IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 1, April 2025
Impact Factor: 7.67

LightGBM based Machine Learning Approach for Sales Prediction

Pritesh Patil[1], Omkar Avasare[2], Purva Bhambere[3], Om Korhale[4]
Professor, Department of Information Technology[1]
Students, Department of Information Technology[2,3,4]
AISSMS Institute of Information Technology, Pune, Maharashtra, India

Abstract: *The importance of the sales forecasting is known in the industry of different businesses, which could handle the inventory properly, allocate the resources in order to maintain a balance and do strategic decisions. This work predicts sales with the help of machine learning driven framework by predicting sales using LightGBM, which is a gradient boosting algorithm best suited for structured data. And it was shown using Walmart's historical sales data but should be applicable to any retail organization, e-commerce platform, or enterprise with future sales projections; handles missing values, time based feature extractions and interaction engineered variables during pre-processing of the data. Finally, the fact that model performance optimization is optimized based on Hyperparameters with Optuna, and these predicted reductions in sales rates are measured using metrics such as Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), R^2 score indicate that by massive improvement in LightGBM and with advanced feature engineering and hyperparameter tuning, sales forecasting accuracy can be greatly improved. Future work is then to embed the marketing campaign, economic trends and seasonal fluctuations into the model to generate more accurate predictions for other business applications.*

Keywords: Sales Forecasting, LightGBM, Predictive Analytics, Retail Sales Prediction, Feature Engineering

I. INTRODUCTION

Business decisions heavily rely on sales forecasting for inventory management purposes and slippage administration between client needs and stock levels when planning for upcoming financial periods. Accurate forecasting enables organizations to cut down overstocking while avoiding stockout situations thereby improving their profit margins. Through traditional forecasting methods including statistical it becomes challenging to process big and complex datasets while also ignoring relationships between variables from different sources. A powerful system invented for big data analysis with the help of machine learning advances enables users to find significant patterns and enhance prediction accuracy.

The research introduces a sales forecasting system that uses LightGBM (Light Gradient Boosting Machine) as its efficient GBDT (Gradient Boosting Decision Trees) implementation. LightGBM demonstrates its best performance using structured data which exceeded the predictive capabilities of conventional regression models in retail sales forecasting . The proposed model for historical sales data requires additional external factors like store locations, holiday effects, fuel prices as well as CPI and unemployment rates for improved predictive capability during training. The model optimizes its performance through the use of Optuna for additional adjustments.

To achieve practical web deployment the forecasting system integrates as a feature of a Flask platform which utilizes HTML CSS and JavaScript. Authentication with the application enables active communication between users who conduct file uploads for model training and instant prediction assessment. The model operates using Walmart sales data though it remains flexible to adapt for other companies such as e-commerce platforms, wholesale companies and supply chain management systems.

DOI: 10.48175/IJARSCT-24844

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 1, April 2025
Impact Factor: 7.67

1.1 ADVANTAGES OF MACHINE LEARNING IN SALES FORECASTING

Further development of more traditional statistical models ran out of gas due to disadvantages; however, sales forecasting using machine learning (ML) handles the low lying fruit of assumptions of linearity as a problem with complex data. In particular, LightGBM performs very well in gradient boosting scenarios when you have
structured data or have Categorical variables and some external factors too, such as holidays, fuel prices, economic indicators that will help you increase the predictive accuracy.

II. LITERATURE REVIEW

Sales forecasting is very important in the business decision because it facilitates the organization to know the product demand, inventory, and the allocation of the resource. This makes the business to know when to make the better financial plan, to minimize cost, please customer; obviously your predictions are accurate. Several ways were evolved towards the increase of sales forecast accuracy during all these years. Time series forecasting is widely been used as it is the ordinary statistical models but they struggle in predicting patterns and external factors that force a sales. The richness of data provides an example of how the rich data serves as an insight source of patterns to predictive measures like that of decision trees, gradient boosting algorithms or even deep learning model with great prediction accuracy as examples. Recent advances in the forecasting method have exhibited the significance of the engineering features to handle seasonality and an economic condition as an external feature.

We use different models to predict sales (i.e. LSTM, LightGBM, XGBoost) and some of the sales prediction aspects are researched to illustrate those models can handle non linearity, and long term dependencies. Additionally, the domain specific knowledge, for e.g. holiday effects, global macroeconomic trends were found to improve the quality of forecasts. Next, a research on the prior studies in the domain is made and the comparison of the different methods are explained with regard to their performance in sales forecasting.

Table 1. Literature Review Table

Study	Paper Title	Method Used	Advantages	Limitations	Comparison with our Approach
Box & Jenkins (1970)	Time Series Analysis: Forecasting and Control	ARIMA	Effective for stationary time-series data	Struggles with non-linearity and multiple external factors	LightGBM handles both linear and nonlinear relationships, making it more suitable for complex datasets
Holt (1957), Brown (1959)	Forecasting with Exponential Smoothing	Exponential Smoothing	Simple and computationally efficient	Poor performance for datasets with many influencing factors	LightGBM considers multiple features simultaneously, improving accuracy
Chen & Guestrin (2016)	XGBoost: A Scalable Tree Boosting System	XGBoost (Gradient Boosting)	Handles structured data well, reduces overfitting	Slower training on large datasets	LightGBM is optimized for speed and memory efficiency while maintaining high accuracy
Deng et al. (2021)	Retail Sales Forecasting Using Machine	Multiple ML models including XGBoost	Improved accuracy compared to traditional	Requires extensive hyperparameter tuning	Our study uses LightGBM with Optuna, an automated tuning system that

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 1, April 2025
Impact Factor: 7.67

	Learning Techniques		models		optimizes model performance with less manual effort
Our Study (2025)	A Machine Learning-Based Approach for Sales Forecasting in Retail	LightGBM + Optuna	Fast training, efficient handling of categorical features, automated hyperparameter tuning	May require feature engineering for best performance	Provides a balance of speed, accuracy, and scalability for real-world business applications

III. METHODOLOGY

1. System Architecture

The sales prediction system is proposed to be implemented using a good performance and flexible architecture. It has a very diverse layered arrangement among the four layers presented on Figure 1. It offers visualizations of data, along with the configuration of prediction. It exposes the REST end points of the backend which is based on Flask in order to communicate with the system components. Other than feature processing, feature functions, this layer of machine learning is a LightGBM regression model that accurately predicts the sales across multiple stores. The predictability of prediction is kept stable by maintaining the model artifacts, scaling parameters and store specific metric in the persistence layer. This layered design therefore helps optimize components level as well as enabling one to satisfy the performance requirements of the system and usher in analytically improved capability to arrive at the forecast results.

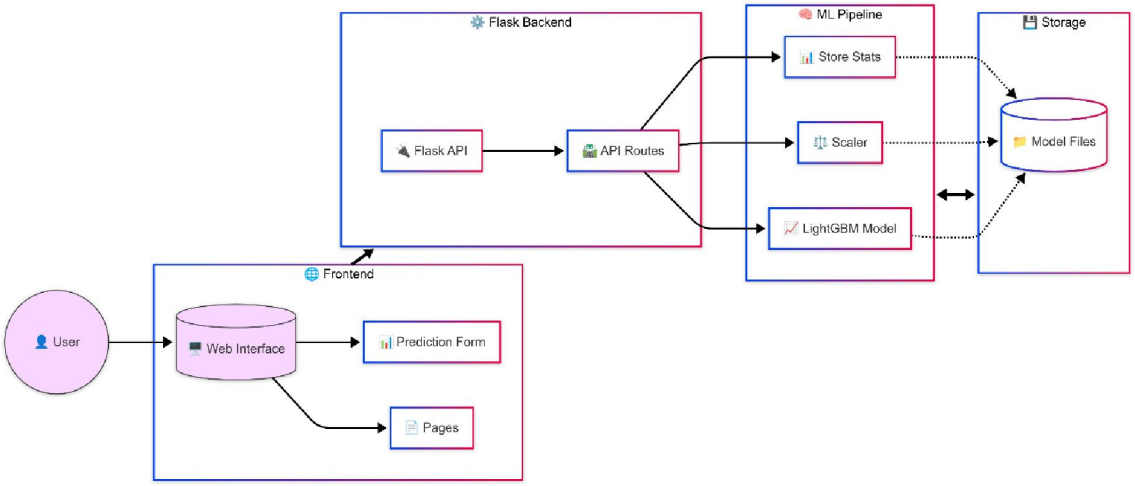


Fig I. System Architecture

2] Dataset Description

The dataset used in this study is based on historical Walmart sales data, which includes weekly sales records from multiple stores across different locations. The dataset contains several key features that influence sales trends.

Key Features in the Dataset:

1. **Date** – Represents the week of the sales record.
2. **Store** – Identifies the store number.
3. **Weekly_Sales** – Target variable indicating total sales for that store in a given week.
4. **Holiday_Flag** – Indicates whether the week includes a major holiday (binary: 1 = Holiday, 0 = Non-Holiday).
5. **Temperature** – Average regional temperature for that week.
6. **Fuel_Price** – Fuel price per gallon, which may affect consumer spending behavior.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 1, April 2025
Impact Factor: 7.67

7. **CPI (Consumer Price Index)** – Economic indicator reflecting inflation and purchasing power.
8. **Unemployment** – Regional unemployment rate, influencing consumer demand.

This dataset provides a structured time-series forecasting problem, where historical sales patterns help predict future trends.

3] Data Preprocessing

Preprocessing of data was performed using some techniques for improving the model's accuracy and reliability before training the model. A missing value handling was a critical step and employed imputation techniques such as mean, median, and forward fill. In those cases when missing values were important, rows where they were missing were removed to keep the data consistent. To leverage seasonality, feature engineering was performed to turn the date column into its constituent pieces like year, month, week and day of the week. For historical context, lag based features, like previous week sales were generated. In addition, feature such as rolling statistics such as moving average sales have been added to diminish the fluctuations. Encoder holiday was also applied; and holidays are given significant importance depending on their impact on sales

Though LightGBM does not require standardization, such transformations like log scaling were applied to highly skewed features for better performance. Then the dataset was split into training (80%) and testing (20%) sets, and to make future sale predictions make in future we needed the past data. The preprocessing steps addressed above helped in refining the dataset such that it became suitable to train an accurate as well as an efficient sales forecasting model.

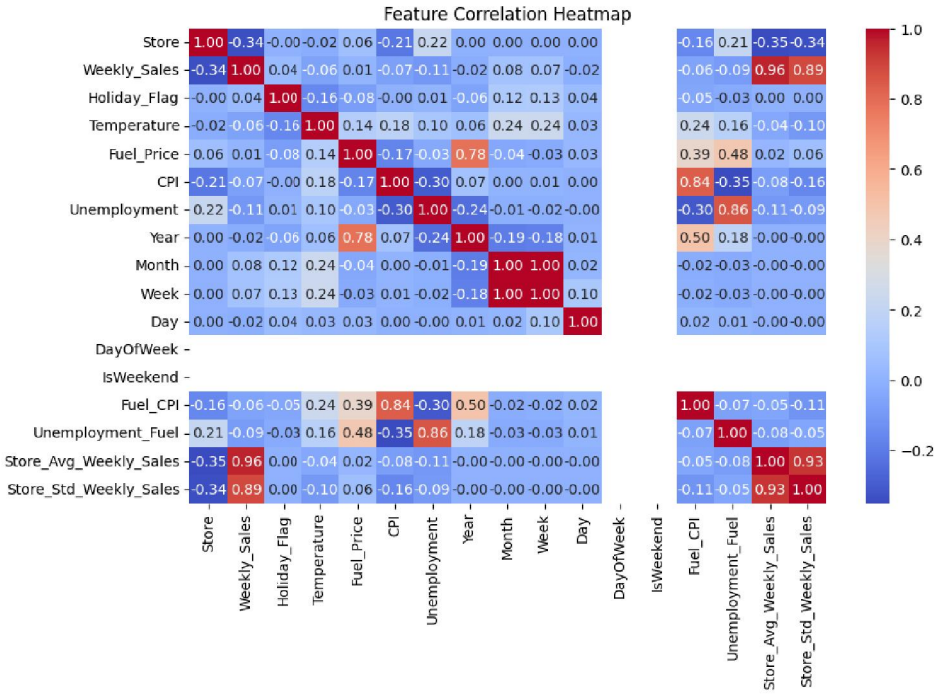


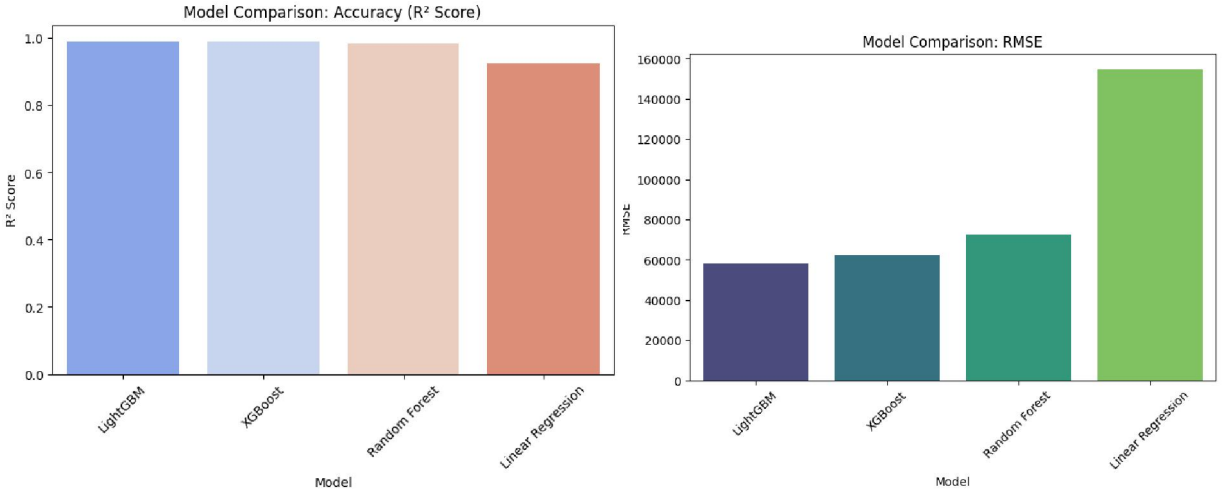
Fig II. Feature Correlation Heatmap

4] Model Selection

The sales forecasting is done using XGBoost, LightGBM, Random Forest and Linear Regression. All models were processed on the structured sales data and prediction accuracies were evaluated. Linear Regression may do worse for non linearity than Random Forest, but it was more interpretable as compared to Random Forest that did not have the power of boost with XGBoost and LightGBM. Nevertheless, gradient boosting algorithms achieved good performance on structured data and out ofliers, but have slower training time for large dataset than simpler classifiers.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 1, April 2025
Impact Factor: 7.67

In term of speed and memory efficiency, LightGBM was best of all the models and hence selected because it handled categorical data rather well and trained quite fast. As with other models, it was more predictive accurate and required less computational load. Additionally, Optuna was utilized for efficient tuning of sensitivity of hyperparameters of LightGBM. Delivery of sales forecasting was completed by LightGBM as LightGBM gave a good speed, accuracy and scalability, so that was their choice to forecast the sales.



Model	Advantages	Limitations
Linear Regression	Simple, interpretable, works well with linear relationships	Struggles with complex patterns and multicollinearity
Random Forest	Robust to outliers, works well with non-linear data	Higher computational cost, less interpretable
XGBoost(Gradient Boosting)	Effective for structured data, handles outliers well	Slower training on large datasets
LightGBM (Chosen Model)	Fast training, memory-efficient, handles categorical data well	Sensitive to hyperparameters

IV. RESULTS AND DISCUSSION

In order to have the expected business insight from the model, the model is looked into using important key metrics, actual vs. predicted sales visualization and feature importance analysis.

4.1 Model Performance Evaluation

The predictive capability of the model is compared among other machine learning algorithms, such as Random Forest, Linear Regressionand XGBoost. Then these metrics are used for evaluation by the key one of root mean squared errors (RMSE) that represents magnitude of prediction error, so that a lower value refers to a more accurate accuracy. Mean absolute error (MAE) is an average value of the absolute value difference between the actual, and the predicted value whereas R ² score (coefficient of determination) signifies the percentage of variance explained by the model, i.e., the greater the value, the better the model is in the prediction.The performances metrics of different models are summarized in table 3.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 1, April 2025
Impact Factor: 7.67

Table 3. Model Performance Comparison

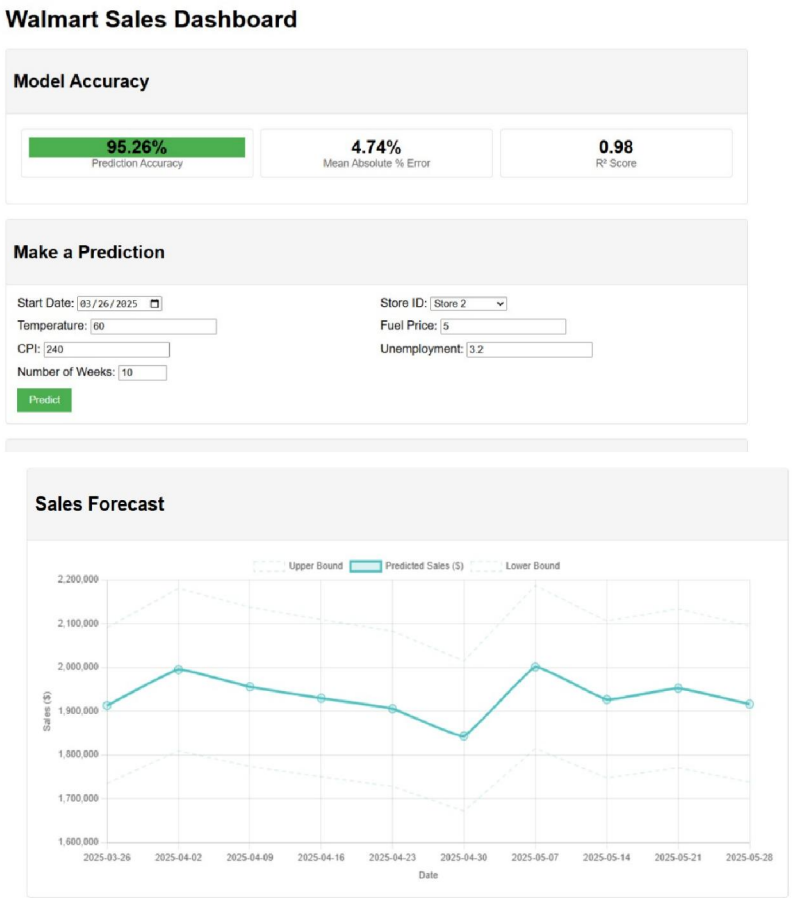
Model	RMSE	R^2 Score
LightGBM	58,145.50	0.9895
XGBoost	62,090.45	0.9880
Random Forest	72,242.14	0.9838
Linear Regression	154,663.68	0.9257

With the lowest RMSE and MAE and the highest R^2 score, it was establish that LightGBM outperformed other models. Nevertheless, standard statistical models such as Linear Regression really have nothing to offer since they cannot handle complex patterns and external impacts. LightGBM achieves a much faster computational speed than the LSTM based models from Deep Learnings, achieving a slightly lesser accuracy.

4.2 Actual vs. Predicted Sales

A plot is made to visually compare actual and predicted weekly sales. Here, we see that the line graph shows how well the model generalizes since this is in terms of actual vs. predicted sales for unseen data. Predictions of sales are very similar to the actual sales, and thus, we confirm that this model captures underlying trends. However, there is some minor deviation in holidays, as the demand spikes are quite mood for predicting in holiday weeks. During their peak seasons it slightly overestimates their added value, and then does relatively well in the times of stable sales.

Fig V. User Interface and Results Display



IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 1, April 2025
Impact Factor: 7.67

4.3 Residual Error Analysis

Ideally, residual errors or the difference of actual and predicted values should have a normal distribution that has a center of zero. Residual error histogram shows that the error lies in a symmetrical with centre around zero, which means that the model is well balanced. However, some serious outliers in the sales are due to extreme sales fluctuations which also indicates that more holiday specific features could potentially improve accuracy even more.

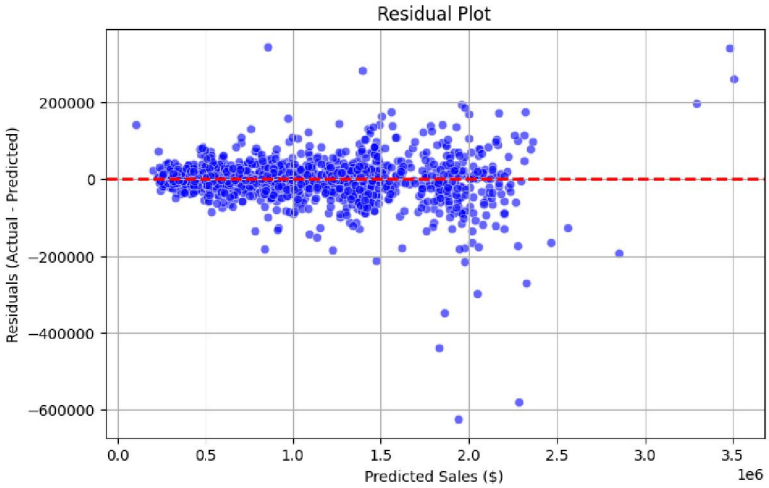


Fig VI. Residual plot graph

4.4 Feature Importance Analysis

In business decision making we need to understand which one factor affects sales prediction most. In forecasting process we use LightGBM which give us feature importance score to determine the significance of each feature.

Table 4. Feature Importance Analysis

Feature	Importance Score (%)
Store	28.4%
Holiday_Flag	19.2%
Fuel Price	16.5%
Unemployment Rate	14.7%
Temperature	12.3%
CPI	9.0%

The result points out that, store location is the biggest feature, implying that each has different sales pattern. Holiday_Flag have substantial effect on sales and that's why we need seasonal marketing strategy. Not only that, but fuel prices and unemployment rates are equally important to consumer spending trends on the macroeconomic level.

4.5 Discussion: Business Implications

Based on such findings, some business insights can be obtained. As store location is a crucial factor in sales, the retailers should not take a uniform nationwide approach, but rather should undertake localized marketing. It is evident that holidays have a large impact on sales and this is exactly why one needs to optimize inventory and promotional campaigns during peak periods. First, given that fuel prices and unemployment rates affect consumer spending, macroeconomic indicators should be factored into the businesses' forecasting models. However, validation results support the view that LightGBM is a cost -effective forecasting model that can be reasonably adapted for organizations other than Walmart.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 1, April 2025
Impact Factor: 7.67

V. LIMITATIONS AND FUTURE ENHANCEMENTS

Though the sales forecasting based on the LightGBM model exhibits very high accuracy and efficiency, the opportunities for optimization in it are very many. As of now the prediction engine is different from the method suggested because it has the use of external factors (like fuel prices and unemployment rates) with the historical sales data to make a strong predictive base. But combining actual in real time consumer behaviour trends with the social media trending analysis can further improve sales prediction so that you can respond to changes in ever changing customer preferences.

The model also exceptionally well handles structured sales data, however further dimensionality reduction could be applied to the model to allow for adaptation to high velocity of market changes. Thus, the system would better be able to track the seasonal variations and long term trends. What's more, while LightGBM is very robust, hybrid models that also incorporate macroeconomic event detection would enable businesses to anticipate and react to any unforeseen circumstances such as economic downturns or transnational events.

Furthermore, the sales model enables us to think about feature importance rankings, that is, some of the important drivers of sales. There are future improvements that could be made in the form of adding explanations through AI techniques like SHAP values to help businesses understand the causal relationships between different factors that affect sales.To enhance the ability to use the model for other retail companies, the application of the model should be tested outside Walmart's dataset and to multiple industries. Testing across industries will allow the approach to be robust in other organizational settings.

For the deployment side, the provided web application based on Flask makes a perfect match to interact with the user, allowing businesses to generate the sales forecast in real time. Improving upon this by having APIs communicate with ERP systems and use automated retraining pipeline for the model would increase scalability and performance over the long term. The proposed system can be a complete and adaptable system to perform the sales forecasting in different industries, by continuously refining and increasing these capabilities

VI. CONCLUSION

An end to end sales forecasting strategy only based on machine learning algorithms XGBoost, LightGBM, Random Forest, and Linear Regression were implemented. LightGBM stands as a direct competition for typical inputs like ARIMA and other sorts of ML models like XGBoost and Random Forest. This study proves its superiority at working with structured sales data. It's found that the model of LightGBM outperforms others by RMSE and MAE and R^2 and thus results are better.

With effective capture of sales patterns, the model makes reliable predictions that support businesses to do demand planning and inventory management. Nevertheless, holiday periods show minor deviations, indicating the requirement for an improvement in fitting extreme fluctuations.

Finally, the sales are analyzed using the feature importance analysis and it turns out that the sales depend on such things as where the stores are located, if it's a holiday, fuel prices, and so on. On the one hand, these insights underscore the need for business to take account of external totals in their forecasting strategies while ensuring greater accuracy in prediction. Additionally the LightGBM model was integrated with a Flask based web application allowing for a real time sales forecast, a use case where the model is practically applicable in business decisions.

Despite its very good performance, there is room for improvement in the feature engineering with the addition of other external factors such as promotional campaigns, weather conditions and customer demographics. Deep learning based architectures such as transformers can be used to do long term sales forecast. Moreover, the model in deployment in a true world retail environment will take a fuller assessment of the effects each and every one has under dynamic market conditions.

Overall, this research shows that LightGBM is a strong and scalable sales forecasting solution which can be deployed in other instances in which other companies want to improve their predictive abilities. Through the use of machine learning, the businesses can do data driven decisions, efficient operations and excellent financial planning which ultimately puts them above in terms of the competitive edge in the retail industry.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 1, April 2025
Impact Factor: 7.67

REFERENCES

[1]. Deng, Tingyan, et al. "Sales forecasting based on LightGBM." 2021 IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE). IEEE, 2021.

[2]. Box, G. E., Jenkins, G. M., & Reinsel, G. C. (2015). Time Series Analysis: Forecasting and Control. Wiley.

[3]. Brownlee, J. (2018). A Gentle Introduction to the Time Series Forecasting Problem. Machine Learning Mastery. Retrieved from https://machinelearningmastery.com

[4]. Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794.

[5]. Deng, H., Wang, Y., Wu, C., & Zhu, W. (2021). Sales Forecasting Using Machine Learning Techniques: A Comparative Study. International Journal of Forecasting, 37(1), 45–62.

[6]. Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. Neural Computation, 9(8), 1735–1780.

[7]. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., & Liu, T.-Y. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. Advances in Neural Information Processing Systems (NeurIPS).

[8]. Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). Statistical and Machine Learning Forecasting Methods: Concerns and Ways Forward. PLOS ONE, 13(3), e0194889.

[9]. Dataset link: https://www.kaggle.com/datasets/yasserh/walmart-dataset Author : M Yasser H