International Journal of Science and Research (IJSR)
ISSN: 2319-7064
Impact Factor 2024: 7.101

Comparative Evaluation of Machine Learning Models for Retail Sales Forecasting: A Multi-Algorithm Approach

Mayank Dwivedi[1], Supriya Mishra[2]

[1]Wayne State University (Computer Science), Detroit, MI (USA).
Corresponding Author Email: *mayank.dwivedi[at]wayne.edu*

[2]University of Massachusetts (Computer Science), Lowell, MA (USA).
Email: *supriya_mishra[at]student.uml.edu*

Abstract: ***Accurate sales forecasting is vital for retail operations, impacting inventory management and strategic planning. This study explores the application of advanced machine learning models- Linear Regression, Random Forest, XGBoost, Support Vector Regression, LSTM, and ARIMA to improve forecasting accuracy. A five-year dataset of retail transactions was preprocessed using differencing and lag features to enhance stationarity. Among all models tested, XGBoost demonstrated the highest predictive accuracy (R^2: 0.989), outperforming traditional methods. The research provides a scalable framework for real-time forecasting, underlining machine learning's transformative role in the retail sector and offering practical implications for inventory optimization and customer demand prediction.***

Keywords: retail forecasting, machine learning, XGBoost, sales prediction, time series analysis

1. Introduction

Forecasting sales for retailers is a critical task that directly impacts inventory management, financial planning, and overall business strategy. Accurate sales forecasting enables retailers to determine the necessary inventory levels to meet consumer demand, thereby avoiding the costly issues of understocking and overstocking. Traditional sales forecasting methods often fall short due to their inability to account for the complex interactions among factors such as price, consumer income, and seasonal trends. These limitations necessitate the adoption of more sophisticated approaches, such as machine learning techniques, which have shown promise in transforming customer data into meaningful insights and facilitating strategic decisions.

Recent advancements in machine learning have markedly improved the accuracy and performance of sales forecasting models. Traditional methods, like time series models, tend to overfit due to limited historical sales records of new items and fail to make accurate predictions (Li et al., 2023) . In contrast, machine learning techniques such as recurrent neural networks offers strong capabilities for capturing complex nonlinear patterns and learning from temporal sequences (Elalem et al., 2023). The integration of sales data into forecasting models has been shown to greatly improve accuracy, especially during periods of maximum risk to the public (Dolan et al., 2023).

Despite these advancements, retail sales forecasting still faces several challenges. The COVID-19 pandemic, for instance, has significantly shifted consumer behavior towards e-commerce, highlighting the need for accurate sales forecasting amidst uncertainties (Abdullhadi et al., 2024). Additionally, there is ongoing debate on whether aggregate forecasts should be generated independently or through hierarchical methods, as each approach has its own implications for accuracy (Oliveira & Ramos, 2019).

In response to these challenges, our study explores the use of various machine learning models to enhance the accuracy of retail sales forecasting. We utilized methods including Linear Regression, Random Forest Generator, XGBoost, Support Vector Regression, LSTM, and ARIMA model. XGBoost demonstrated superior performance in predicting customer buying behavior with high accuracy.

The purpose of this study is to evaluate the predictive effectiveness of various machine learning algorithms for retail sales forecasting and propose a robust forecasting framework based on comparative model performance. This study contributes to retail analytics by showcasing how machine learning enhances forecasting accuracy, which is critical for optimizing stock levels, reducing costs, and improving customer satisfaction in a dynamic market environment.

2. Materials and Methods

Data Collection

Data collection incorporated both internal and external sources. The analysis began with sales data for specific items at various locations on given dates. This dataset spans five years and includes multiple stores and a diverse range of products. It captures trends, seasonality factors, and the influence of advertising and marketing embedded within the sales figures. Some algorithms also utilize customer reviews along with sales figures to bring the customer satisfaction index into the forecast. These reviews typically include numerical product ratings along with textual content, providing valuable depth and insight into consumer opinions (Schneider & Gupta, 2016).

Data Cleaning and Transformation

We performed data cleaning and transformation to address issues like duplicate entries, missing values, and outliers. Duplicate data were removed, missing values were imputed

using statistical methods, and outliers were identified. The data was then normalized to ensure consistency and comparability. We removed the trend and seasonality using differencing techniques, which were crucial for improving forecasting accuracy. By applying the differencing method, we effectively addressed these irregularities, enhancing the model's robustness (Parpoula, 2024). Once the model is trained and starts making predictions, we add the trend back to get the final predictions.

Feature Selection and Engineering

Feature selection and engineering were conducted to enhance model performance. We used feature selection algorithms and similarity-based methods to identify the most relevant features for sales prediction. Independent component analysis was employed to extract features from sales data, which were then used to improve the performance of linear regression models. For our models, we have created a new data frame where each feature represents the previous month's sales. To determine the number of months to include in our feature set we chose to look-back period to be 12 months, therefore, generated a data frame that has 13 columns, 1 column for each of the 12 months and the column for our dependent variable, difference in sales.

Machine Learning Algorithms

We explored various machine learning algorithms, including ARIMA (Autoregressive Integrated moving average), Linear Regression, Random Forest Generator, XGBoost, LSTM (Long Short-term memory), Support Vector Regression. Each algorithm was configured and trained independently using the collected and preprocessed data. These models were trained and optimized to capture patterns and relationships within the data. To ensure efficient deployment and reproducibility, the trained model has been separately bundled using a .pkl file and is made available for further use. The comparison technique was used to forecast sales using different models, to cover a wide range of retailers and varying features and data sets.

Training and Validation Process

The training and validation process utilized cross-validation techniques to ensure robust model performance. The training dataset consisted of 1,801 distinct datasets incorporating 13 features. It was constructed using a 12-month lag time series analysis, drawing from five years of historical sales data segmented by date, location, and item. In total, the dataset comprised approximately 913,000 rows, providing a comprehensive foundation for model training and evaluation. The validation set included 12 different data sets including 13 features. Additionally, we employed MinMaxScalar to scale the data between the feature range of (-1, 1) so that all our variables fall within defined range. After running our models, we used a helper function to reverse scale the data to its original values.

Model Evaluation

We used root mean square error and mean absolute error to evaluate model performance. R^2 score was utilized to get the score predicted values compared to actual sales. R^2 is the square of correlation coefficient between observed and predicted values in a regression model.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}|$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2$$

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2}$$

$$R^2 = 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2}$$

$$where: \quad \hat{y} - predicted\ value\ of\ y$$
$$\bar{y} - mean\ value\ of\ y$$

3. Results

Data Collection and Preprocessing

In this study, the data set comprises sales transactions recorded by date, store, and item, providing a granular view of sales patterns across different locations and products. The data preprocessing is required for data quality contributes to better prediction, achieve legal compliances and boost operational efficiencies (Tawakuli et al., 2024).

The data is first aggregated and validated to ensure completeness, consistency, and accuracy. Missing values, if present, are handled through interpolation or imputation techniques. The data below is aggregated by month and year.

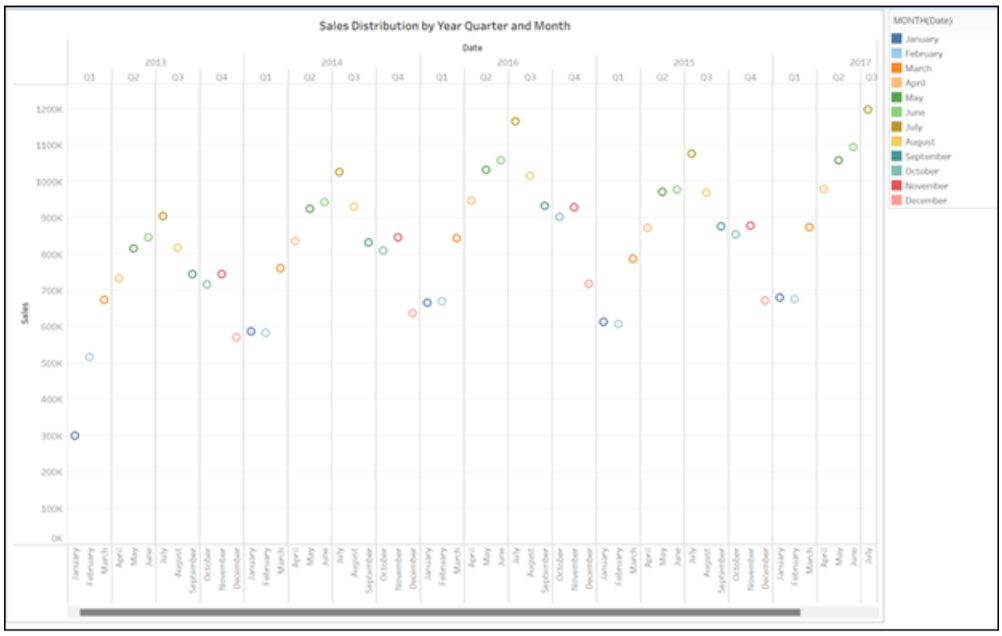

Figure 1: Results of sales distribution per year, quarter, month

Data Exploration (EDA – Exploratory Data Analysis)

Preprocessing optimized the dataset for effective model training. We aggregate the sales further by date at company level. To make it stationary, we calculated the difference between the sales of subsequent days to form a time series using differencing techniques and added it to our data frame as new columns. The figure below represents the Monthly Sales before difference transformation.

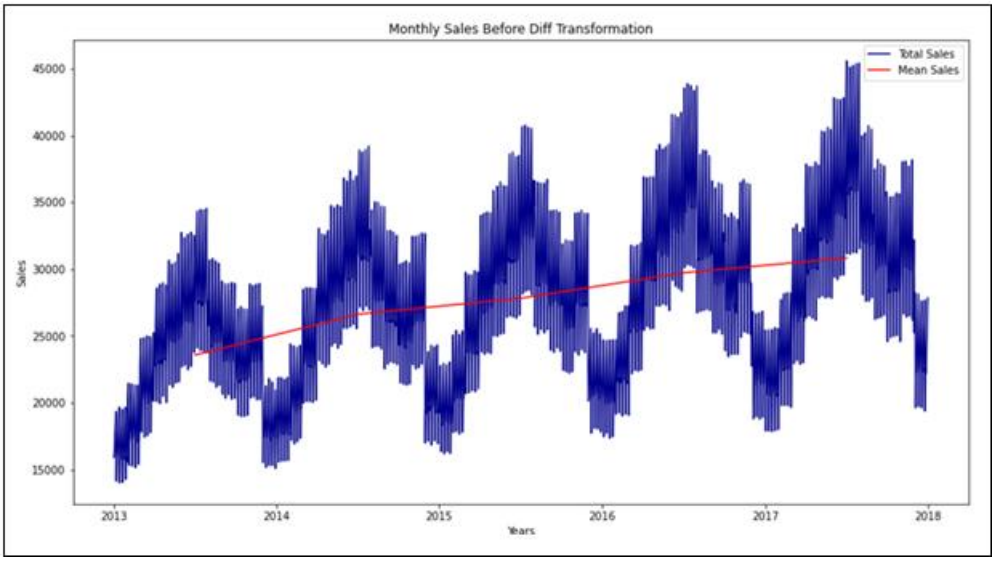

Figure 2: Results of sales distribution per month before differencing

We applied the differencing method to compute the difference between consecutive terms in the series and to get rid of the varying mean. The sales diff is aggregated at month level and the figure below represent the month sales after differencing

Sales data by day along with sales diff will be made available on Git repository.

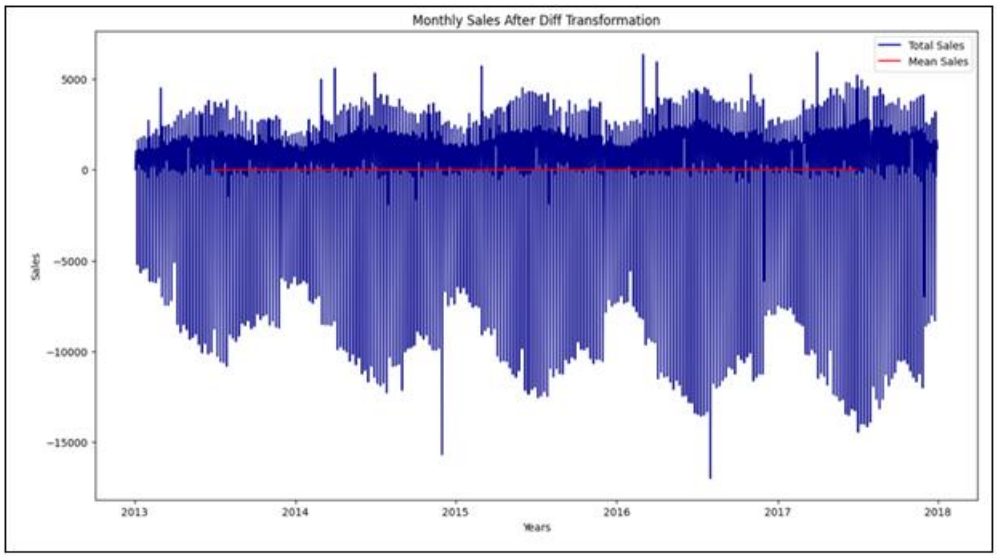

Figure 3: Results of sales distribution per month after differencing

Data Sources and Description

The data set used in this study consists of daily aggregated sales records across multiple stores and items, providing a comprehensive view of sales trends over time. Each data point includes the total sales for a given date, capturing fluctuations in demand across different periods. By analyzing the sales differences instead of raw sales values, the model can better identify underlying patterns and dependencies. Further transformation involved restructuring the dataset, where each row represented a day, and columns included total sales, dependent variables, and lagged sales values for different delays. Based on exploratory data analysis (EDA), 12 lagged sales features were introduced to capture temporal dependencies, enabling the model to learn from past sales behavior.

Table 1

Sales data by day with dependent variables, and previous sales for each delay (The sample model data is provided in the model development section and additional data will be made available on request)

date	sales	sales_diff	lag_1	lag_2	lag_3	lag_4	lag_5	lag_6	lag_7	lag_8	lag_9	lag_10	lag_11	lag_12
1/14/2013	14000	-5665	926	1119	789	730	320	1654	-5221	1005	584	1057	777	11
1/15/2013	15772	1772	-5665	926	1119	789	730	320	1654	-5221	1005	584	1057	777
1/16/2013	15799	27	1772	-5665	926	1119	789	730	320	1654	-5221	1005	584	1057
1/17/2013	16840	1041	27	1772	-5665	926	1119	789	730	320	1654	-5221	1005	584
1/18/2013	17666	826	1041	27	1772	-5665	926	1119	789	730	320	1654	-5221	1005

Model Development

Selection of Machine Learning Algorithms

Multiple machine learning algorithms were employed to train on the existing dataset and predict future sales. Specifically, Linear Regression, Random Forest Regressor, XGBoost Gradient, Support Vector Regression, LSTM- a recurrent neural network and ARIMA- a time series-based model was selected. Linear Regression served as a baseline model, providing insights into linear dependencies between features. The Random Forest Regressor, an ensemble learning method, was utilized to capture nonlinear patterns and interactions between variables, improving predictive accuracy. XGBoost, a gradient boosting algorithm known for its high performance and robustness, was employed to further enhance the model forecasting capability by effectively managing missing data, handling outliers, and reducing overfitting. These models were trained using the preprocessed dataset, and their performance was evaluated to determine the most effective approach for accurate sales forecasting. The interesting arguments have been documented on the forecasting models, presenting the framework for regression-based models on how complexities can be reduced and how models and methods can be further evaluated (Bojer, 2022).

Linear Regression:

Our independent features consist of sales differences along with time series data, including a 12-month lag (denoted as **X**), while the corresponding sales (**Y**) serve as the dependent variable.

Assuming a linear relationship between **X** and **Y**, the sales forecast can be represented as:

$$\hat{Y} = \boldsymbol{\theta_1} + \boldsymbol{\theta_2 X}$$

or equivalently,

$$\hat{y}_i = \boldsymbol{\theta_1} + \boldsymbol{\theta_2 x_i}$$

where:

- $y_i \epsilon Y$ are the actual sales values (labels in supervised learning).
- $x_i \epsilon X$ for are the input independent training. Mathematically training set can be represented as $\boldsymbol{X_t} = \boldsymbol{f(X_{t-1}, X_{t-2}, \ldots, X_{t-12})}$

where $\boldsymbol{X_t}$ represents the sales data at time t, and the lagger feature $\boldsymbol{X_{t-1}, X_{t-2}, \ldots, X_{t-12}}$ captures past trends over the previous 12 months.

The model determines the best regression fit to predict the sales over the span of time by finding the best values of $\boldsymbol{\theta_1}$ & $\boldsymbol{\theta_2}$where:

- $\boldsymbol{\theta_1}$ is the intercept
- $\boldsymbol{\theta_2}$ is the coefficient of X.

The Linear Regression model has been trained on the same dataset referenced above. The trained model has been saved as "**linear_regression_trained_model.pkl"** and is included in the data and code repository mentioned in the Supplementary Materials section.

Table 2: Train and Test data sample

sales_diff	lag_1	lag_2	lag_3	lag_4	lag_5	lag_6	lag_7	lag_8	lag_9	lag_10	lag_11	lag_12
-5665	926	1119	789	730	320	1654	-5221	1005	584	1057	777	11
1772	-5665	926	1119	789	730	320	1654	-5221	1005	584	1057	777
27	1772	-5665	926	1119	789	730	320	1654	-5221	1005	584	1057
1041	27	1772	-5665	926	1119	789	730	320	1654	-5221	1005	584
826	1041	27	1772	-5665	926	1119	789	730	320	1654	-5221	1005

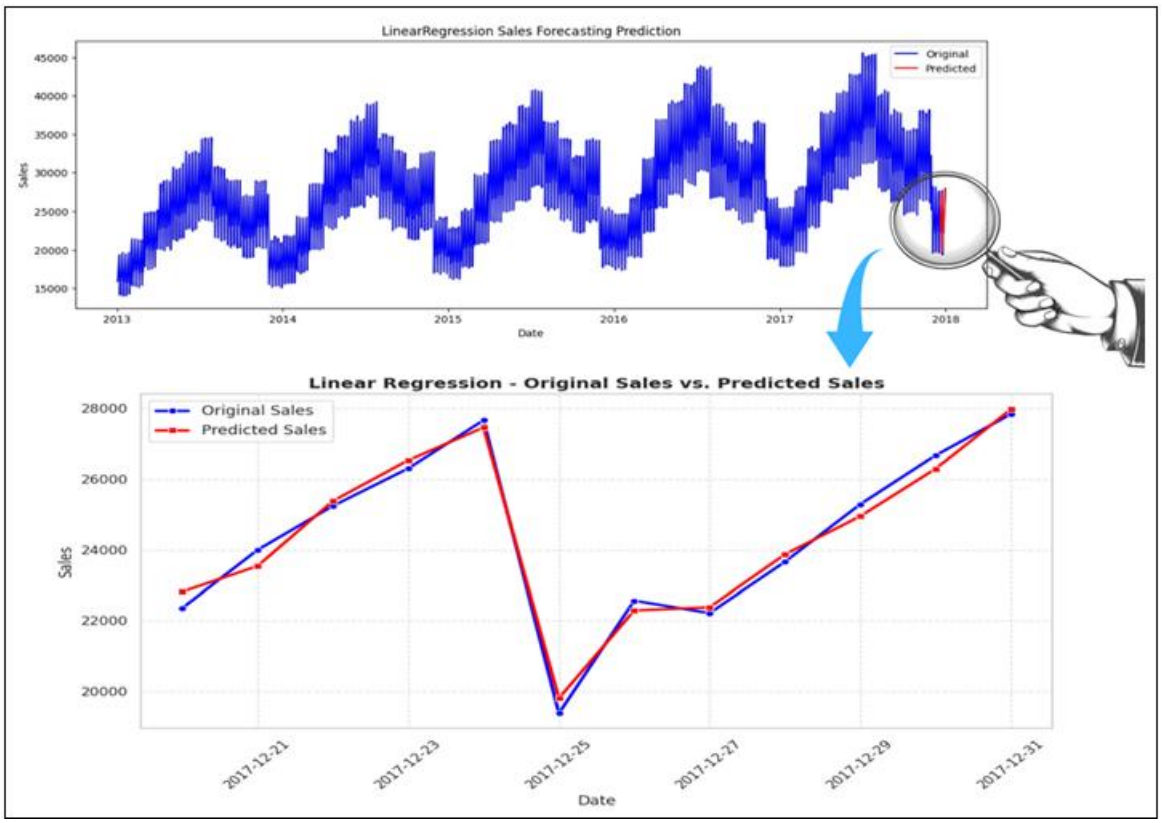

Figure 4: Linear Regression Model on Sales Data - Original vs Prediction

Linear Regression Model Score	
Shape of Train	(1801, 13)
Shape of Test	(12, 13)
RMSE (Root Mean square error)	308.9278503
MAE (Mean absolute error)	285.4166667
R2 Score	0.983923872

Random Forest Regression:

Unlike a linear relationship, the **Random Forest Regressor** model leverages an ensemble of decision trees to capture complex patterns in the data. The sales forecast is generated by aggregating the predictions from multiple decision trees, reducing variance and improving accuracy. Mathematically, the training set can be represented as:

$$\boldsymbol{X_t = f(X_{t-1}, X_{t-2}, \ldots, X_{t-12})}$$

where $\boldsymbol{X_t}$ represents the sales data at time t, and the lagger feature $\boldsymbol{X_{t-1}, X_{t-2}, \ldots, X_{t-12}}$ captures past trends over the previous 12 months.

The **Random Forest Regressor** model learns patterns by constructing multiple decision trees and averaging their outputs to predict future sales. It determines the best possible regression fit by minimizing errors through feature selection and tree-based learning mechanisms.

The random forest regressor model has been trained on the same dataset referenced above. The trained model has been saved as **"random_forest_trained_model.pkl"** and is included in the data and code repository mentioned in the Supplementary Materials section.

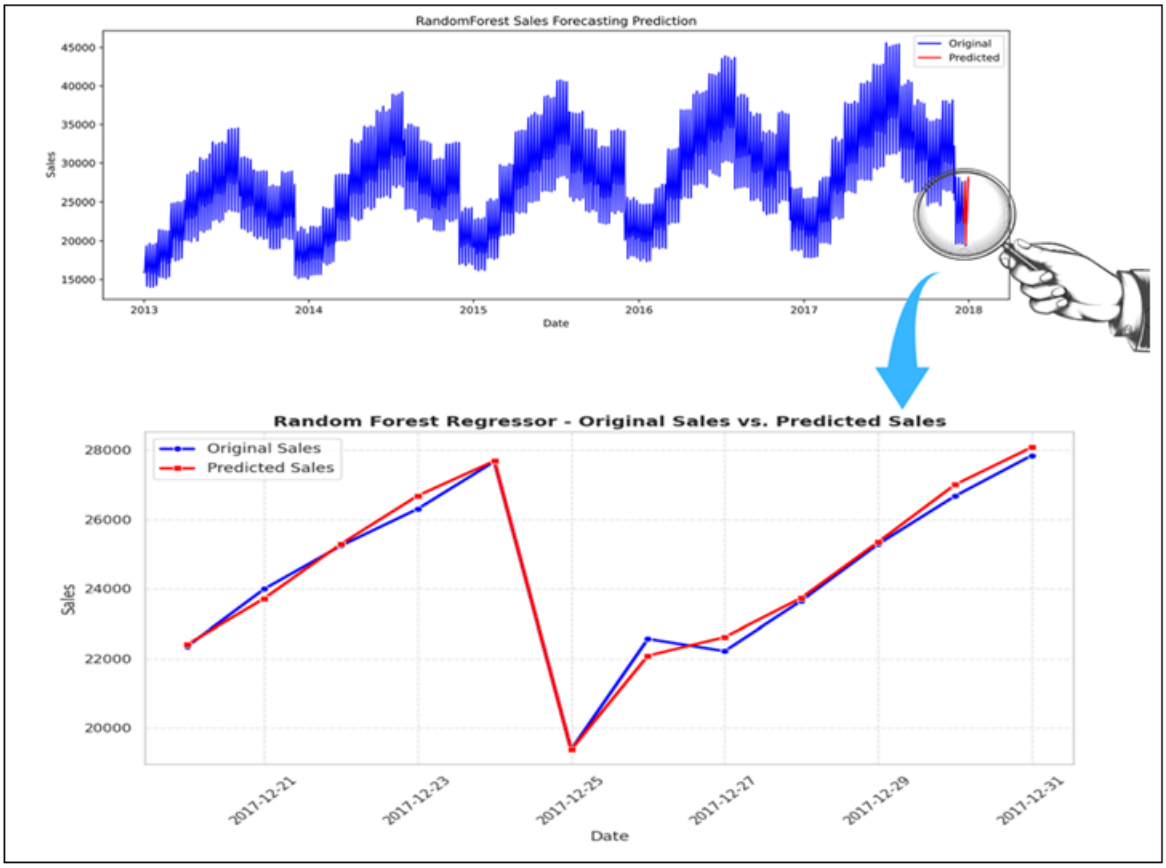

Figure 5: Random Forest Regressor on sales data - Original vs Prediction

Random Forest Regressor Model Score	
Shape of Train	(1801, 13)
Shape of Test	(12, 13)
RMSE (Root Mean square error)	270.67231110
MAE (Mean absolute error)	215.16666667
R2 Score	0.9876588684

XGBoost Gradient:

The Linear Regression and Random Forest are easy to interpret but often with accuracy on complex data sets. The **XGBoost**, short for eXteme Gradient Boosting model employs an optimized gradient boosting algorithm to capture complex patterns and dependencies in the data. By iteratively improving weak learners (decision trees), it enhances predictive accuracy while controlling overfitting through regularization techniques.

Mathematically, the training set can be represented as:

$$\boldsymbol{X_t = f(X_{t-1}, X_{t-2}, \ldots, X_{t-12})}$$

where $\boldsymbol{X_t}$ represents the sales data at time t, and the lagger feature $\boldsymbol{X_{t-1}, X_{t-2}, \ldots, X_{t-12}}$ captures past trends over the previous 12 months.

The **XGBoost Regressor** model optimizes the prediction process by sequentially minimizing errors using a gradient boosting framework. The model determines the best regression fit by combining multiple decision trees in an additive manner and refining the predictions at each step.

Instead of fitting the model all at once XGBoost optimizes the model iteratively. The model begins with an initial prediction $\hat{y}_t^{(0)} = 0$, refining it iteratively by adding new trees to improve the model. The updated predictions after adding the n^{th} tree can be written as:

$$\boldsymbol{\hat{y}_t^{(i)} = \hat{y}_t^{(i-1)} + f(X_{t-1}, X_{t-2}, \ldots, X_{t-12})}$$

Where $\hat{y}_t^{(i-1)}$ is the prediction from the previous iteration.

The **regularization term** $\Omega(f(X_{t-1}, X_{t-2}, \ldots, X_{t-12})$ simplify complex trees by penalizing the number of leaves in the tree and the size of the leaf. It is defined as:

$$\boldsymbol{\Omega(f(X_{t-1}, X_{t-2}, \ldots, X_{t-12}) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2}$$

T is the number of leaves in the tree.The depth of tree impacts model complexity.

γ: Regularization that controls complexity and prevents overfitting.

λ is a parameter that penalizes the squared weight and prevents complexity of the model.

The **XGBoost model** has been trained on the same dataset referenced above. The trained model has been saved as **"XGBoost_trained_model.pkl"** and is included in the data and code repository mentioned in the Supplementary Materials section.

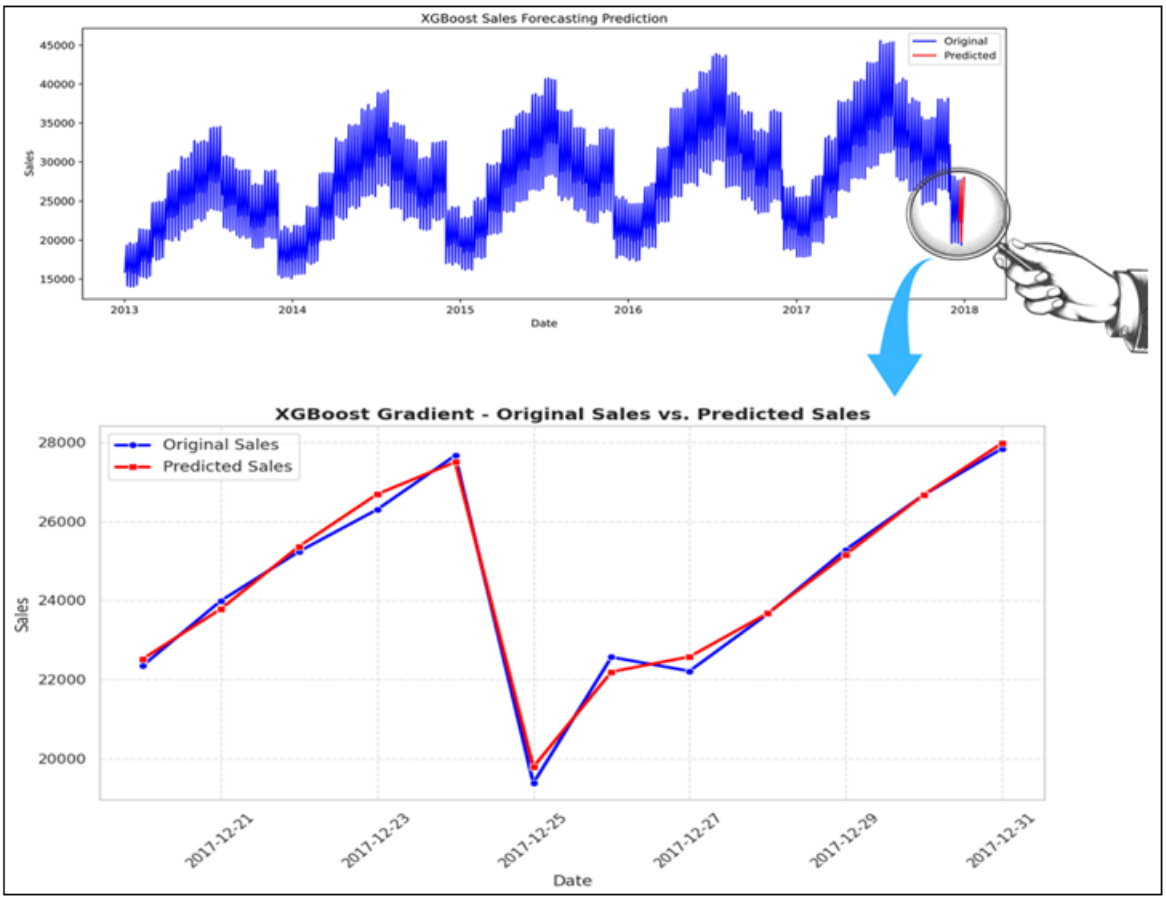

Figure 6: XGBoost on sales data - Original vs Prediction

XGBoost Gradient	Model Score
Shape of Train	(1801, 13)
Shape of Test	(12, 13)
RMSE (Root Mean square error)	247.416652633
MAE (Mean absolute error)	206.0
R2 Score	0.98968842100

Support Vector Regression:
The Linear Support Vector Machine (SVM) Regression model identifies the best-fitting hyperplane that minimizes error while maintaining a margin of tolerance.

Mathematically, the training set can be represented as:

$$X_t = f(X_{t-1}, X_{t-2}, \ldots, X_{t-12})$$

where X_t represents the sales data at time t, and the lagger feature $X_{t-1}, X_{t-2}, \ldots, X_{t-12}$ captures past trends over the previous 12 months.

The **Support Vector Regressor model** has been trained on the same dataset referenced above. The trained model has been saved as **" SVM_regression_trained_model.pkl"** and is included in the data and code repository mentioned in the Supplementary Materials section.

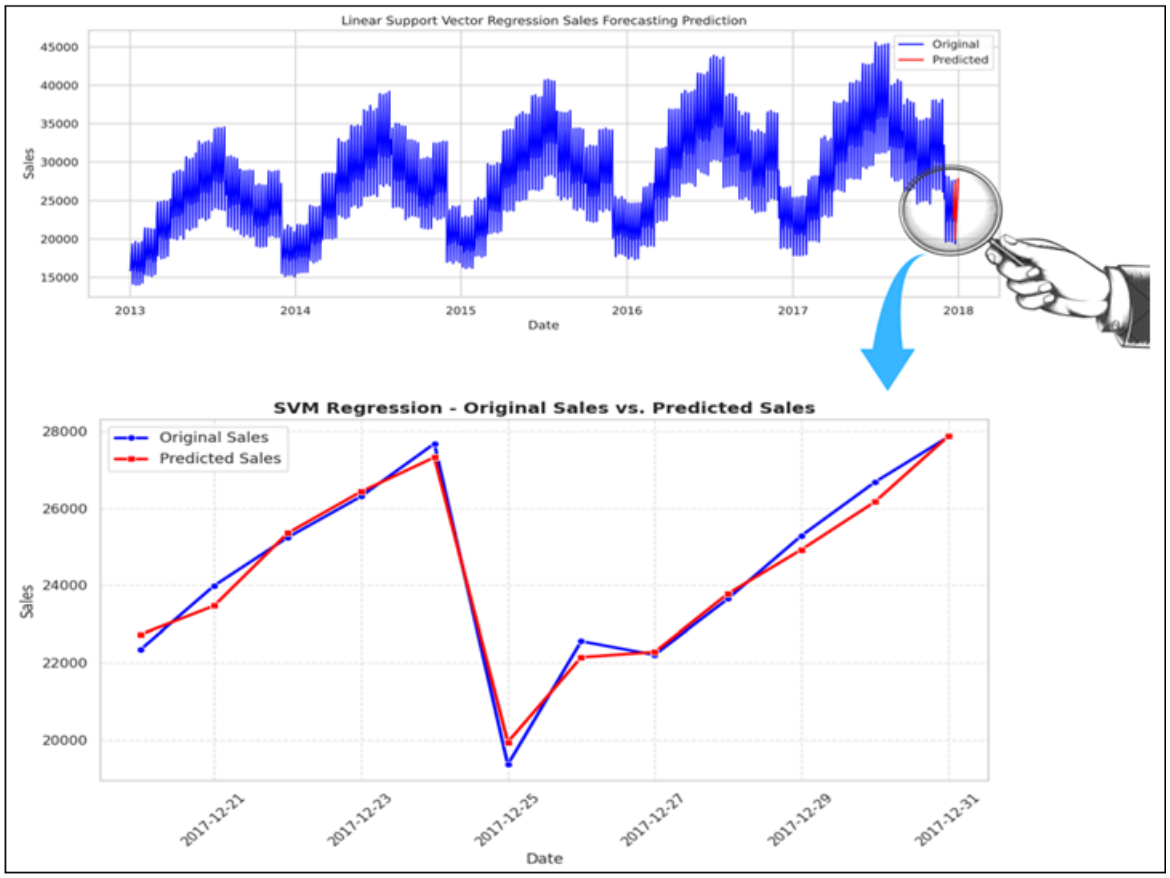

Figure 7: SVM Regression on sales data - Original vs Prediction

Support Vector Regressor	Model Score
Shape of Train	(1801, 13)
Shape of Test	(12, 13)
RMSE (Root Mean square error)	353.28281872
MAE (Mean absolute error)	298.75
R2 Score	0.9789761449

LSTM (Long Short-Term Memory):
Long Short-Term Memory is an improved version of the recurrent neural network and excels in time series prediction task. Unlike traditional regression-based models, the LSTM (Long Short-Term Memory) model is a type of recurrent neural network (RNN) specifically designed to capture long-term dependencies in sequential data. LSTMs are highly effective in time series forecasting because they maintain memory of past observations and learn complex temporal relationships without suffering from vanishing gradient issues, which commonly affect standard RNNs.

Mathematically, the training set can be represented as:

$$\boldsymbol{X_t = f(X_{t-1}, X_{t-2}, \ldots, X_{t-12})}$$

where $\boldsymbol{X_t}$ represents the sales data at time t, and the lagger feature $\boldsymbol{X_{t-1}, X_{t-2}, \ldots, X_{t-12}}$ captures past trends over the previous 12 months.

The **LSTM model** has been trained on the same dataset referenced above. The trained model has been saved as **" lstm _trained_model.pkl"** and is included in the data and code repository mentioned in the Supplementary Materials section.

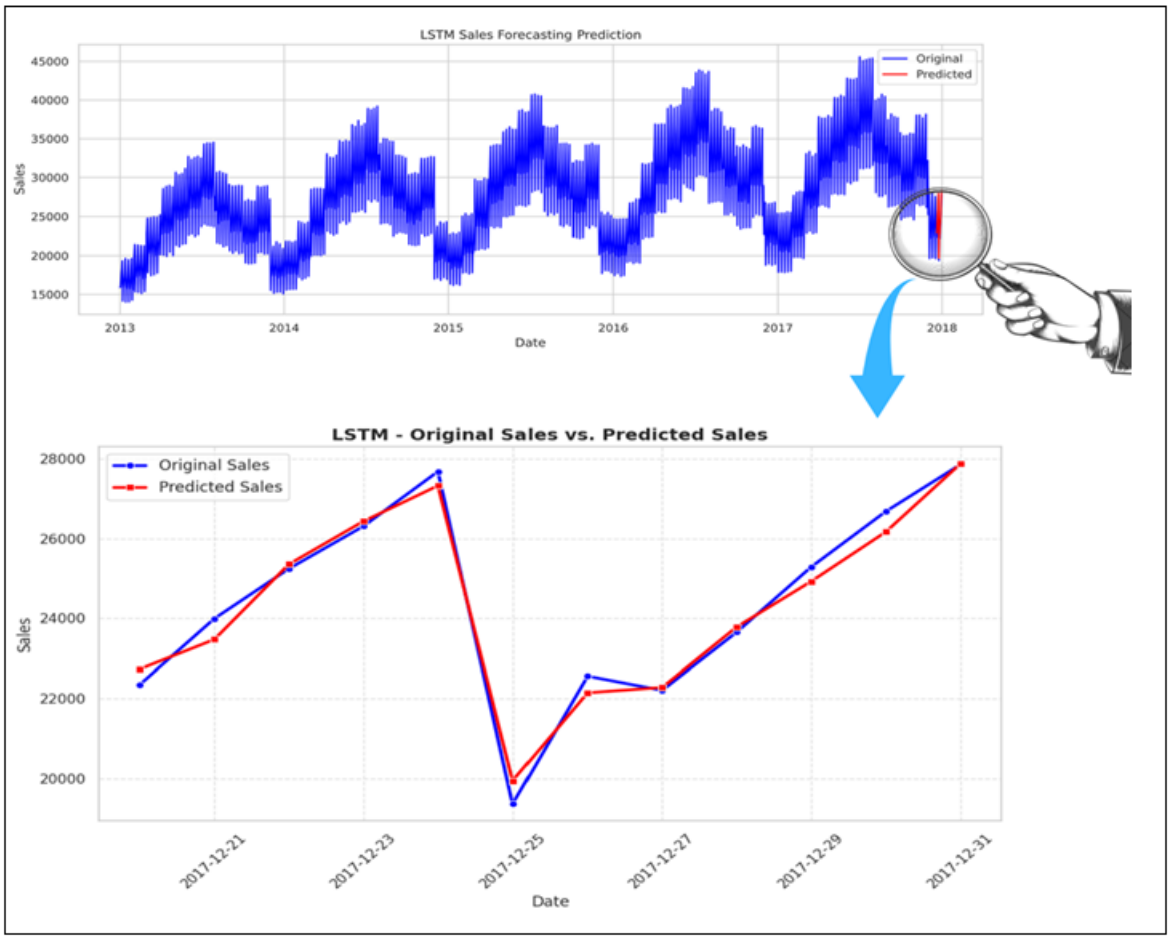

Figure 8: LSTM Regression on sales data - Original vs Prediction

LSTM	Model Score
Shape of Train	(1801, 13)
Shape of Test	(12, 13)
RMSE (Root Mean square error)	440.71778574
MAE (Mean absolute error)	399.0
R2 Score	0.9672818699

ARIMA (Auto Regressive Integrated Moving Average) Model:

ARIMA modeling effectively captures underlying trends, autocorrelation, and seasonality while offering flexible modeling for various types of impacts (Schaffer et al., 2021). Unlike deep learning-based models, the ARIMA (Auto Regressive Integrated Moving Average) model is a classical statistical approach specifically designed for time series forecasting. ARIMA effectively captures trends, seasonality, and autocorrelations in sales data through its three main components:

- Auto Regressive (AR) term: Captures relationships between past observations.
- Integrated (I) term: Differencing step to make the series stationary.
- Moving Average (MA) term: Accounts for past forecast errors to improve predictions

Mathematically, The ARIMA model parameters can be represented as:

$$X_t = f(X_{t-1}, X_{t-2}, \ldots, X_{t-12})$$

where X_t represents the sales data at time t, and the lagger feature $X_{t-1}, X_{t-2}, \ldots, X_{t-12}$ captures past trends over the previous 12 months.

The ARIMA model leverages auto-correlation to identify relationships between past and present values in a time series, while partial auto-correlation helps determine the direct influence of past observations by filtering out intermediate lags, enabling optimal selection of AR and MA terms for accurate forecasting.

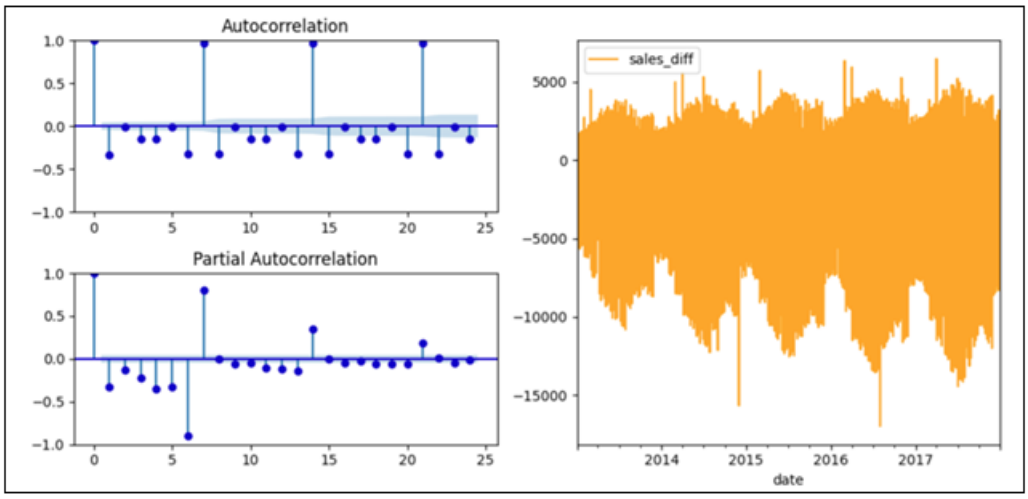

Figure 9: Time Series Auto Correlation and Partial Correlation

The **ARIMA model** has been trained on the same dataset referenced above. The trained model has been saved as **" arima_trained_model.pkl"** and is included in the data and code repository mentioned in the Supplementary Materials section.

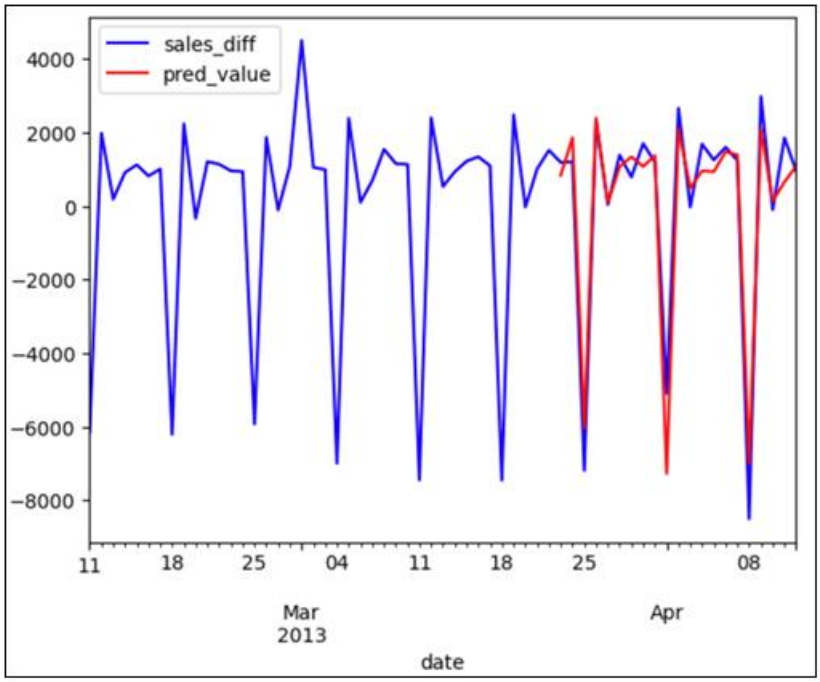

Figure 10: ARIMA Model on sales data - Original Sales vs Prediction

ARIMA Model Score	
Shape of Train	(1801, 13)
Shape of Test	(12, 13)
RMSE (Root Mean square error)	937.95186466
MAE (Mean absolute error)	713.08111124
R2 Score	0.9169801656

Training and Validation Process

The sales forecasting model was trained and validated using a structured approach to ensure reliable and accurate predictions. The dataset was split into training and test sets, with the training data consisting of 1,801 records and 13 features (Shape: (1801, 13)), while the test set contained 12 records with the same number of features (Shape: (12, 13)). The training data was used to fit the selected machine learning models, allowing them to learn patterns and relationships within historical sales data. The test set served as an independent dataset for validation, assessing the model's ability to generalize to unseen data. Performance metrics such as Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) were used to evaluate the effectiveness of each model in predicting future sales. This rigorous training and validation process ensured that the model could effectively capture trends and fluctuations in sales, leading to improved forecasting accuracy.

Comparison of Different Models

The performances were assessed using root mean square error and mean absolute error (Zhou et al., 2023). We compared Linear Regression, Random Forest Regressor, XGBoost using, Support Vector Linear Regressor, LSTM, and ARIMA. The R^2 score, or coefficient of determination having values closer to 1 indicate better predictive performance. Linear Regression provided a baseline performance, capturing only linear relationships in the data. We can see the XGBoost Gradient has less error when compared to other models.

Model	RMSE (Root Mean Square Error)	MAE (Mean Absolute Error)
Linear Regression	308.9278503	285.4166667
Random Forest Regressor	270.6723111	215.1666667
XGBoost Gradient	247.4166526	206
Support Vector Regressor	353.2828187	298.75
LSTM	440.7177857	399
ARIMA	937.9518647	713.0811112

Performance Metrics

To evaluate the effectiveness of machine learning models' various performance metrics were utilized. Below graph represent the comparison using R^2 score

Model	R^2 Score
Linear Regression	0.983923872
Random Forest Regressor	0.987658868
XG Boost Gradient	0.989688421
Support Vector Regressor	0.978976145
LSTM	0.96728187
ARIMA	0.916980166

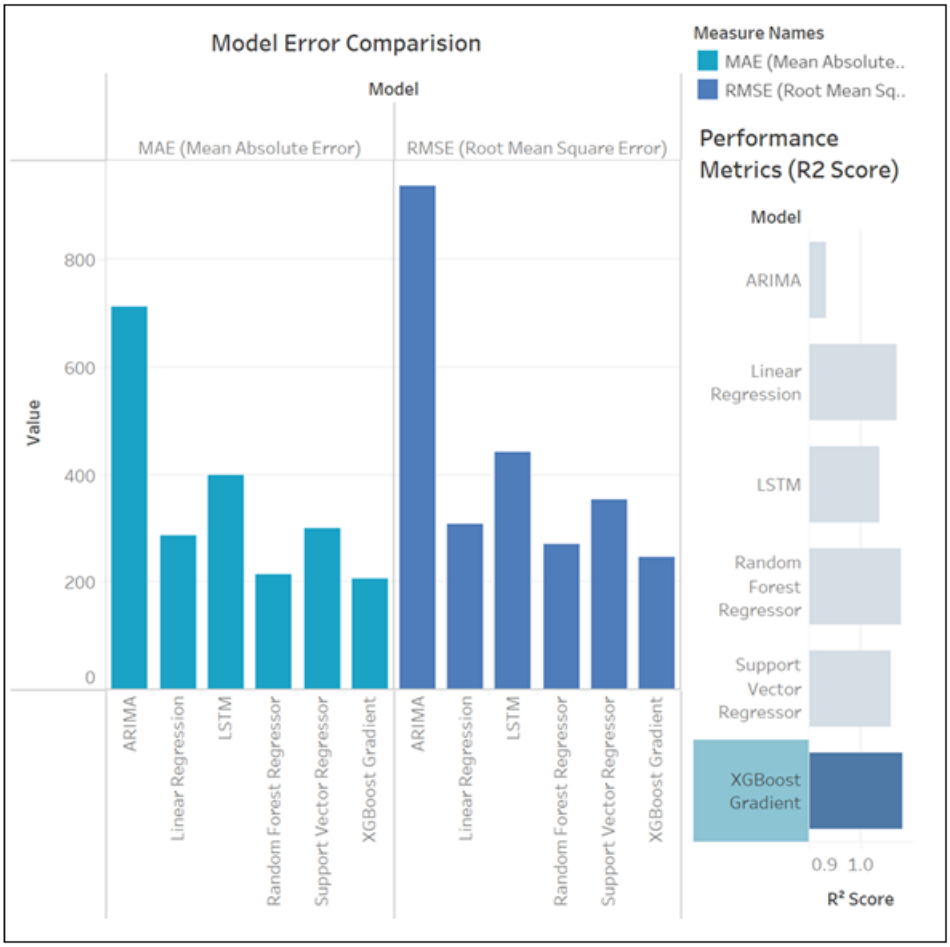

Figure 11: Comparison of Error and Performance between models

Sales Forecasting Results and Practical Implications

Short-term Sales Forecasting

In this study, short-term forecasting focuses on predicting sales for the upcoming weeks or months based on historical data patterns, seasonal trends, and external factors. The short-term sales forecasts are well suited for retailers in highly competitive business models and for them the data quality becomes pertinent to train the models and make accurate predictions. There can be approach utilized to derive reliable insights with limited information, one of such method is discussed in the article (Chee et al., 2022) By leveraging short-term sales forecasting, businesses can make data-driven decisions to optimize stock levels, reduce operational costs, and meet customer demand more effectively.

Long-term Sales Forecasting

Long-term sales forecasting is essential for strategic business planning, budgeting, and capacity management, as it provides insights into future sales trends over extended periods. Unlike short-term forecasting, which captures immediate fluctuations, long-term forecasting focuses on identifying broader patterns, seasonal cycles, and market trends that influence sales over time. In this study, machine learning models were trained using historical sales data, aggregated features, and lagged variables to capture long-term dependencies.

Seasonal and Trend Analysis

Seasonal and trend analysis was essential for understanding patterns and making accurate predictions. Machine learning algorithms provided valuable insights into data-driven decision-making, especially when considering long forecast horizons (Maccarrone et al., 2021). In contrast, there are

situations where seasonality did not affect the series over time (Da'ar & Ahmed, 2018). Our research shows that seasonality affects retail sales and sales forecasts too. Algorithms may perform differently on distinct data sets, such as retail sales versus infectious disease data are two different use cases. Due to their varying natures, result may vary.

Integration with Retail Systems
Integrating machine learning models with retail systems involved setting up data pipelines, deploying models, and monitoring performance to ensure accuracy and efficiency. Seamless integration allowed for real-time data processing and decision-making support. The trained model .pkl file along with train and test data sets has been provided in the Model Development section.

Scalability and Real-time Performance
Scalability and real-time performance were critical for practical application. The models effectively handled large volumes of data, providing timely and accurate predictions to support business decision-making. The model can be enhanced further with the addition of new features and new variables and all the models can be trained and tested to find which performs better.

Business Impact and ROI Analysis
This study can help retailers bring accuracy to sales forecasts. The accurate forecasts will trigger the replenishment of the required quantity at stores or online ordering. The availability of the right product at the right time directly relates to customer retention, satisfaction as well as the growth of sales revenue for retailers.

4. Discussion

The increasing complexity and dynamic nature of the retail industry underscores the significance of our study to forecast sales using machine learning models. Traditional sales forecasting methods, often based on simple models, fail to account for intricate interactions such as price, consumer income, and seasonal trends. In contrast, machine learning techniques provide strong capabilities for capturing complex nonlinear patterns and learning from temporal sequences.

Our study introduces an innovative integration of various machine learning models to enhance the accuracy of retail sales forecasting and has shown significant improvements in forecasting accuracy. This approach addresses the limitations of traditional models, offering a more robust framework for managing inventory levels and mitigating understocking or overstocking risks. Machine learning techniques uncover meaningful patterns in vast and diverse data, a task that would be nearly impossible for traditional predictive models or even highly skilled individuals (Mitra et al., 2022).

In our study, we utilized machine learning methods such as Linear Regression, Random Forest Regressor, XGBoost Gradient, Support Vector Regression, LSTM, ARIMA. The Gradient boosting exhibited superior performance in predicting, achieving an R^2 score: 0.9892.

These advanced algorithms provide valuable guidance for data-driven decision-making, especially when considering long forecast horizons (Maccarrone et al., 2021). Our findings indicate that machine learning models produce more accurate forecasts compared to traditional models. This aligns with existing literature highlighting the superiority of machine learning models in handling complex and dynamic datasets.

However, discrepancies exist between our findings and current knowledge. While we observed exceptional performance with gradient boosting, other studies have reported mixed results regarding the superiority of different machine learning models. These variations could be attributed to differences in datasets, feature selection, and model tuning parameters. Further research is needed to standardize these factors and provide a more comprehensive comparison of machine learning models in retail sales forecasting.

Our study contributes to the existing body of knowledge by demonstrating the practical applications and benefits of machine learning in retail sales forecasting. The integration of support vector regression, for example, outperformed comparison models in terms of forecasting error, indicating the promise of hybrid models in sales forecasting (Dai et al., 2014). This finding is consistent with previous research emphasizing the importance of hybrid models in improving forecasting accuracy (Lu & Chang, 2014).

5. Conclusion

This study validates the effectiveness of advanced machine learning algorithms in improving retail sales forecasting. Among all models, XGBoost demonstrated the most accurate predictions, emphasizing its utility in operational planning. By addressing data preprocessing and model selection challenges, this research contributes a replicable forecasting framework. Future studies should integrate real-time data streams and broader external variables for dynamic market responsiveness.

Declaration of Competing Interest
The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

This article represents personal views and research findings conducted independently by correspondence and co-author and does not reflect the opinions, policies, or positions of any current or former organizations with which I have been affiliated.

Data availability
The sample model data along with the trained model has been provided in the Model Development section. Additional data will be made available on request.

Supplementary Materials
1) Linear Regression
2) Random Forest Regressor
3) XGBoost Gradient
4) Support Vector Regression
5) LSTM
6) ARIMA

7) Code Availability: https://github.com/MayankDw/salesforecast_using_LinearRegression.git

References

[1] Abdullhadi, S., Al-Qudah, D. A., & Abu-Salih, B. (2024). Time-aware forecasting of search volume categories and actual purchase. *Heliyon*, *10*(3), e25034. https://doi.org/10.1016/j.heliyon.2024.e25034

[2] Bojer, C. S. (2022). Understanding machine learning-based forecasting methods: A decomposition framework and research opportunities. *International Journal of Forecasting*, *38*(4), 1555–1561. https://doi.org/10.1016/j.ijforecast.2021.11.003

[3] Chee, C. C. F., Leng Chiew, K., Sarbini, I. N. B., & Jing, E. K. H. (2022). Data Analytics Approach for Short-term Sales Forecasts Using Limited Information in E-commerce Marketplace. *Acta Informatica Pragensia*, *11*(3), 309–323. https://doi.org/10.18267/j.aip.196

[4] Da'ar, O. B., & Ahmed, A. E. (2018). Underlying trend, seasonality, prediction, forecasting and the contribution of risk factors: An analysis of globally reported cases of Middle East Respiratory Syndrome Coronavirus. *Epidemiology and Infection*, *146*(11), 1343–1349. https://doi.org/10.1017/S0950268818001541

[5] Dai, W., Wu, J.-Y., & Lu, C.-J. (2014). Applying different independent component analysis algorithms and support vector regression for IT chain store sales forecasting. *TheScientificWorldJournal*, *2014*, 438132. https://doi.org/10.1155/2014/438132

[6] Dolan, E., Goulding, J., Marshall, H., Smith, G., Long, G., & Tata, L. J. (2023). Assessing the value of integrating national longitudinal shopping data into respiratory disease forecasting models. *Nature Communications*, *14*(1), 7258. https://doi.org/10.1038/s41467-023-42776-4

[7] Elalem, Y. K., Maier, S., & Seifert, R. W. (2023). A machine learning-based framework for forecasting sales of new products with short life cycles using deep neural networks. *International Journal of Forecasting*, *39*(4), 1874–1894. https://doi.org/10.1016/j.ijforecast.2022.09.005

[8] Li, C., Jiang, W., Yang, Y., Pan, S., Huang, G., & Guo, L. (2023). Predicting Best-Selling New Products in a Major Promotion Campaign Through Graph Convolutional Networks. *IEEE Transactions on Neural Networks and Learning Systems*, *34*(11), 9102–9115. https://doi.org/10.1109/TNNLS.2022.3155690

[9] Lu, C.-J., & Chang, C.-C. (2014). A hybrid sales forecasting scheme by combining independent component analysis with K-means clustering and support vector regression. *TheScientificWorldJournal*, *2014*, 624017. https://doi.org/10.1155/2014/624017

[10] Maccarrone, G., Morelli, G., & Spadaccini, S. (2021). GDP Forecasting: Machine Learning, Linear or Autoregression? *Frontiers in Artificial Intelligence*, *4*, 757864. https://doi.org/10.3389/frai.2021.757864

[11] Mitra, A., Jain, A., Kishore, A., & Kumar, P. (2022). A Comparative Study of Demand Forecasting Models for a Multi-Channel Retail Company: A Novel Hybrid Machine Learning Approach. *Operations Research Forum*, *3*(4), 58. https://doi.org/10.1007/s43069-022-00166-4

[12] Oliveira, J. M., & Ramos, P. (2019). Assessing the Performance of Hierarchical Forecasting Methods on the Retail Sector. *Entropy (Basel, Switzerland)*, *21*(4), 436. https://doi.org/10.3390/e21040436

[13] Parpoula, C. (2024). An analytical approach for identifying trend-seasonal components and detecting unexpected behaviour in psychological time-series. *International Journal of Psychology*, *59*(6), 1307–1316. https://doi.org/10.1002/ijop.13244

[14] Schaffer, A. L., Dobbins, T. A., & Pearson, S.-A. (2021). Interrupted time series analysis using autoregressive integrated moving average (ARIMA) models: A guide for evaluating large-scale health interventions. *BMC Medical Research Methodology*, *21*(1), 58. https://doi.org/10.1186/s12874-021-01235-8

[15] Schneider, M. J., & Gupta, S. (2016). Forecasting sales of new and existing products using consumer reviews: A random projections approach. *International Journal of Forecasting*, *32*(2), 243–256. https://doi.org/10.1016/j.ijforecast.2015.08.005

[16] Tawakuli, A., Havers, B., Gulisano, V., Kaiser, D., & Engel, T. (2024). Survey:Time-series data preprocessing: A survey and an empirical analysis. *Journal of Engineering Research*, S2307187724000452. https://doi.org/10.1016/j.jer.2024.02.018

[17] Zhou, Y., Ahmad, Z., Almaspoor, Z., Khan, F., Tag-Eldin, E., Iqbal, Z., & El-Morshedy, M. (2023). On the implementation of a new version of the Weibull distribution and machine learning approach to model the COVID-19 data. *Mathematical Biosciences and Engineering: MBE*, *20*(1), 337–364. https://doi.org/10.3934/mbe.2023016