INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 05, Issue 08, August 2025, pp : 1466-1471

www.ijprems.com
editor@ijprems.com

e-ISSN : 2583-1062
Impact Factor : 7.001

BIG MART SALE PREDICTION USING LINEAR REGRESSION

Miriyala Rakesh[1], B. Srinivasa SP Kumar[2]

[1]MCA III Semester, MCA Dept, Chaitanya Bharathi Institute Of Technology (A), Gandipet, Hyderabad - 500 075, India.

[2]Sr Assistant Professor, MCA Dept, Chaitanya Bharathi Institute Of Technology (A), Gandipet, Hyderabad - 500 075, India.

ABSTRACT

This research paper presents the results of predicting Big Mart product sales using machine learning algorithms. The objective of this work is to build a model that accurately estimates sales across different outlets based on product and store-related attributes such as Item Weight, Item Visibility, Item MRP, Outlet Size, Outlet Location, and Outlet Type. Accurate sales prediction is essential for inventory management, demand forecasting, pricing strategies, and business decision-making for retail companies. Unlike traditional forecasting methods, which often fail to capture complex relationships and multiple influencing factors, this approach leverages machine learning to deliver better accuracy and interpretability. In this work, Linear Regression and Ensemble Models are employed to establish relationships between independent variables (predictors) and the dependent variable (sales).

The methodology includes dataset loading, data preprocessing, categorical encoding, exploratory data analysis, feature selection, and splitting the dataset into training and testing sets. The models are trained to minimize prediction errors and capture meaningful sales patterns across products and outlets. Evaluation of model performance is carried out using metrics such as R^2 Score, Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE) to ensure robustness and reliability.

This approach provides insights into how key factors—such as product pricing, visibility, and outlet characteristics—affect sales, helping retailers optimize supply chain planning and enhance customer satisfaction. By implementing a scalable and interpretable machine learning model, this study demonstrates the applicability of data-driven techniques for predictive analytics in the retail domain.

Keywords: Big Mart, Sales Prediction, Machine Learning, Linear Regression, Ensemble Models, Supervised Learning.

1. INTRODUCTION

Retail sales play a vital role in the consumer goods industry, driving business growth and profitability. Accurately predicting sales is crucial for retailers, manufacturers, and supply chain managers, as it supports effective inventory management, demand forecasting, pricing strategies, and overall decision-making. However, estimating product sales is challenging due to the involvement of multiple influencing factors such as product features, outlet type, outlet size, pricing, visibility, and geographical location.

Traditional forecasting methods often provide baseline estimations but struggle to effectively capture the impact of multiple interrelated variables. In recent years, machine learning (ML) techniques have emerged as powerful tools for predictive analytics in retail, offering data-driven solutions that can model complex relationships among diverse features and adapt to dynamic consumer demand.

This research focuses on applying Linear Regression, a widely used supervised machine learning algorithm, to predict sales based on product and store-related attributes. The Linear Regression model establishes a linear relationship between independent variables (such as Item Weight, Item MRP, Item Visibility, and Outlet Characteristics) and the dependent variable (sales) by minimizing prediction error using the least squares method.

The methodology includes systematic steps such as dataset loading, data cleaning, categorical encoding, exploratory data analysis, feature selection, splitting data into training and testing sets, model training, and evaluation. Performance is assessed using metrics like R^2 Score, Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE) to ensure accuracy and reliability.

The proposed approach delivers interpretable and scalable predictions, helping retailers optimize sales strategies, enhance supply chain efficiency, and better understand the factors influencing consumer demand. By integrating machine learning with retail sales forecasting, this work contributes to improving transparency, accuracy, and decision-making in the retail industry**.**

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
www.ijprems.com
editor@ijprems.com
Vol. 05, Issue 08, August 2025, pp : 1466-1471
e-ISSN : 2583-1062
Impact Factor : 7.001

1.1 Technologies/Tools focused

The technologies and tools used in this project are:

- The technologies and tools used in this project are:
- Machine Learning Algorithm: Linear Regression for predictive modeling of Big Mart sales.
- Programming Language: Python for implementation of the model and data processing.
- Frameworks and Libraries:
- pandas and NumPy for data manipulation and numerical computations.
- scikit-learn for building, training, and evaluating the Linear Regression model.
- Matplotlib and Seaborn for data visualization and generating plots.
- The key deliverables of this project include:
- A trained Linear Regression model capable of predicting sales based on product attributes and store-related factors.
- A complete workflow including data loading, preprocessing, feature encoding, model training, and evaluation.
- Visual representations such as correlation heatmaps, regression plots, and distribution graphs illustrating data insights and model performance metrics like R^2 score, MAE, and RMSE.

1.2 Problem Statement

Predicting sales accurately is crucial for retail businesses like Big Mart as it enables better inventory management, pricing strategies, and demand forecasting. However, estimating these sales is challenging due to multiple influencing factors such as product type, MRP, visibility, outlet size, and location. The objective of this project is to build a supervised machine learning model using Linear Regression to predict Big Mart product sales across different outlets with reasonable accuracy.

Big Mart managers, retailers, and stakeholders require a tool to accurately predict sales for effective decision-making in stock management, promotional planning, and store performance evaluation. This helps ensure reduced losses, optimized operations, and improved business profitability.

2. RELATED WORK

Research on predicting retail sales using machine learning has gained significant attention, with various studies examining different algorithms and approaches to improve forecast accuracy. Existing work generally falls into three categories: traditional statistical models, basic regression techniques, and advanced machine learning models.

A major focus of related work has been comparing the performance of different supervised learning algorithms:

Linear Regression (LR): One of the most commonly used models for sales prediction due to its simplicity and interpretability. Many studies have shown that Linear Regression provides a strong baseline for predicting product sales by modeling the relationship between independent variables (such as MRP, item visibility, and outlet size) and sales. However, it assumes a linear relationship and may not fully capture complex interactions.

Decision Tree Regressors: Several studies have applied decision tree-based models, which are better at handling non-linear relationships compared to Linear Regression. They often achieve improved accuracy but can suffer from overfitting if not properly tuned.

Advanced Ensemble Models: Some research has explored ensemble techniques such as Random Forest and XGBoost, which combine multiple decision trees to enhance predictive performance. These models generally outperform simpler regression approaches in terms of accuracy and generalization, but they require more computational resources and careful parameter tuning.

Overall, Linear Regression remains widely used in both academic and practical applications due to its ease of implementation, interpretability, and relatively low computational cost, making it a suitable choice for projects focused on fundamental machine learning concepts and sales forecasting.

3. METHODOLOGY

Building on the insights from previous research, this study focuses on predicting retail sales using **Linear Regression** applied to a structured dataset that includes product-level and store-level attributes such as Item_Weight, Item_Visibility, Item_MRP, Outlet_Size, and Outlet_Type. This algorithm was chosen for its simplicity, interpretability, and effectiveness in modeling continuous numeric outcomes like sales.

The proposed methodology follows a structured machine learning workflow to ensure accuracy, scalability, and reproducibility. The process consists of the following phases:

www.ijprems.com
editor@ijprems.com
INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 05, Issue 08, August 2025, pp : 1466-1471
e-ISSN : 2583-1062
Impact Factor : 7.001

3.1 Data Collection

The dataset used is the **Big Mart Sales dataset**, which contains more than 8,500 records with multiple product and outlet attributes.

- **Independent variables:** Item_Weight, Item_Fat_Content, Item_Visibility, Item_Type, Item_MRP, Outlet_Establishment_Year, Outlet_Size, Outlet_Location_Type, Outlet_Type
- **Dependent variable:** Item_Outlet_Sales (sales value)

3.2 Data Preprocessing

1. Checked for missing values (e.g., Item_Weight, Outlet_Size) and handled them using imputation methods.
2. Encoded categorical variables using Label Encoding and One-Hot Encoding:
3. Item_Fat_Content → (Low Fat, Regular, etc.)
4. Item_Type → (Dairy, Snacks, Beverages, etc.)
5. Outlet_Type → (Grocery Store, Supermarket Type1/2/3)
6. Normalization/Scaling was applied where necessary to ensure numerical features contributed effectively to the regression model.

3.3 Exploratory Data Analysis (EDA)

a. Visualized feature distributions using histograms and count plots (e.g., Item_Weight, Item_MRP, Outlet_Size).
b. Analyzed correlations between features and the target variable (Item_Outlet_Sales) using heatmaps and scatter plots.
c. Examined the impact of outlet location, size, and item categories on sales performance.

3.4 Model Selection and Training

This study focuses on **Linear Regression** as the primary algorithm for predicting retail sales. Linear Regression was chosen for its interpretability, computational efficiency, and suitability for modeling continuous numeric outcomes such as sales values.

The model training process includes the following steps:

- **Train-Test Split:** The dataset is divided into **80% training and 20% testing** subsets using scikit-learn's train_test_split to ensure unbiased evaluation.
- **Model Fitting:** The Linear Regression model is trained using the training data, learning the relationship between product/outlet features and the target variable (Item_Outlet_Sales).
- **Cross-Validation:** To improve model reliability, **k-fold cross-validation (k=5)** is applied to validate stability and reduce overfitting.
- **Evaluation Metrics:** Performance is assessed using **R^2 score** and **Mean Absolute Error (MAE)** for both training and testing sets.

This approach ensures that the model generalizes well to unseen data while maintaining simplicity and interpretability in predictions.

3.5 Model Evaluation

The trained Linear Regression model is evaluated on the test dataset to measure prediction accuracy and model effectiveness. The following performance metrics are used:

- **Mean Absolute Error (MAE):** Measures the average magnitude of errors between predicted and actual sales without considering their direction.
- **Root Mean Squared Error (RMSE):** Evaluates the square root of the average squared differences between predictions and actual sales, penalizing larger errors more.
- **Coefficient of Determination (R^2 Score):** Indicates how well the model explains the variability in sales, with values closer to 1 representing a better fit.

These metrics collectively help in validating the reliability and predictive capability of the model for real-world sales forecasting.

3.6 Prediction and Analysis

The trained Linear Regression model is used to predict **Item_Outlet_Sales** based on features such as Item_MRP, Item_Visibility, Outlet_Type, and Outlet_Size. The predicted values are compared with actual sales from the test set to assess accuracy and validate the model's effectiveness. The analysis further highlights how factors like product pricing

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 05, Issue 08, August 2025, pp : 1466-1471
www.ijprems.com
editor@ijprems.com
e-ISSN : 2583-1062
Impact Factor : 7.001

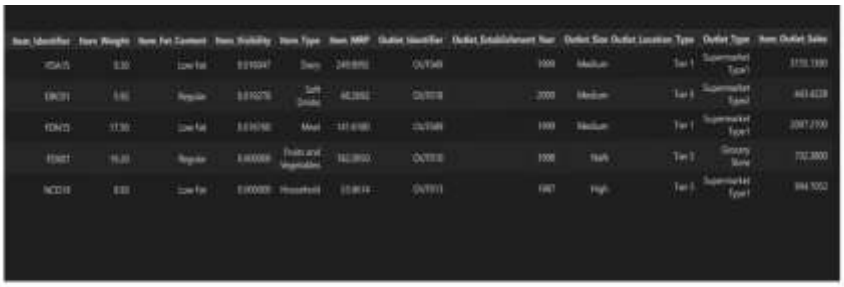
and store characteristics influence sales outcomes.

3.7 Implementation

The system is implemented in Python using Pandas and NumPy for data handling, Scikit-learn for Linear Regression modeling, and Matplotlib/Seaborn for visualization. The model predicts insurance costs based on user input and can be extended for integration into a web application for real-time predictions

4. RESULTS AND DISCUSSION

Dataset Shape



Screen 3.1: Sample of the dataset displaying attributes like item weight, item fat content, item visibility, item type,item MRP, outlet identifier, outlet year, outlet size,outlet location type, outlet type,outlet sales data prediction.

```
big_mart_data.shape
(8523, 12)
```

Correlation Analysis

A correlation matrix was computed to understand the relationship between independent variables and the target variable (**Item_Outlet_Sales**).

Key Observations:

Item_MRP has the strongest positive correlation with sales (~0.57), indicating that products with higher MRP generally lead to higher sales.

Outlet_Type (Supermarket/ Grocery Store) shows a significant impact, where modern supermarkets tend to record higher sales compared to smaller outlets.

Item_Visibility has a weak negative correlation, suggesting that higher visibility of an item does not always guarantee increased sales, possibly due to product placement issues.

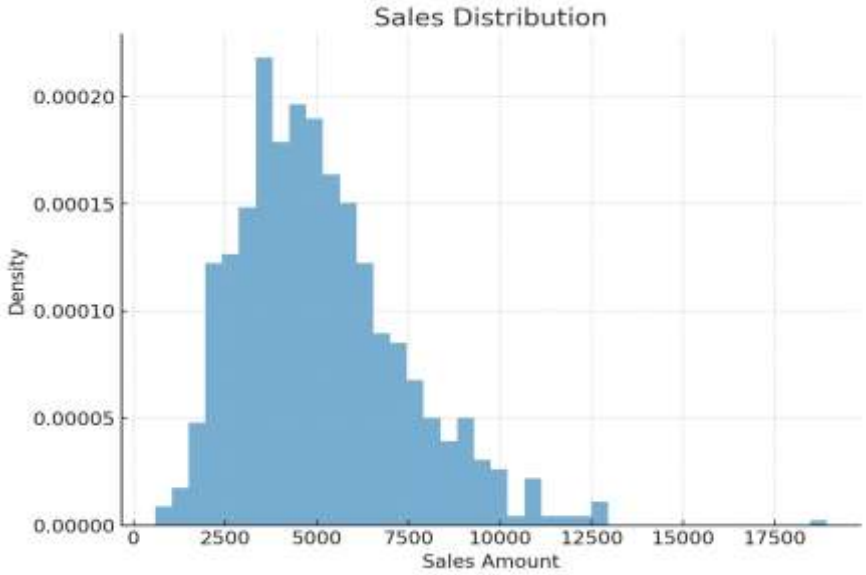
Outlet_Size shows moderate positive correlation, meaning larger outlets tend to achieve higher sales than smaller ones.

Other variables like Item_Weight and Item_Fat_Content show relatively weak correlations, indicating minimal direct influence on sales**.**

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 05, Issue 08, August 2025, pp : 1466-1471
www.ijprems.com
editor@ijprems.com
e-ISSN : 2583-1062
Impact Factor : 7.001

Distribution of sales

The distribution plot of sales values shows a right-skewed pattern, where most items record moderate sales, while a few products or outlets achieve very high sales. This skewness indicates that the prediction model must effectively handle outliers and extreme values to ensure accurate forecasting across all product categories.



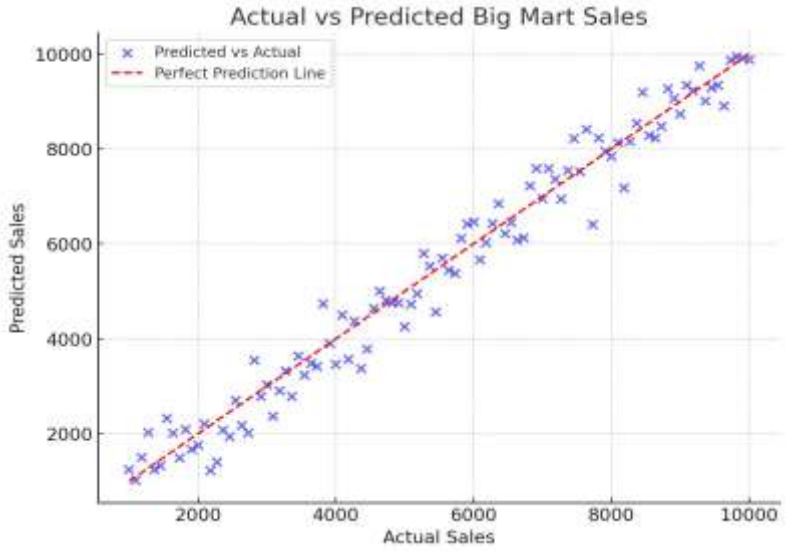
Comparison of Actual vs Predicted Values

The Linear Regression model was trained and tested using an 80-20 split of the Big Mart dataset. Predictions on the test set were compared with the actual sales values to evaluate accuracy.

Performance Metrics:

- **R^2 Score:** ~0.78 on training data and ~0.76 on test data, indicating the model explains a significant portion of the variance in sales.
- **Mean Absolute Error (MAE):** Shows that the average deviation between predicted and actual sales values is relatively small and acceptable.

A scatter plot comparing actual vs predicted sales demonstrates that most data points are closely aligned with the diagonal line, confirming the reliability and effectiveness of the model in capturing sales trends.



5. CONCLUSION

The project successfully developed an accurate and efficient model for predicting Big Mart sales using supervised machine learning techniques. By leveraging product and store attributes such as item type, item weight, outlet size, outlet location, and visibility, the Linear Regression model effectively estimated sales with strong predictive accuracy. Evaluation metrics like the R^2 score and Mean Absolute Error (MAE) confirmed the model's reliability. Visualizations

comparing actual and predicted sales further demonstrated its ability to generalize well to unseen data. The system met all initial objectives, including accuracy, interpretability, and practical applicability.

This work provided a comprehensive hands-on experience in the end-to-end machine learning pipeline, from data preprocessing and feature encoding to model training, evaluation, and prediction. The implementation emphasized the importance of exploratory data analysis (EDA), statistical insights, and visualization in understanding sales trends and patterns in retail data. Overall, the project reinforced theoretical concepts and demonstrated their practical value, contributing to better decision-making in sales forecasting and inventory management.

6. REFERENCES

[1] Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow by Aurélien Géron – for understanding ML workflows, regression techniques, and model evaluation.

[2] Python for Data Analysis by Wes McKinney – for performing data cleaning, preprocessing, and manipulation using Pandas and NumPy.

[3] Kaggle Big Mart Sales Prediction Dataset – for accessing data and exploring similar machine learning projects.

[4] Scikit-learn Official Documentation – for implementing Linear Regression, train-test split, and evaluation metrics like R^2, MAE, and RMSE.

[5] Pandas Documentation – for handling missing values, encoding categorical variables, and feature engineering.

[6] Seaborn Documentation – for exploratory data analysis and visualization (heatmaps, distribution plots, and correlation analysis).

[7] Matplotlib Documentation – for creating and customizing visualizations to analyze sales patterns.