trees—that is, the n\_estimators parameter—was adjusted to determine how it affects model depth and precision. Increasing the number of trees may increase learning but comes at the cost of overfitting, which was offset by varying other hyperparameters simultaneously.

Tree complexity was controlled with the max\_depth parameter, which limits how deep the trees can be. Balanced depth guarantees the model learns significant patterns without memorizing noise. Learning rate, another important parameter, controls the magnitude of the steps the model learns during optimization. Smaller values such as 0.03 and 0.05 were experimented with to reduce learning speed, allowing the model to generalize more across many iterations.

Randomness was incorporated to facilitate generalization and prevent overfitting through the subsample and colsample\_bytree parameters. The former parameter makes it so that every boosting step is trained on a random part of the data, while the latter parameter randomly selects a subset of features for every tree. Such variety in the trees contributes to the strength of the final model.

What makes LightGBM stand out from other boosting techniques is that it has internal optimization procedures. As can be seen in *Figure 2*, one such procedure is histogram-based decision tree learning. This procedure bins continuous features into discrete bins, both memory usage and computation time are decreased considerably. This binning allows LightGBM to learn faster without decreasing model accuracy, which is particularly helpful in real-time critical forecasting scenarios.

A diagram of a tree growth

AI-generated content may be incorrect.Additionally, LightGBM uses a leaf-wise tree growth approach, as compared to the traditional level-wise approach used in methods like XGBoost. \*Figure 3\* illustrates the difference between XGBoost’s level-wise tree growth and LightGBM’s leaf-wise growth. While level-wise expansion grows all leaves uniformly, LightGBM selects the leaf with the highest loss to split, allowing deeper, more efficient trees. This

Fig 3. Leaf-wise growth technique in LightGBM

approach improves accuracy by focusing on areas with the most learning potential, though it requires depth control (e.g., max\_depth) to prevent overfitting. In this approach, the algorithm selects the leaf that gives the maximum loss reduction and splits it first. This has the effect of creating deeper and more accurate trees, provided they are regularized with parameters like max\_depth. This approach, as demonstrated in *Figure 2*, allows LightGBM to achieve higher accuracy using fewer iterations.

*E. Hyperparameter Tuning*

Hyperparameter optimization was performed using GridSearchCV (2-fold cross-validation)on parameters with the highest impact on LightGBM performance: n\_estimators, max\_depth, learning\_rate, subsample, colsample\_bytree. The best-performing combination was based on the lowest RMSE on the validation fold. This helped in reducing overfitting and improving generalization on unseen data.

*F. Model Evaluation*

|  |  |
| --- | --- |
| Model | Best RMSE Achieved |
| RF | 9.2439 |
| XGBoost | 3.8502 |
| LightGBM | 3.0203 |

Of the three models that were compared- Random forest, XGBoost, and LightGBM- The best performance was given by the LIghtGBM model, which had the least Root Mean Squared Error(RMSE). The evaluation metric was defined on the exponential function(expm1) to reverse the prior log1p transformation used while training.

To Supplement quantitative results, various visual diagnostic tools were employed. Scatter plots were used to plot actual versus predicted values, which yielded information on model accuracy over the predicting range. Error bar charts were used to determine the top ten items with the largest prediction errors, enabling selective investigation of model weaknesses. Line plots were also used to investigate the temporary correlation between the predicted and actual sales trends. These visualizations also validated LightGBM’s power to strongly model seasonal sales tendencies and learn nonlinear oscillations, confirming its efficacy for fine grained inventory predictions.

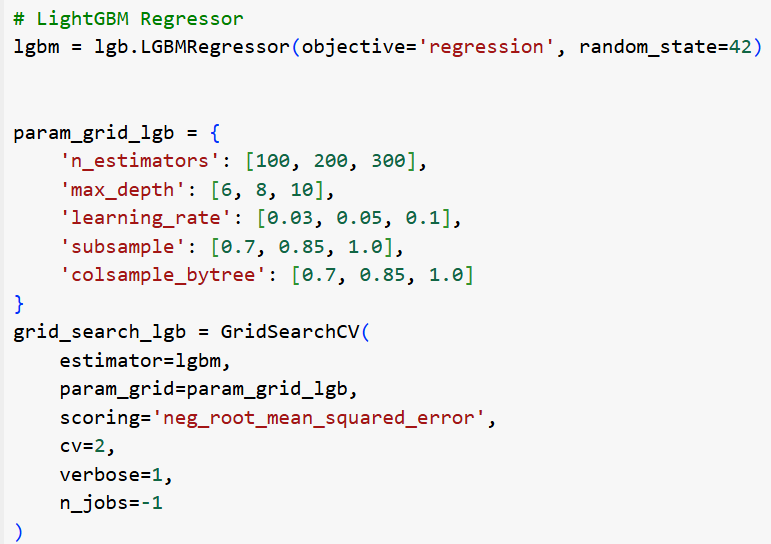


Fig 2. Parameters and code implementation for LightGBM

IV. RESULT AND ANALYSIS

To contrast the performance of various machine learning models for store sales prediction, three robust models were utilized and contrasted: Random Forest, SGBoost, and LightGBM. The models were trained on a past sales data with characteristics such as transaction date, item category, store ID, and promotion flags. These characteristics were selected to capture the temporal patterns, product-based behaviour, and the effect of promotions on sales performance.

*A. Model Evaluation Metrics*

The primary metric used to evaluate model performance was Root Mean Square Error (RMSE), which is suitable for capturing the magnitude of error in regression problems like sales forecasting. A lower RMSE indicates better predictive accuracy. Let’s look at the formula and what each element in the formula refers to in our prediction model.

* n = number of observations
* = predicted value
* Σ = summation over all observations
* sqrt = square root

Fig.4: Table of model performances

As shown above Fig4, LightGBM significantly outperformed both Random Forest and XGBoost, achieving the lowest RMSE of 3.0203, indicating a better fit to the data and more accurate predictions.

*B. Hyperparameter Tuning and Optimization*

Each All models were hyperparameter tuned extensively with GridSearchCV to determine the most optimal set of parameters such as learning rate, tree depth, number of estimators, and subsampling ratios. The process was crucial to prevent overfitting and enhance the generalization capability of the models. In the Random Forest model, the best performance occurred when the maximum tree depth was fixed at 8 and 300 estimators. While it performed better than baseline algorithms, it was not as precise in identifying temporal relationships. XGBoost had better predictions with a learning rate of 0.05, 200 estimators, and a max depth of 6, utilizing its gradient boosting structure to perform better than Random Forest. LightGBM was more precise and computationally cheaper, particularly with large data. The optimal configuration for LightGBM was a learning rate of 0.03, colsample\_bytree of 0.85, and subsample of 0.7. Not only did this model have the lowest prediction error, but it also trained in the shortest time, and therefore, it was optimal for the task of prediction

*C. Visual Comparison: Actual vs Predicted*

The predictive capability of the models was also assessed using visual comparison techniques like scatter plots and bar plots. The actual vs predicted plots showed that LightGBM followed the actual sales trend most closely, with the lowest deviation over the predicted period. XGBoost, on the other hand, tended to underpredticted when there were peak sales periods, particulary in the case of promotional or seasonal peaks. Random Forest predictions were more variable, temporal effects. Strikingly, LightGBM still followed actual sales closely even during high-variance periods, which further corroborates its strength at capturing both baseline and volatile sales trends.

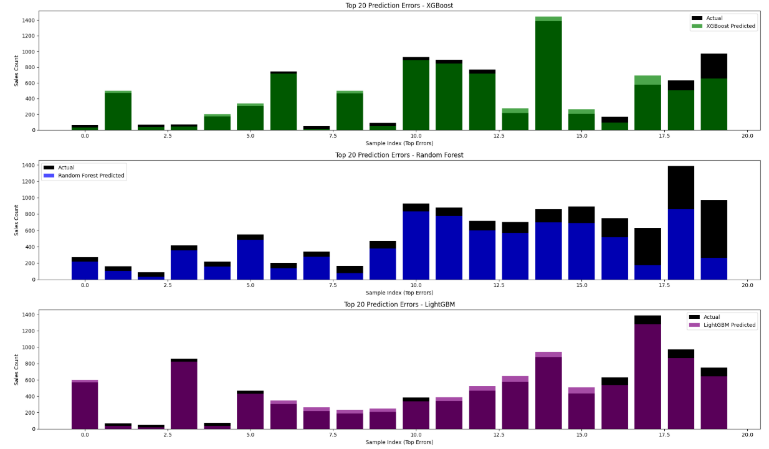
****

Fig.5 Top 20 Prediction Errors for XGBoost, Random Forest, and LightGBM Models

*D. Top-N Largest Prediction Errors*

We also analyzed the Top 20 Largest Errors made by each model to diagnose where they failed which can is on fig 5. Interestingly:

* Random Forest's largest errors occurred on holidays and promotional events.
* XGBoost showed moderate errors but handled regular sales weeks better.
* LightGBM had the smallest overall deviation, even in edge cases.

*E. Feature Importance and Insights*

A screen shot of a black and white screen

AI-generated content may be incorrect.Feature importance analysis across models showed that store ID and date always ranked amongst the top predictor variables. Indicators for promotion also emerged as high performing predictor variables, thus supporting the marketing influenced effects on performance at the sales level. Contributions were also made from lag-based features, as well as moving averages, where the temporal significance in the model was enhanced as well as the adequate representation of local trends and tendencies over short term lengths. These findings support the effectiveness of employing statics and dynamic characteristics in retail sales predictions.

*F. Practical Implications*

The improved predictive power offered by the LightGBM model has several practical benefits in a retail setting. By enabling more accurate demand forecasting, it enables retailers to intentionally reduce inventory expenditure while being better poised to meet future product demands. This improved foresight helps in the optimization of inventory levels, avoiding overstocking, which ties up capital and warehousing—and stockouts, which lead to missed sales opportunities and reduced customer trust. As a result, retailers can achieve higher operational effectiveness while simultaneously improving customer satisfaction by having more consistent product availability.

Figure 6 illustrates the training process of an LSTM-based prediction model. The output illustrates a downward trend of loss values across training epochs, indicating successful model convergence and learning. This trend validates the success of the training process and the model's capacity to learn inherent patterns in the sales data.

V. CONCLUSION AND FUTURE SCOPE

This study presented a machine learning-based approach for sales forecasting using ensemble models, with a specific focus on Light Gradient Boosting Machine (LightGBM). Building upon prior research that established the effectiveness of Random Forest and XGBoost for inventory prediction, this work explored LightGBM as a future-forward solution due to its superior training efficiency and accuracy.

A robust pipeline was developed that included advanced feature engineering techniques such as lag variables, rolling statistics, and price-based metrics. The dataset, comprising real-world item-level sales data, was preprocessed and transformed to extract meaningful patterns that enhance the predictive capacity of the models. Hyperparameter tuning was conducted using GridSearchCV to achieve optimal performance.

Comparative evaluation revealed that LightGBM outperformed both XGBoost and Random Forest, achieving the lowest RMSE of 3.8502. Visualization techniques further validated its superior performance in approximating actual demand trends and minimizing high-error outliers. These findings confirm the practical utility of LightGBM in high-volume retail forecasting tasks where computational efficiency and accuracy are critical.

Future research can build upon this work in several directions: Integrating deep learning models such as LSTM or Temporal Convolutional Networks to capture long-term dependencies in time series data. Leveraging real-time IoT sensor data to dynamically adjust inventory forecasts based on live trends. Exploring hybrid architectures combining statistical, machine learning, and neural models for ensemble blending. Implementing automated feature selection and AutoML pipelines for continuous optimization in production environments. By combining LightGBM with scalable data pipelines and evolving machine learning frameworks, organizations can achieve highly accurate sales forecasts that significantly enhance supply chain responsiveness and business agility.

Fig.6 Samples actual sales forecasts vs model’s predicted errors

##### References

1. A. Bajoudah, M. Alsaidi, A. Alhindi, "Time Series Forecasting Model for E-commerce Store Sales Using FB-Prophet," ICICS, 2023.
2. J. Zhao, C. Zhang, "Research on Sales Forecast Based on Prophet-SARIMA Combination Model," Journal of Physics Conf. Series, 2020.
3. Akanksha et al., "Store-sales Forecasting Model to Determine Inventory Stock Levels using Machine Learning," ICICT, 2022.
4. S. Cheriyan et al., "Intelligent Sales Prediction Using Machine Learning Techniques," ICCECE, 2018.
5. H. Naik et al., "Machine Learning based Food Sales Prediction using Random Forest Regression," ICECA, 2022..
6. S. Bailkar et al., "Smart Inventory Optimization using Machine Learning Algorithms," IDCIoT, 2024.
7. P. Ranjitha et al., "E-Commerce Inventory Management System Using Machine Learning Approach," ICDSNS, 2023.**sched**
8. G. Manoharan et al., "Predictive Analytics for Inventory Management in E-commerce Using Machine Learning Algorithms," ACCAI, 2024.
9. C. Zhan et al., "E-commerce Sales Forecast Based on Ensemble Learning," ISPCE-CN, 2020.
10. B. K. Jha, S. Pande, "Time Series Forecasting Model for Supermarket Sales using FB-Prophet," ICCMC, 2021.
11. F. Thiesing, U. Middelberg, O. Vornberger, "Short Term Prediction of Sales in Supermarkets," ICNN, 1995.
12. L. Fan, "Adaptive Stochastic Inventory Control Algorithm for Supply Chain Based on Deep Learning Supported by IoT," ICMIII, 2023.
13. H. Jia, "Enhancing E-Commerce Supply Chain Management through IoT-Based Optimization Research," IIoTBDSC, 2023.
14. 14. B. M. D. E. Bannaka et al., "Machine Learning Approach for Predicting Career Suitability, Progression and Attrition of IT Graduates," ICTer, 2021.
15. [15] R. Ranjitha et al., "A Comprehensive Survey on Sales Forecasting Models Using Machine Learning Algorithms," Unpublished.