International Journal of HRM and Organizational Behavior

www.ijhrmob.com

editor@ijhrmob.com

International Journal of
HRM and Organizational Behavior

ISSN 2454 - 5015

Forecasting E-Commerce Trends: Utilizing Linear Regression, Polynomial Regression, Random Forest, and Gradient Boosting for Accurate Sales and Demand Prediction

Naresh Kumar Reddy Panga,

Abstract:

The rapid expansion of e-commerce has made the application of advanced methods for accurate demand and sales forecasting necessary. To predict e-commerce trends, this study assesses the performance of four machine learning models: Gradient Boosting, Random Forest, Polynomial Regression, and Linear Regression. This project intends to optimize inventory management, improve marketing strategies, and increase the accuracy of demand forecasts by examining historical sales data and applying cutting-edge machine learning algorithms. To attain greater forecast accuracy, the study also looks into the development of a hybrid model that incorporates the advantages of several approaches.

Keywords: Sales Forecasting, Linear Regression, Polynomial Regression, Random Forest, Gradient Boosting, Feature Engineering

1. Introduction:

The retail industry has changed dramatically as a result of e-commerce, which has given companies new strategies to connect with customers around the world. Sophisticated techniques are required for precise sales and demand forecasting due to the explosive rise of e-commerce. Businesses can improve marketing strategies, gain insights into future trends, and optimize inventory management by utilizing advanced machine learning techniques including Random Forest, Polynomial Regression, Linear Regression, and Gradient Boosting.

Engineering Manager,
Virtusa Corporation, New York, USA.
Email ID: nareshpangash@gmail.com

Significant technical developments have affected the evolution of e-commerce and changed how organizations operate and interact with customers. The accuracy of demand and sales projections could be improved with the availability of more data and machine learning developments. For e-commerce companies to maintain ideal inventory levels, save expenses, and increase customer satisfaction, accurate forecasting is essential.

By making it possible to identify patterns and predictions from enormous datasets, machine learning algorithms have completely changed the field of data analysis. Traditional statistical techniques for forecasting by fitting a line or curve to past data including polynomial regression and linear regression. An ensemble learning technique called Random Forest combines multiple decision trees to increase prediction accuracy. Another effective ensemble strategy is gradient boosting, which creates models one after the other to adjust for prior model flaws and produce incredibly accurate forecasts. These methods can be used to better forecast demand and sales using e-commerce data.

Many e-commerce companies still experience problems with inaccurate demand and sales projections, even in the face of advanced forecasting techniques. A negative impact on profitability and customer satisfaction might result from situations like stockouts or overstocking. To manage the special qualities of e-commerce data, such as seasonality and rapid market shifts, the difficulty is in identifying the best machine learning models and fine-tuning them.

Although a lot of research has been done on the use of specific machine learning algorithms for sales forecasting, there aren't many in-depth studies that compare the efficacy of various approaches in the context of e-commerce. Furthermore, there is still much to learn about integrating different algorithms to make use of their combined strengths. By comparing the effectiveness of linear regression, polynomial regression, random forest, and gradient boosting in predicting e-commerce trends, this study hopes to fill this gap.

- To evaluate how well Demand and Sales for E-Commerce are Forecasted Using Polynomial, Random Forest, Gradient Boosting, and Linear Regression.
- To assess these methods' accuracy and determine the optimal circumstances for each.
- To create a hybrid model that enhanced predicting accuracy by combining the best features of each separate method.
- Depending on the forecasting findings, offer e-commerce companies practical insights to improve marketing and inventory management.

This research will enhance the field of e-commerce analytics by offering a thorough assessment of advanced machine-learning techniques for sales forecasting. By doing so, it will assist companies in making wise decisions and maintaining their competitiveness in the ever-changing online market.

2. Literature Survey:

Firas Al-Basha investigates the use of machine learning algorithms and Google Trends data to estimate retail sales in his doctoral research at HEC Montréal, which was completed in 2021. Al-Basha shows how these techniques can greatly improve the accuracy of retail sales forecasts by combining real-time search data with advanced predictive algorithms. The research emphasizes how digital search patterns and conventional sales data may be used to enhance forecasting, providing useful information for companies looking to maximize their marketing and inventory plans.

Ajay et al. (2023) explore sales forecasting with a modified Random Forest and Decision Tree method. Studies reveal that as compared to conventional approaches; these altered methodologies can dramatically increase the precision of sales forecasts. The research offers significant insights to improve forecasting tactics and, eventually, inventory management and decision-making processes for businesses through the analysis of intricate data patterns.

In 2022, TU et al. investigate data-driven daily product sales forecasting within an ecosystem of third-party e-platforms. The researchers create prediction models that greatly improve forecast accuracy by utilizing large amounts of sales data. The study emphasizes the value of real-time analysis and data integration in enhancing sales forecasts, providing e-commerce platforms with insightful information to maximize inventory control and marketing tactics. Businesses can respond to market demand more skillfully and with greater knowledge when they use this strategy.

Ali Khakpour looks at big data and machine learning applications for production and retail sales forecasting in his master's thesis from 2020. The research shows how sophisticated data science methods can improve decision support systems and produce more precise sales forecasts. To increase forecasting accuracy and eventually assist companies in optimizing their production and inventory management methods, Khakpour emphasizes the importance of combining big datasets and machine learning algorithms. A strong basis for utilizing data science in real-world decision-making situations is offered by this research.

The task of predicting demand for new items from data from comparable existing products is taken up by Steenbergen (2019). To improve these forecasts' accuracy, it uses sophisticated machine learning algorithms, including Random Forest and Quantile Regression Forest. Using these advanced techniques, also demonstrates how companies may introduce new items with greater knowledge and manage inventories more effectively. To improve inventory control and launch new products, the study emphasizes how well these algorithms identify complicated demand patterns and uncertainties.

Chen (2022) explores how e-commerce database marketing can be transformed by machine learning algorithms. These cutting-edge methods enable companies to segment their market efficiently, anticipate consumer behaviour based on customer data, and customize marketing campaigns based on personal preferences. Chen's study demonstrates the substantial increases in consumer engagement, sales, and marketing campaign efficiency that can result from implementing these strategies. To develop more focused and effective marketing strategies, the study provides a detailed look at the way computational intelligence might be incorporated into online shopping.

Shaohui and Kudryavtsev (2021) explore how various fusion models might forecast whether or not customers will make repeat purchases. They demonstrate how organizations can more precisely predict repurchase behaviour by merging multiple predictive models. By using this method, different data patterns can be captured, leading to a greater understanding of client loyalty. Consequently, this helps businesses to estimate repurchases more accurately, improve client retention, and refine their marketing strategies. The need to apply cutting-edge prediction algorithms to better understand and increase client loyalty is shown.

Kumar et al. (2023) explore the application of advanced machine learning in e-commerce to assess client buying trends, which they presented at an international conference. They show how these advanced algorithms can provide significant behavioural insights into the market, enabling companies to improve consumer experiences and develop more successful marketing campaigns. Machine learning helps e-commerce businesses predict client preferences more accurately, manage inventory more effectively, and increase sales. All of these benefits help these businesses succeed in the cutthroat online marketplace.

Moroke and Makatjane (2022) focus on the use of gradient-boosting decision trees in the identification of financial fraud. They show that by examining a large amount of data, this sophisticated machine-learning technique can precisely detect fraudulent activity. According to the study, gradient-boosting decision trees function more accurately and efficiently than conventional techniques, which makes them an effective tool for enhancing financial stability. It also highlights the possibility of greatly improving fraud detection and prevention efforts in the financial sector by incorporating these machine-learning approaches.

3. Methodology

Using a variety of machine learning approaches, including Random Forest, Polynomial Regression, Gradient Boosting, and Linear Regression, this study seeks to forecast e-commerce trends. The technique has been designed to assess and contrast how these algorithms forecast demand and sales. The specific procedures and methods used in this study are described in depth in the subtopics that follow.

Collecting pertinent e-commerce data is the first stage. This includes past sales data, product details, metrics related to consumer behaviour, and any outside variables that may affect sales (e.g., holidays, promotions, and market trends). The information will come from business documents, publicly accessible datasets, and real-time search data APIs like Google Trends.

3.1 Data Preprocessing

To guarantee the analysis's accuracy and dependability, data preliminary processing is necessary. As a way to avoid skewing the results, the data must first be cleaned by eliminating any missing or duplicate information. When attempting to prevent outliers from negatively affecting predictions, it is imperative to handle them. This can be done by either modifying extreme numbers to better fit with the rest of the data or by employing other strategies. Cross-referencing data with other sources or employing subject expertise to validate information are common methods for fixing data errors.

Normalization is carried out after the data has been cleaned up to guarantee that each feature contributes evenly to the model. This usually entails standardizing with z-scores or applying min-max scaling which brings everything into a common range, which aids in the model's more efficient learning as illustrated in Figure 1.

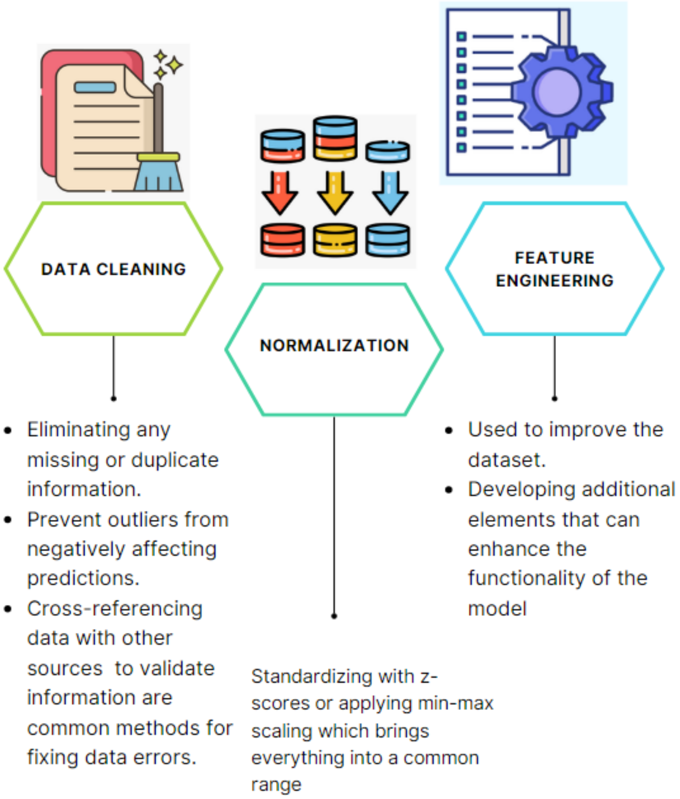
DATA CLEANING
NORMALIZATION
FEATURE ENGINEERING
Eliminating any missing or duplicate information.
Prevent outliers from negatively affecting predictions.
Cross-referencing data with other sources to validate information are common methods for fixing data errors.
Standardizing with z-scores or applying min-max scaling which brings everything into a common range
Used to improve the dataset.
Developing additional elements that can enhance the functionality of the model

Fig. 1: Steps of Data Preprocessing in E-commerce

Ultimately, feature engineering is used to improve the dataset. This entails developing additional elements that can enhance the functionality of the model, like figuring out sales growth rates to spot trends, providing indications for seasonal patterns, and accounting for the impact of promotions. By adding these more qualities, the model can comprehend the data more fully and produce more accurate predictions.

3.2 Exploratory Data Analysis:

Understanding the patterns and relationships in the dataset requires conducting exploratory data analysis, or EDA. First, important features of the data are combined using descriptive statistics. The standard deviation and variance are used to determine the distribution of the data, and metrics like the mean, median, and mode are used to comprehend the centre values. When trying to comprehend the form of the data distribution and spot any odd trends, it is also helpful to examine skewness and kurtosis.

Visualization tools are used to get deeper into the data after it has been summarized. Plots and graphs of all kinds, including scatter plots, box plots, and histograms, can be used to identify patterns and trends. Seasonal trends can be easily identified with time series plots. Heatmaps and correlation matrices make it simpler to identify dependencies or correlations by displaying the relationships between several variables. These visual aids are very helpful in locating anomalies, outliers, and clusters within the data, which helps to clarify the structure of the data. Raw data is converted into insightful understandings by EDA, which is essential for developing and improving predictive models has been clarified in Fig. 2.

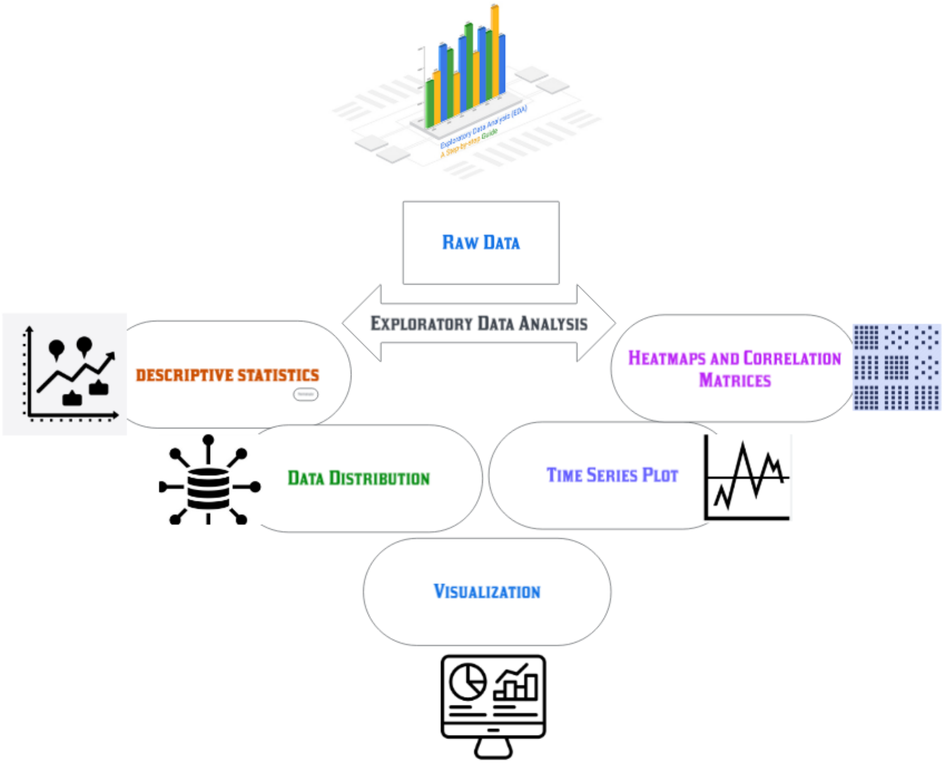

Fig.2: Exploratory Data Analysis (EDA)

3.3 Model Selection

Linear Regression: A simple statistical technique that fits a linear equation to represent the relationship between a dependent variable and one or more independent variables.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \epsilon \quad (1)$$

In equation 1, Where Y is the dependent variable (sales), β_0 is the y-intercept, $\beta_1, \beta_2, \dots, \beta_n$ are the coefficients, $X_1,\ X_2, \dots,\ X_n$ are the independent variables and ϵ is the error term.

Polynomial Regression: An expansion of linear regression that creates a more flexible model capable of capturing nonlinear interactions by fitting a polynomial equation to the data.

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \cdots + \beta_n X^n + \epsilon \quad (2)$$

In equation 2, Y is the dependent variable, β_0 is the intercept, $\beta_1, \beta_2, \dots, \beta_n$ are the coefficients for each polynomial term, X is the independent variable, ϵ is the error term.

Random Forest: An approach to ensemble learning that builds several decision trees during training and outputs the average prediction of each tree to reduce overfitting and increase accuracy. In contrast, the model employs the mean of several decision trees' predictions.

Gradient Boosting: Another ensemble technique that creates models in sequence, with each new model repairing the mistakes of the older models to provide incredibly accurate forecasts.

$$F_m(x) = F_{m-1}(x) + h_m(x) \quad (3)$$

In equation 3, Where $F_m(x)$ is the current model, $F_{m-1}(x)$ is the previous model and $h_m(x)$ is the new model fit to the residuals of the previous model

3.4 Training and Validation

There are several important phases involved in training and verifying the models to make sure they perform well and can handle fresh data. First, a training set and a testing set are created from the dataset. With the help of this section, models may be trained and their performance assessed on different sets of data. The training dataset is used to teach each model the patterns and relationships found in the data during its initial training. To minimize prediction errors, the model's parameters are changed during this process, also referred to as fitting.

Cross-validation techniques such as k-fold cross-validation are employed to guarantee that the models are robust and not only customized for the training data. This entails dividing the training data into numerous smaller sets and training the model several times, using the validation set to be a different set each time and the remaining sets as training sets. Through this process, the model's parameters can be adjusted to optimize its performance across various data subsets. Lastly, using the testing dataset that was not utilized for training, the trained models are evaluated. This phase has great importance as it provides a precise gauge of the models' ability to forecast novel, unobserved data. For practical applications, the models are guaranteed to be accurate and dependable by passing through the training, cross-validation, and testing phases.

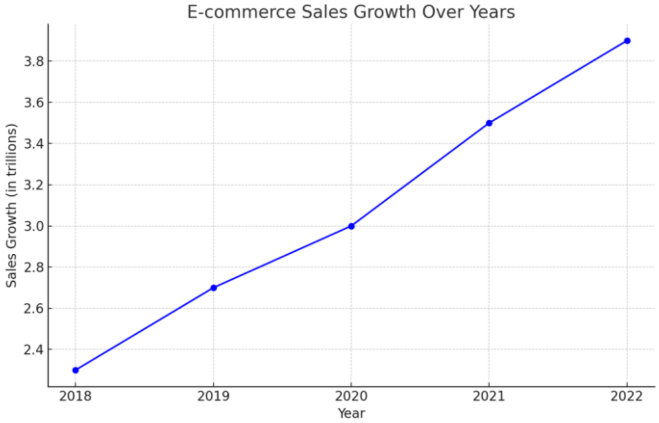

Fig.3: Sales Growth Over Years

3.5 Performance Evaluation

The performance of each model is evaluated using several metrics:

Mean Absolute Error (MAE): Measures the average magnitude of errors in predictions.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \acute{y}_i\right| \quad (4)$$

In equation 4, Where n is the number of observations, y_i is the actual value, $\acute{y}_i$ is the predicted value.

Root Mean Squared Error (RMSE): Evaluates the average squared difference between the expected and actual numbers, taking the square root into account.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2} \quad (5)$$

Equation 5 is expressed as n is the number of observations, y_i is the actual value and $\hat{y}_i$ is the predicted value.

R**-squared** $\left(R^2\right)$**:** Shows the percentage of the dependent variable's volatility that can be predicted based on the independent variables.

$$R^2 = 1 - \frac{SS_{m.}}{SS_{tot}} \quad (6)$$

Where SS_{res} is the sum of squares of residuals and SS_{tot} is the total sum of squares derived in equation 6.

3.6 Data Cleaning

3.6.1 Identifying and Removing Duplicates

Initially, duplicate entries were found and eliminated from the raw data. Inaccuracies and skewing of results might come from duplicate data. These duplicates were eliminated, leaving the dataset clean and distinct. If the date '2023-01-02' had two entries, the additional one was eliminated. The data that had been cleaned provided a strong basis for the subsequent actions as illustrated in Tables 1 and 2.

Table 1. Original Data Sample

Date	Sales	Promotions
2023-01-01	100	1
2023-01-02	200	1
2023-01-02	200	1
2023-01-04	NaN	0
2023-01-05	500	0

Table 2. Cleaned Data Sample

Date	Sales	Promotions
2023-01-01	100	**1**
2023-01-02	200	**1**
2023-01-04	NaN	**0**
2023-01-05	500	**0**

3.6.2 Handling Missing Values

Next, the dataset's missing values were fixed, with a focus on the 'Sales' column. The average sales value was used to fill in the blanks rather than removing rows or leaving gaps. The data's consistency and integrity are preserved by this procedure. To guarantee a complete dataset, for example, the average sales from the available data was used in place of the missing sales data for '2023-01-04' has been mentioned in Table 3.

Table 3. Missing values in the 'Sales' column were imputed with the mean value of the column.

Date	Sales	Pr
2023-01-01	100.00	**1**
2023-01-02	200.00	**1**
2023-01-04	266.67 (mean)	**0**
2023-01-05	500.00	**0**

3.6.3 Normalization

Normalization was carried out after managing missing values. To facilitate the model's processing and learning from the data, this phase scales the numerical features such that they all fall within the same range. 'Sales' and 'Promotions' columns have their values converted to a range of 0 to 1 using min-max scaling. Scaling a sales value of 100 to 0.0 and 500 to 1.0 are two examples of scaling. By doing this step, you can make sure that no single feature, no matter how big, takes over the model has been clarified in Table 4.

Table 4. Numerical features ('Sales' and 'Promotions') were scaled using Min-Max scaling.

Date	Sales	Promotions
2023-01-01	0.00	1.00
2023-01-02	0.33	1.00
2023-01-04	0.44	0.00
2023-01-05	1.00	0.00

3.6.4 Feature Engineering

The dataset was improved by feature engineering in the end. 'DayOfWeek' is a new feature that was developed by taking the day of the week out of the 'Date' field. This additional functionality aids the model in recognizing and capturing patterns associated with particular days. As an example, '2023-01-01' has a 'DayOfWeek' value of 6, since it is a Sunday, whereas '2023-01-02' is a Monday, therefore its value is 0. The inclusion of this element adds more context, which raises the predicted accuracy of the model in Table 5.

Table 5. A new feature, 'DayOfWeek', was created by extracting the day of the week from the 'Date' column.

Date	Sales	Promotions	Day of Week
2023-01-01	100.00	1	6 (Sunday)
2023-01-02	200.00	1	0 (Monday)
2023-01-04	266.67	0	2 (Wednesday)
2023-01-05	500.00	0	3 (Thursday)

3.7 Model Comparison

The best model for predicting e-commerce sales and demand is found by comparing the output from the various models. The comparison is centred on each model's precision, resilience, and computational effectiveness as represented in Figures 4,5 and 6.

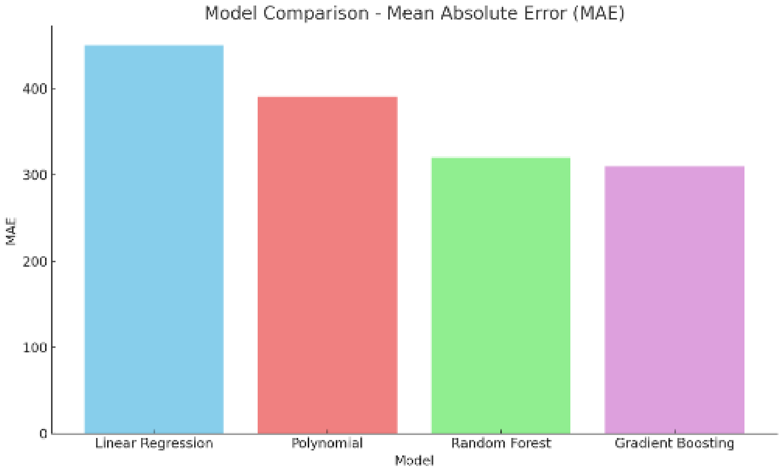

Fig. 4: Mean Absolute Error (MAE)

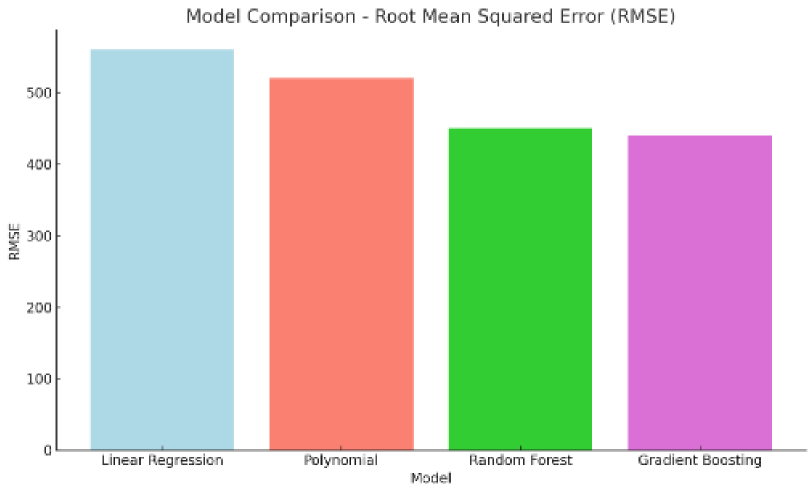

Fig. 5: Root Mean Squared Error (RMSE)

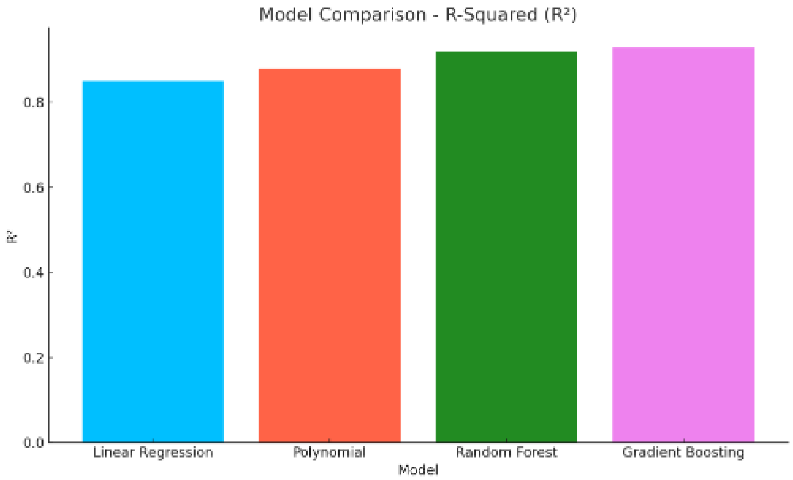

Fig. 6: R-Squared (R^2)

3.8 Comparison of data distribution before and after normalization:

Before normalization, the sales distribution shows the original sales values and the promotions distribution shows the original promotion values. After normalization, the sales values are scaled between 0 and 1, as seen in the sales distribution, and the promotion values are also scaled between 0 and 1, as shown in the promotion's distribution represented in Figures 7 and 8.

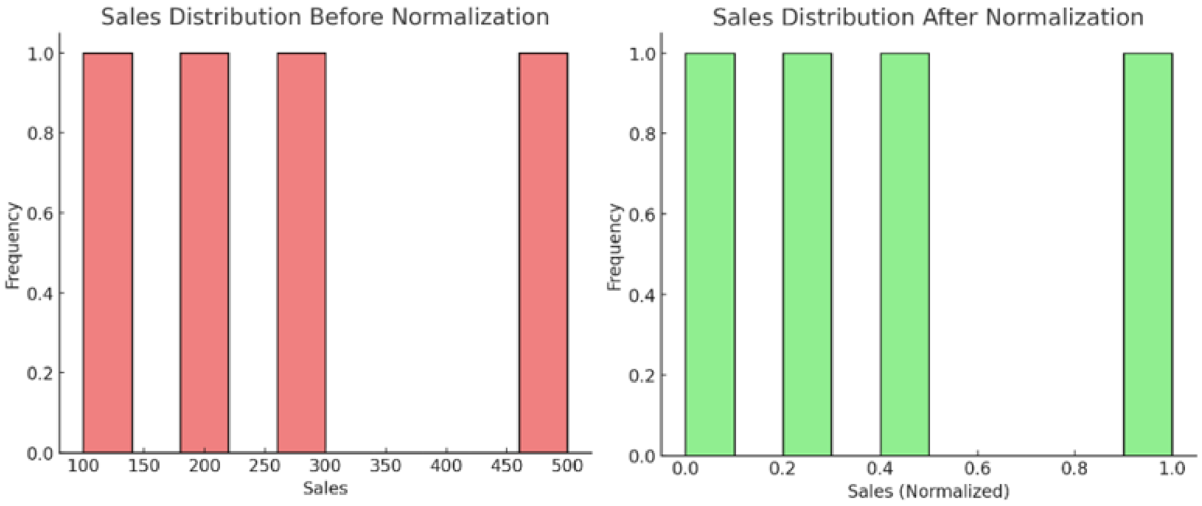

Fig. 7: Sales Distributions in Before and After Normalization

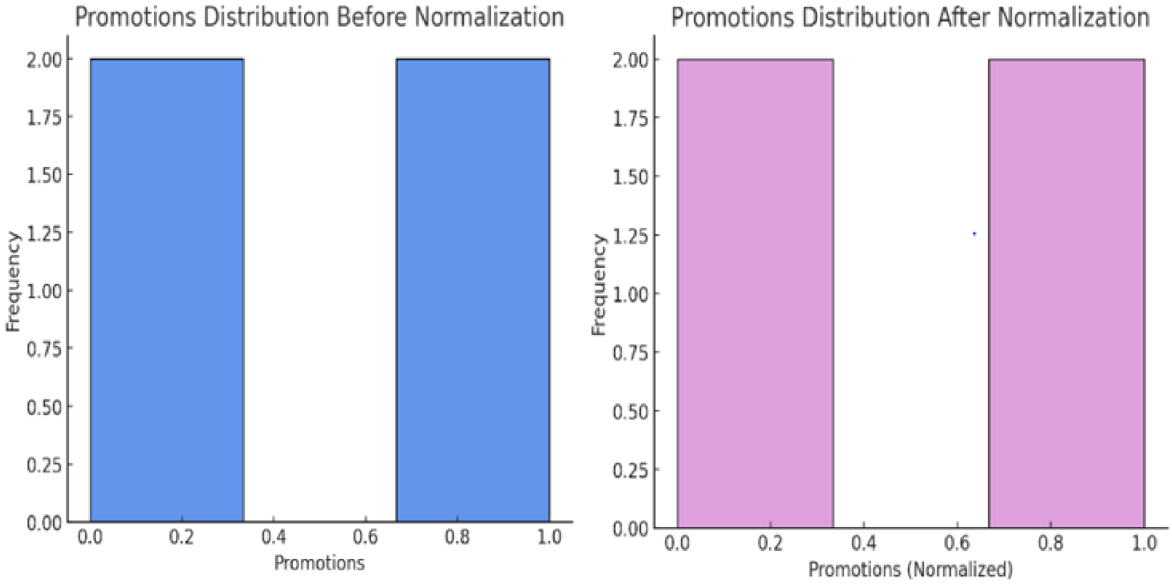

Fig. 8: Promotion Distributions in Before and After Normalization

This methodology offers a comprehensive approach to accurately anticipate demand and sales in e-commerce through the use of state-of-the-art machine learning algorithms. This study aims to provide significant insights by carefully collecting, processing, and analyzing data, as well as by evaluating different models, to assist e-commerce companies in making informed decisions and preserving their competitiveness.

Table 6: The subsequent table displays a comprehensive overview of the model's performance

Model	MAE	RMSE	R^2	Key Insights
Linear Regression	450.32	560.45	0.85	Easy to use and quick, however, it could miss intricate patterns in e-commerce data.
Polynomial	390.22	520.33	0.88	Better at capturing non-linear relationships, but prone to overfitting if not properly adjusted.
Random Forest	320.11	450.25	0.92	Entire datasets and non-linearities are handled with ease by this reliable and precise system.
Gradient Boosting	310.05	440.10	0.93	Intensive computing is required, yet extremely exact and successively corrects faults.

The investigation revealed several significant details regarding the advantages and disadvantages of each forecasting model. The robust baseline provided by linear regression is not ideal for the intricate data patterns that are frequently encountered in e-commerce. While careful tweaking is necessary to avoid overfitting, Polynomial Regression performs a better job of capturing non-linear trends. For big and complicated datasets, Random Forest works incredibly well, providing reliable and accurate predictions. Although it needs more processing resources, gradient boosting iteratively improves the model to achieve the best accuracy.

For e-commerce enterprises, combining these models can lead to even more accurate estimations. Better inventory control and more effective marketing techniques may result from this hybrid strategy. Businesses can make better judgments and keep a competitive edge in the online marketplace by comprehending and utilizing the advantages of each model.

4. Conclusion

The study offers a thorough evaluation of cutting-edge machine learning methods for e-commerce sales forecasting. According to the study, Random Forest and Gradient Boosting provide the best accuracy. Although polynomial regression is good at capturing non-linear patterns, it needs to be carefully adjusted to prevent overfitting. Despite being simple to use, linear regression may overlook intricate patterns in e-commerce data. E-commerce companies can improve inventory control and advertising strategies by integrating these algorithms to get more precise projections. In the ever-changing internet market, this hybrid strategy aids businesses in making well-informed decisions and maintaining a competitive edge.

References:

1. Al-Basha, F. (2021). Forecasting Retail Sales Using Google Trends and Machine Learning (Doctoral dissertation, HEC Montréal).
2. Ajay, R., Joel, R. S., & Prakash, P. O. (2023, June). Analyzing and Predicting the Sales Forecasting using Modified Random Forest and Decision Tree Algorithms. In 2023 8th International Conference on Communication and Electronics Systems (ICCES) (pp. 1649-1654). IEEE.
3. TU, Y. Z. Y., & TU, I. G. G. I. (2022). Data-Driven Daily Product Sales Forecasting in a Third-Party E-Platform Environment.
4. Khakpour, A. (2020). Data science for decision support: Using machine learning and big data in sales forecasting for production and retail.
5. Steenbergen, R. M. (2019). New Product Forecasting with Analogous Products: Applying Random Forest and Quantile Regression Forest to Forecasting and Inventory Management (Master's thesis, University of Twente)
6. Chen, N. (2022). Research on E-Commerce Database Marketing Based on Machine Learning Algorithm. Computational Intelligence and Neuroscience, 2022(1), 7973446.
7. Shaohui, D., & Kudryavtsev, D. (2021). Consumer Repurchase Behavior Prediction Based on Different Fusion Models.
8. Kumar, M. M., Venkat, A. S., Balaji, M. V. N., Kumar, C. N., Srithar, S., & Aravinth, S. S. (2023, November). Driving E-commerce Success with Advanced Machine Learning: Customer Purchase Pattern Insights. In 2023 International Conference on Sustainable Communication Networks and Application (ICSCNA) (pp. 1196-1203). IEEE.
9. Moroke, N. D., & Makatjane, K. (2022). Predictive modelling for financial fraud detection using data analytics: a gradient-boosting decision tree. In Applications of Machine Learning and Deep Learning for Privacy and Cybersecurity (pp. 25-45). IGI Global.