Case Study of AI Project

Sales Prediction of Big Mart Based on Linear Regression, Random Forest, and Gradient Boosting

Ranveer Patil (22305107)
Ravi Virani (22307118)
Arshita Thummer (12402540)
Nevil Rafaliya (22305913)
Allan Johns (22304075)

Applied AI for Digital Production Management
Technische Hochschule Deggendorf Cham, Deutschland.
July 09, 2024

Abstract

Accurate forecasting of sales is one of the fundamental tasks for any retailer in the current state of developing machine learning and data analytics; it enhances strategic planning and maximization of profits. The present case study handles a small retail supermarket called Big Mart and it applies machine learning methods to predict sales. This study will assess the performance of three different models—Linear Regression, Random Forest, and Gradient Boosting—through the analysis of a relatively small and specific dataset. Model comparison was based on error metrics: Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and the R^2 score. We can see that a Linear Regression model usually underfits data causing significant errors in prediction. On the other hand, Decision Tree-based models such as Random Forest or Gradient Boosting can get much better accuracy; Gradient Boosting prevails slightly in this race. In conclusion, despite some weaknesses in the data, the current study has revealed that valuable results could be derived for sales forecasting in small retail environments by employing machine learning techniques. The research also shows some challenges and recommendations for improvement in the prediction of sales with machine learning.

1 Introduction

In today's retail environment, accurately predicting sales is vital for a business's success. Effective sales forecasting allows retailers to make well-informed decisions about inventory management, pricing strategies, and resource allocation, which in turn optimizes operational efficiency and boosts profitability [1]. The rise of artificial intelligence (AI) and machine learning (ML) has transformed this field, providing advanced tools that can analyze extensive datasets and detect patterns that traditional methods might miss [2].

This report explores the use of AI for sales forecasting in a small retail setting. It focuses on evaluating the feasibility and performance of different machine learning models in predicting sales for Big Mart, a fictional small retail supermarket. Small-scale businesses often face challenges such as limited data and resources, making accurate sales prediction more difficult but no less important [3]. We use three well-known machine learning techniques in our case study: Linear Regression, Random Forest, and Gradient Boosting. These models are applied to a dataset containing sales figures, product details, and store characteristics. Each model has its own advantages: Linear Regression is simple and easy to interpret, Random Forest is robust and good with non-linear data, and Gradient Boosting offers high predictive accuracy through iterative refinement [9].

The main goals of this study are to determine how suitable these machine learning models are for small retail environments, compare their performance based on accuracy and reliability, and identify any challenges and limitations they present. We evaluate the models using key performance metrics like Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and the coefficient of determination (R^2 score). This introduction sets the groundwork for a comprehensive analysis of how AI can improve sales forecasting for small retailers. By leveraging machine learning, even businesses with limited data and resources can significantly enhance their forecasting abilities [7]. The results of this study are intended to provide practical insights and useful guidelines for small retailers aiming to use AI for better decision-making and strategic planning.

2 Materials and Methods

2.1 Materials

2.1.1 Data Source

The dataset for this study was sourced from Kaggle and represents a fictional retail supermarket named Big Mart. It includes sales data for 1559 products sold across 10 different stores in various locations. Each entry in the dataset provides several attributes pertaining to both the products and the stores. These attributes are divided into dependent and independent variables:

- **Dependent Variable**: The quantity of each product sold at the various locations.
- **Independent Variables**: A total of 11 variables which are categorized as follows:
  - **Categorical Variables**: Product ID, product type, fat content, outlet ID, store size, location type, and outlet type.
  - **Numerical Variables**: Product weight, product visibility, maximum retail price (MRP), and the year the outlet opened.

2.1.2 Library and Software

- **Python 3.9**: This version was selected for data analysis and machine learning due to its robust libraries and ease of use.
- **Pandas**: A powerful library for data manipulation and analysis, crucial for tasks such as managing the dataset, preprocessing, and cleaning.
- **Scikit-learn**: This toolkit provides essential functions for clustering, regression, and classification, among other tools needed for building machine learning models.

- **Matplotlib**: A versatile plotting library used to create static, interactive, and animated visualizations in Python.
- **Seaborn**: Based on Matplotlib, this statistical data visualization library is used to produce attractive and informative statistical graphics.
- **Jupyter Notebook**: An open-source web application that enables the creation and sharing of documents containing live code, equations, visualizations, and narrative text.
- **SciPy**: A library used for scientific and technical computing, offering additional functionality to complement NumPy.

Project flow chart

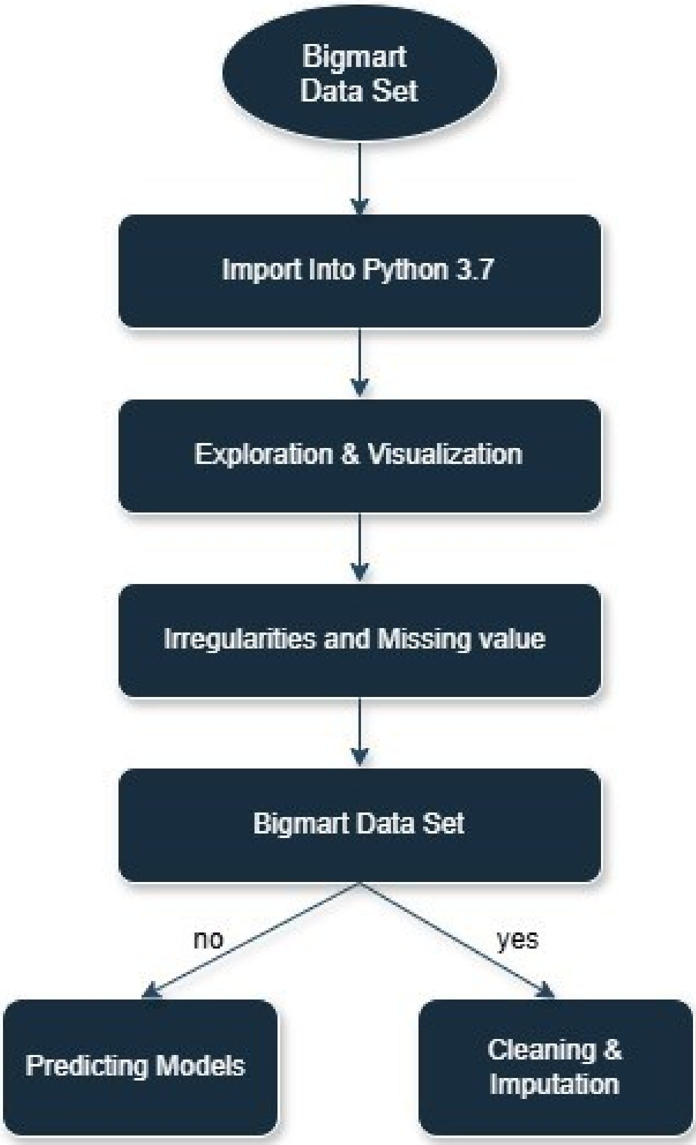


Figure 1: Heatmap of Correlation Coefficients

The flow diagram illustrates a structured process for managing the Bigmart dataset in an AI case study. It starts with importing the dataset into Python 3.7, followed by data exploration and visualization to understand its characteristics. The next step involves identifying and addressing any irregularities and missing values. If issues are detected,the data undergoes cleaning and imputation to ensure high quality. Once the dataset is prepared, predictive models are developed to make accurate forecasts. This This systematic approach ensures that the AI models produce reliable and meaningful results.

2.2 Methods

2.2.1 Data Preprocessing

To make the dataset suitable for machine learning models, several preprocessing steps were carried out:

- **Handling Missing Values**: Missing values were found in the 'Item_Weight' and 'Outlet_Size' columns. To maintain data integrity and avoid introducing bias, rows with missing values were removed instead of imputing them.

- **Encoding Categorical Variables**: Categorical variables such as 'Item_Fat_Content', 'Item_Type', 'Outlet_Size', 'Outlet_Location_Type', and 'Outlet_Type' were converted into numerical values using one-hot encoding. This step was essential for the machine learning algorithms to process the categorical data effectively.

- **Feature Selection**: Features were selected based on their relevance and correlation with the target variable 'Item_Outlet_Sales'. Features that were redundant or had low correlation were either consolidated or excluded to enhance model performance.

- **Train-Test Split**: The dataset was divided into training and testing sets with an 80%-20% split. This ensures that the models are trained on a significant portion of the data while being tested on unseen data to evaluate their performance accurately.

2.2.2 Machine Learning Models

Three machine learning models [4] were chosen for this study due to their diverse characteristics and effectiveness in regression tasks:

- **Linear Regression**: This basic model assumes a linear relationship between the independent variables and the dependent variable (sales). It is straightforward to implement and interpret but might not capture complex patterns in the data.

- **Random Forest**: An ensemble learning method that constructs multiple decision trees and combines their results to enhance prediction accuracy and control overfitting. It handles non-linear relationships well and is robust to outliers [6].

- **Gradient Boosting**: Another ensemble technique that builds models sequentially, with each new model correcting the errors made by the previous ones. Gradient Boosting is known for its high predictive accuracy and ability to handle complex data patterns [5].

2.2.3 Performance Metrics

The models were evaluated using three key performance metrics:

- **Root Mean Square Error (RMSE)**: This metric measures the average magnitude of the errors between predicted and actual values. It is sensitive to large errors and provides a sense of prediction accuracy.
- **Mean Absolute Error (MAE)**: This metric represents the average absolute difference between predicted and actual values. It indicates the average prediction error in the same units as the target variable.
- **Coefficient of Determination ($\mathbf{R^2}$ Score)**: This metric indicates the proportion of the variance in the dependent variable that can be predicted from the independent variables. A higher R^2 score signifies better model performance.

2.2.4 Procedure

After preprocessing the data and selecting the models, the next steps were feature selection and model training. Features were chosen based on their correlation with the target variable 'Item_Outlet_Sales'. The correlation coefficients were calculated using Cramér's V [8], which measures the association between nominal variables. This statistic was used as the criterion for feature selection.

Additionally, the Dython library in Python was used to calculate these correlation coefficients. This library provides tools to measure the strength and direction of relationships between categorical features and the target variable.

3 Results

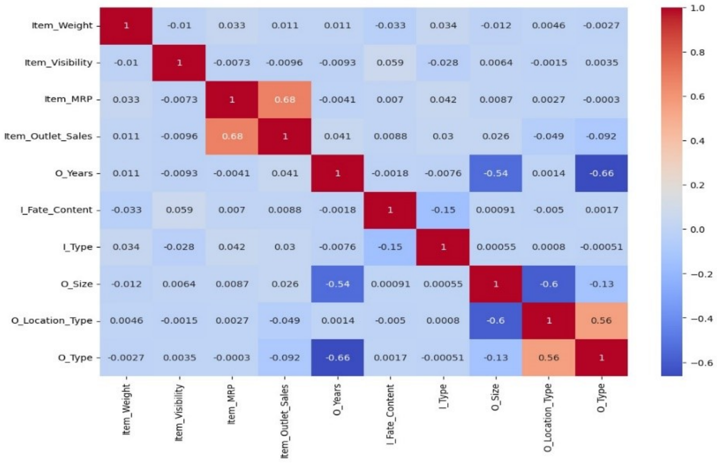


Figure 2: Heatmap of Correlation Coefficients

The heatmap in Figure 2 displays the correlation coefficients among the variables. The item MRP (Maximum Retail Price) exhibited the highest correlation with the target variable outlet sales with a coefficient of 0.68. Other variables such as outlet size and outlet type also showed some correlation. However, variables like item weight, visibility, and fat content had very low correlation with sales suggesting they have limited predictive value for the models.

3.1 Model Performance

3.1.1 Linear Regression

The performance of the Linear Regression model is shown in the figures for both the training and test data. The model displayed significant underfitting. The scatter plots demonstrate a linear trend, but there's a substantial spread around the ideal prediction line (red dashed line).

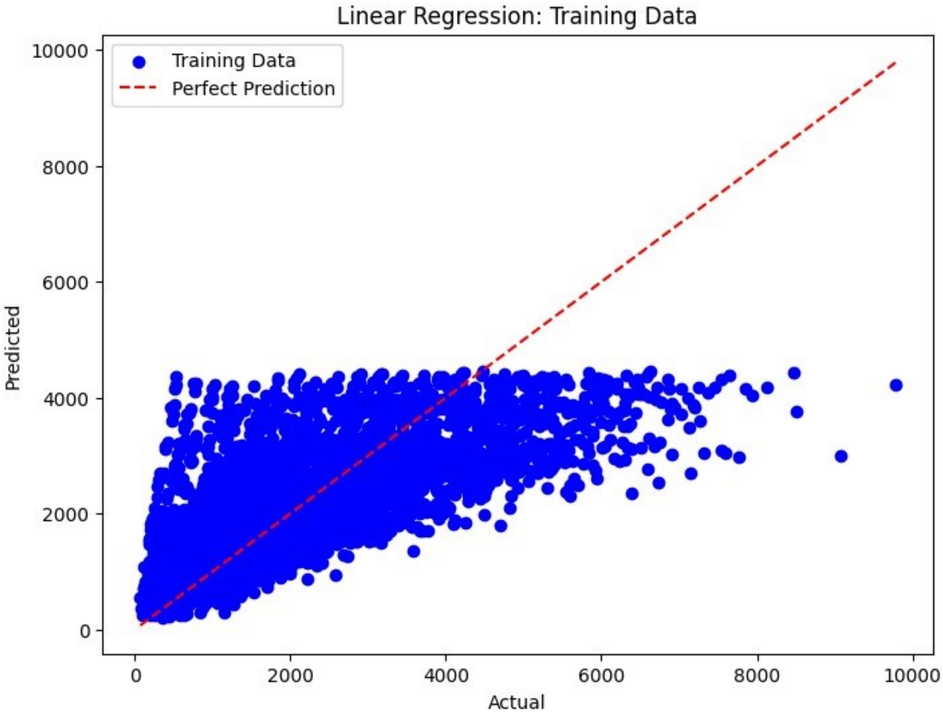


Figure 3: Scatter Plot of Linear Regression (Train)

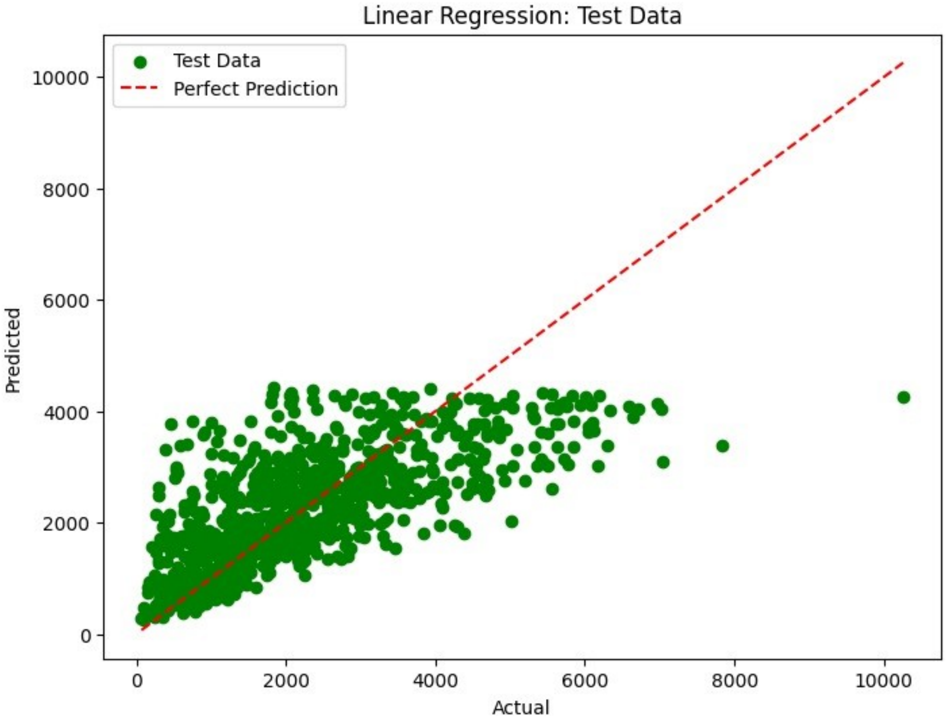


Figure 4: Scatter Plot of Linear Regression (Test)

For higher sales values, the model's predictions become less accurate with larger deviations from the actual values as seen from the wider dispersion of points away from the ideal line. Additionally, the model sometimes predicts negative sales values which are unrealistic and highlight its poor performance.

3.1.2 Random Forest

The Random Forest model performed better than Linear Regression as depicted in the figures. The scatter plots show that the Random Forest model's predictions are tightly clustered around the ideal prediction line, particularly for the training data indicating a good fit. However, the model exhibits overfitting as the test data shows a wider dispersion of points compared to the training data [4].

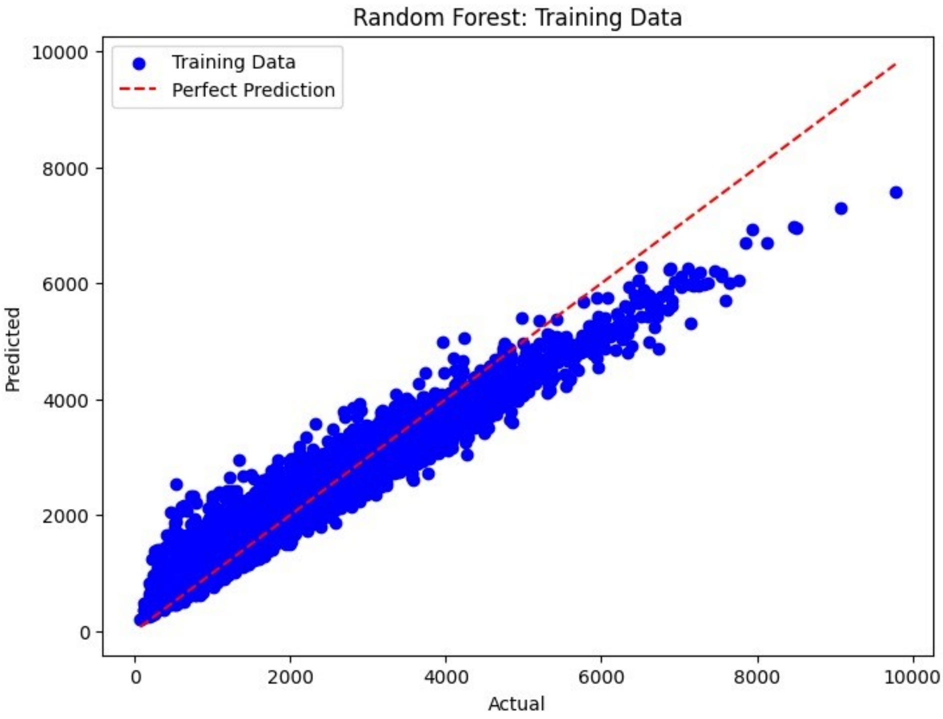


Figure 5: Scatter Plot of Random Forest (Train)

The model is accurate for sales values up to around 6000 but struggles with higher sales values resulting in larger prediction errors for these cases.

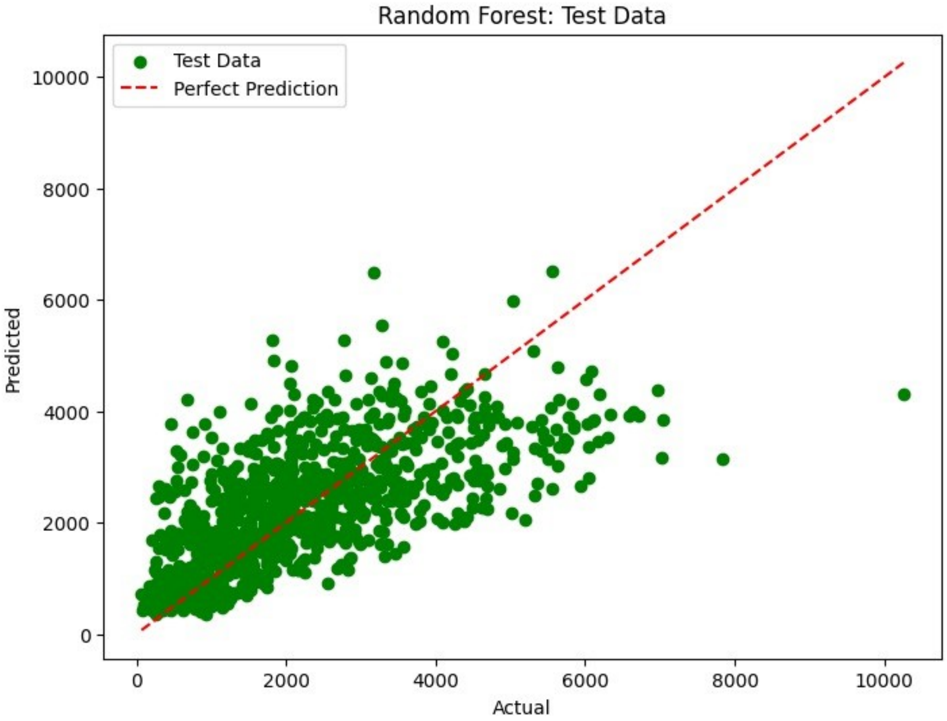


Figure 6: Scatter Plot of Random Forest (Test)

3.1.3 Gradient Boosting

Gradient Boosting demonstrated a balanced performance effectively navigating between underfitting and overfitting as shown in the figures [6].

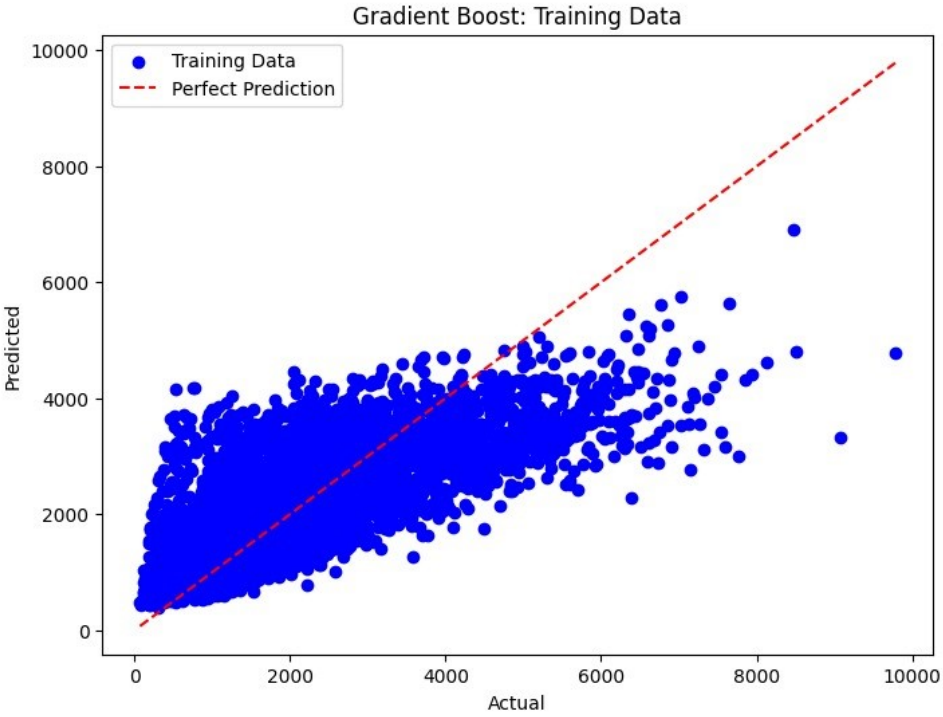


Figure 7: Scatter Plot of Gradient Boosting (Train)

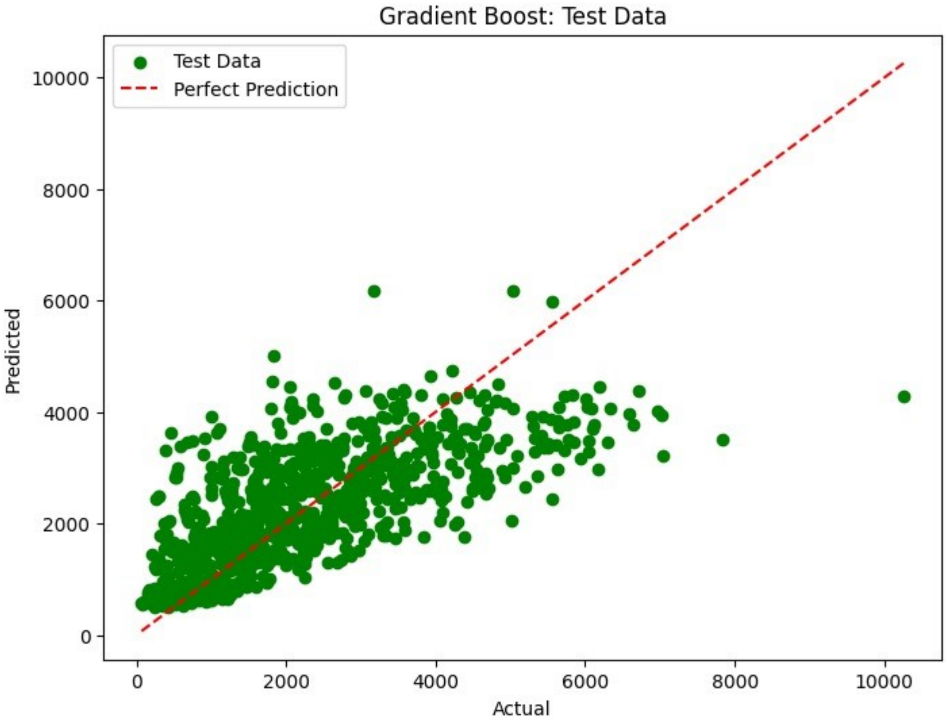


Figure 8: Scatter Plot of Gradient Boosting (Test)

The scatter plots indicate a relatively good fit with points clustering more closely around the ideal prediction line compared to Linear Regression. Additionally, it performed better on the test data than the Random Forest model. The model exhibits a linear trend with less deviation suggesting it generalizes better than Random Forest. However, similar to Random Forest, Gradient Boosting also struggles with higher sales values showing larger prediction errors in these cases.

3.2 Distribution of Predicted Values

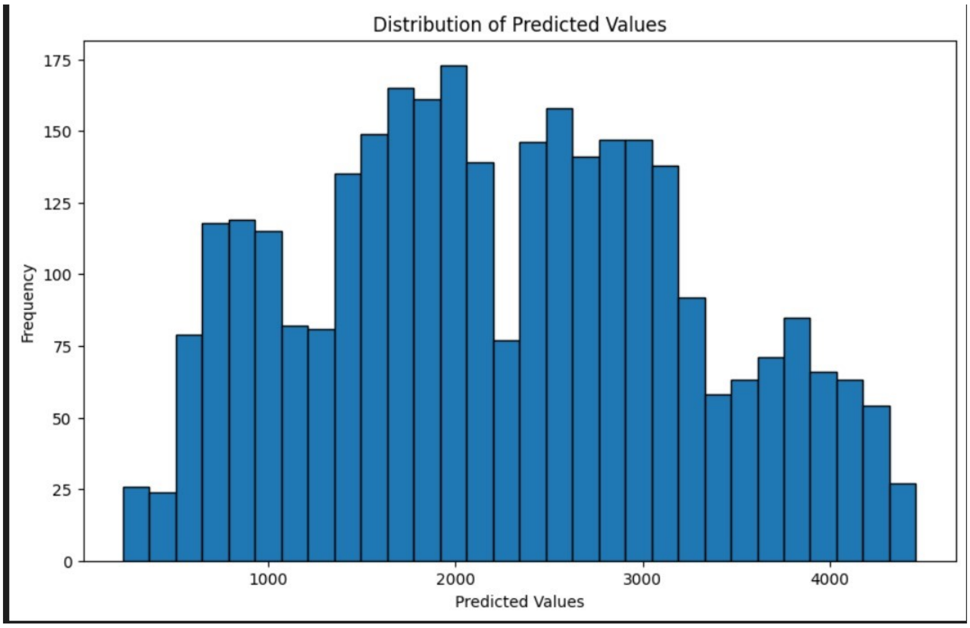


Figure 9: Histogram of Predicted Values

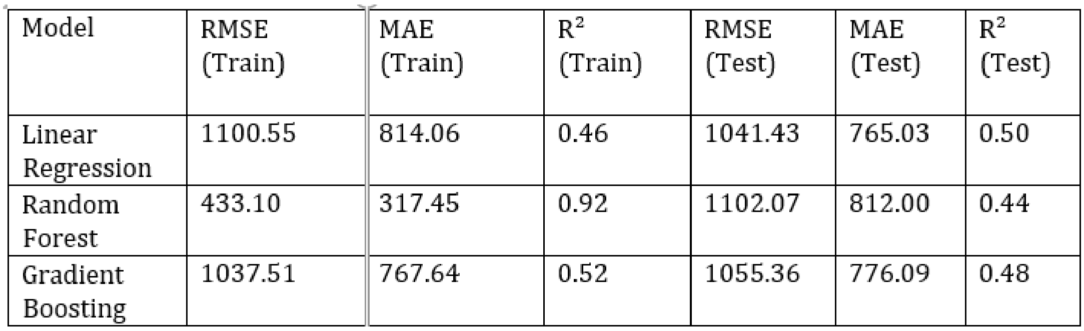
The histogram in Figure 9 illustrates the distribution of predicted sales values generated by the sales prediction model. The x-axis represents the predicted sales values while the y-axis indicates how frequently these values occur. This visualization helps to understand the spread of the model's predictions across different sales ranges.

From the histogram, it's clear that the predicted sales values span a wide range from below 1000 to over 4000. The frequency of predictions peaks in the 1500 to 2000 and 2500 to 3000 ranges indicating that the model often predicts sales within these intervals. Predictions are less common at the extremes, below 1000 and above 4000, suggesting these sales values are less frequently anticipated by the model.

This distribution offers insights into the model's behavior, revealing its tendency to predict certain sales ranges more often. Analyzing the distribution of predicted values is crucial for assessing the model's performance and identifying any potential biases, ensuring that the predictions align well with the actual sales patterns observed in the dataset.

3.3 Summary of Model Performance

In summary, Gradient Boosting outperformed the other models on the test set, striking a balance between fitting the training data and generalizing well to the test data. Random Forest showed the best performance on the training set but suffered from overfitting,



Model	RMSE (Train)	MAE (Train)	R^2 (Train)	RMSE (Test)	MAE (Test)	R^2 (Test)
Linear Regression	1100.55	814.06	0.46	1041.43	765.03	0.50
Random Forest	433.10	317.45	0.92	1102.07	812.00	0.44
Gradient Boosting	1037.51	767.64	0.52	1055.36	776.09	0.48

Figure 10: Performance of MOdel

failing to generalize as effectively. Linear Regression performed the worst, demonstrating significant underfitting. Both the RMSE (1055.36) and MAE (776.09) metrics indicate that Gradient Boosting had the best generalization performance among the three models.

4 Discussion

This study aimed to evaluate the performance of three machine learning models, Linear Regression, Random Forest, and Gradient Boosting—in predicting sales for a small-scale retail supermarket, Big Mart. Additionally, we developed a web-based application to provide sales predictions using these models, allowing users to input data via an Excel file or a single row. The research question focused on identifying the model that offers the best balance of accuracy and generalizability given the constraints of a limited dataset and practical implementation in a web tool.

4.1 Key Insights

- **Linear Regression**: This model exhibited significant underfitting with a wide spread of predictions, especially for higher sales values. It often produced unrealistic negative predictions, highlighting its limitations with categorical features and variables with low correlation.

- **Random Forest**: While it outperformed Linear Regression, it showed signs of overfitting. It excelled on the training set but struggled on the test set, particularly with higher sales values. The web application indicated better accuracy for mid-range sales but issues with extreme values.

- **Gradient Boosting**: This model delivered the most balanced performance, fitting both training and test data well. It demonstrated superior accuracy with fewer errors, though it still struggled with higher sales values. The web application confirmed its reliability for a wide range of sales values, providing consistent and realistic predictions.

4.2 Strengths of the Study

- **Model Variety and Comprehensive Evaluation**: The use of three distinct machine learning models allowed for a thorough comparison and multiple performance metrics (RMSE, MAE, R^2) enabled a robust assessment [9].

- **Practical Application**: The web-based tool demonstrated practical applicability, offering an accessible platform for small retailers to input data and receive sales predictions.

4.3 Limitations

- **Data Constraints**: The small, undiversified dataset contributed to underfitting in Linear Regression and overfitting in Random Forest. Low correlation among features also limited predictive power [1].
- **Exclusion of External Factors**: The absence of key external factors such as market conditions, customer behavior, and competition further constrained model accuracy [10].
- **Performance on Extreme Values**: All models showed larger prediction errors for higher sales values, affecting their suitability for predicting extreme outcomes as reflected in the web tool's performance [3].

4.4 Future Research and Development

- **Integration of External Data**: Incorporating market trends, customer demographics, and competitive actions to enhance model accuracy [2, 5].
- **Development of Hybrid Models**: Investigating hybrid models and time series analysis to improve results [5].
- **Advanced Techniques**: Employing sophisticated models like LSTM networks or deep learning techniques [7].
- **Expanding Dataset**: Increasing the dataset size and diversity to reduce underfitting and overfitting.
- **Enhancements to the Web Tool**: Improving the web application to handle larger datasets and provide detailed analytics.

5 Conclusion

This case study explored the application of machine learning techniques for sales prediction in a small-scale retail setting using Big Mart as the focus. We evaluated three models: Linear Regression, Random Forest, and Gradient Boosting to determine their predictive performance. The Random Forest model performed well during training but exhibited overfitting when tested on new data. In contrast, Gradient Boosting provided a more balanced performance across both the training and test datasets. Linear Regression, being the simplest model, showed the lowest accuracy.

Our research demonstrated that machine learning can effectively forecast sales even with a limited dataset, emphasizing the significance of model selection and feature relevance. In conclusion, our research question was answered affirmatively: machine learning techniques, particularly Gradient Boosting, can be used to predict sales in a small retail environment, offering valuable insights and supporting strategic decision-making for small retailers.

References

[1] Dalrymple, D. J. (1987). Sales forecasting practices: Results from a United States survey. *International Journal of Forecasting*, 3(3-4), 379-391.

[2] Liu, X., & Ichise, R. (2017, July). Food sales prediction with meteorological data—a case study of a Japanese chain supermarket. In *International Conference on Data Mining and Big Data* (pp. 93-104). Springer, Cham.

[3] Sharma, S. K., Chakraborti, S., & Jha, T. (2019). Analysis of book sales prediction at Amazon marketplace in India: a machine learning approach. *Information Systems and e-Business Management*, 17(2), 261-284.

[4] Hu, C. Y., & Griffith, D. A. (2022). Incorporating spatial autocorrelation into house sale price prediction using random forest model. *Transactions in GIS*, 26(5), 2123–2144.

[5] Manikandan, S., Deetshiha, A., Sushmitha, D. J., et al. (2022). Intelligent sales prediction using ARIMA techniques. *AIP Conference Proceedings*, 2444(1).

[6] Xia, X., Wu, S., Sun, L., et al. (2020). ForeXGBoost: passenger car sales prediction based on XGBoost. *Distributed and Parallel Databases: An International Journal*, 38(3), 713–738.

[7] Canton, C. R., Gibaja, D. E., & Caballero, S. O. (2019). Sales Prediction through Neural Networks for a Small Dataset. *International Journal of Interactive Multimedia and Artificial Intelligence*, 5(4), 35–41.

[8] Cramér, H. (2016). Mathematical Methods of Statistics (PMS-9). Princeton: Princeton University Press.

[9] Caruana, R., & Alexandru, N. M. (2006). An Empirical Comparison of Supervised Learning Algorithms. *Proceedings of the 23rd International Conference on Machine Learning – ICML 06.*

[10] Armstrong, J. S. (1999). Sales Forecasting. *The IEBM Encyclopedia of Marketing*, 278-290. Retrieved from `https://repository.upenn.edu/marketing_papers/237`

6 Acknowledgments

Student Name	Contribution
Ranveer Patil	Contributes to establishing the project objective, Software/Package Implementation, Project Flowchart, Results Analysis: Ranveer played a vital role in analyzing the outcomes, interpreting performance metrics, and extracting key insights from the productivity prediction model.
Ravi Virani	Responsible for Data modeling, back-end and front-end (user interface) programming, and deployment.
Nevil Rafaliya	Methodology Design: Step-by-Step Instructions, Report Writing: Nevil led the development of the report's content, ensuring it was clear, coherent, and complied with the specified format and guidelines.
Arshita Thummer	Arshita played a crucial role in feature selection and engineering, identifying relevant features, and transforming raw data into meaningful inputs for the machine learning models.
Allan Johns	Responsible for Data collection.