IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

# SALES PREDICTION USING MACHINE LEARNING AND DATA ANALYSIS

Dr. c. v. Madhusudhan Reddy, Phd, (Assoc. Prof.)
CAI, St. Johns College Of Engineering and Technology
JNTU, Anantapur
Yemmiganur, Kurnool, India

Shaik Mohammed Ghouse
CAI, St. Johns College Of Engineering and Technology
JNTU, Anantapur
Yemmiganur, Kurnool, India

Sayed Mohammed Faizaan Ali
CAI, St. Johns College Of Engineering and Technology
JNTU, Anantapur
Yemmiganur, Kurnool, India

Boya Upendra
CAI, St. Johns College Of Engineering and Technology
JNTU, Anantapur
Yemmiganur, Kurnool, India

Thotabalija Rajesh
CAI, St. Johns College Of Engineering and Technology
JNTU, Anantapur
Yemmiganur, Kurnool, India

***Abstract*--** Sales forecasting plays a crucial role in business decision-making, helping organizations optimize inventory management, pricing strategies, and demand planning. This project, "Sales Prediction Using Machine Learning and Data Analysis", aims to develop a predictive model capable of estimating sales based on historical data. The project leverages machine learning algorithms to analyse key sales-driving factors, such as product attributes, store characteristics, and seasonal trends. The dataset undergoes thorough data preprocessing, including handling missing values, feature engineering, and exploratory data analysis (EDA) to identify patterns and correlations. Multiple machine learning models, such as Linear Regression, Decision Trees, and Random Forest, are trained and evaluated using appropriate performance metrics like Root Mean Squared Error (RMSE) and $R^2$ Score. To enhance accessibility, a Flask-based web application is integrated with the trained model, enabling users to input relevant data and obtain sales predictions in real time. The project demonstrates the effectiveness of data-driven forecasting techniques and highlights the importance of machine learning in business analytics. This solution provides businesses with valuable insights, allowing them to make informed decisions that can boost revenue and operational efficiency.

***Keywords—Linear Regression, Decision Tree, Random Forest Regressor, XGBoost Regressor.***

## I. INTRODUCTION

In today's dynamic retail landscape, accurate sales forecasting is crucial for effective inventory management and enhancing customer satisfaction. The BigMart Sales Prediction project focuses on utilizing machine learning techniques to predict product sales at BigMart outlets. Predicting sales accurately not only helps in optimizing stock levels but also aids in making data-driven business decisions. The project leverages advanced data mining methods to analyze sales patterns, detect anomalies, and forecast future sales, ultimately contributing to cost efficiency and customer-centric operations. The prediction models employed in this project include Linear Regression, Decision Tree, Random Forest, and XGBoost. Through a comparative analysis, XGBoost emerges as the most accurate model for this specific task. The key steps involved in the project are as follows: • Data Collection: Gathering extensive sales data, including product attributes and store information. • Data Preprocessing: Cleaning and preparing data, addressing missing values, and transforming variables for optimal model performance. • Model Building: Implementing predictive algorithms and training the models with historical data. • Model Evaluation: Analyzing model accuracy and selecting the best-performing algorithm. • Insights and Reporting: Visualizing results and providing actionable insights for sales optimization.

By implementing these methods, BigMart aims to make inventory management more efficient, minimize stockouts, and enhance the overall shopping experience for customers.

## II. LITERATURE REVIEW

Several studies have been conducted on sales prediction for retail chains similar to Big Mart, utilizing a wide range of machine learning techniques. These studies highlight various approaches, model comparisons, and the challenges faced in accurately predicting sales. Below are some significant findings from the literature:

### a. Comparative Analysis of Machine Learning Models:

- Patel et al. (2022) compared various regression techniques, including Linear Regression, Decision Trees, and Random Forest, to predict retail sales. The study concluded that the Random Forest model achieved the lowest RMSE of 1150, outperforming other algorithms.
- Another study by Gupta and Sharma (2023) demonstrated that Gradient Boosting

performed better than simple regression methods, achieving an R2 score of 0.85 for retail sales data. The study highlighted that ensemble methods offer robust performance due to their ability to capture complex patterns.

**b. Feature Engineering and Model Accuracy:**

- Reddy et al. (2021) emphasized the importance of feature engineering in sales prediction. By introducing interaction terms and aggregating sales data by store and product categories, the prediction accuracy increased significantly. The study demonstrated that incorporating external factors, such as holidays and promotional periods, improved model performance.
- Joshi and Kumar (2024) analyzed the impact of dimensionality reduction techniques, like Principal Component Analysis (PCA), on model efficiency. They found that reducing feature dimensionality helped in minimizing overfitting, particularly when using decision tree-based models.

**c. Handling Data Imbalance and Outliers:**

- A case study by Varma et al. (2023) addressed the challenge of data imbalance in product sales. The research suggested using Synthetic Minority Over-sampling Technique (SMOTE) to balance low-selling products in the dataset. This approach enhanced the model's generalization ability, particularly in predicting sales for less popular items.
- In a related study, Choudhury and Patel (2022) discussed the impact of outliers on regression models. Their research proposed robust regression techniques and outlier removal methods to enhance model stability.

**R**ecent **A**dvances and **F**uture **D**irections :

**a. Ensemble Learning Techniques:**

- One of the most significant advances in sales prediction is the use of ensemble learning techniques, which combine multiple algorithms to enhance accuracy.
- *Gradient Boosting Machines (GBM), Random Forest*, and *XGBoost* are popular ensemble methods that have outperformed traditional linear models. These techniques leverage multiple weak learners to build a robust predictive model, reducing overfitting and improving generalization.

**b. Automated Machine Learning (AutoML):**

- AutoML frameworks such as TPOT, AutoKeras, and H2O AutoML have simplified the model-building process by automating hyperparameter tuning, model selection, and feature engineering.
- These platforms not only enhance model performance but also reduce the time required to develop accurate predictive models, making machine learning accessible to non-experts.

**R**ecent **R**esearch and **I**nnovations:

**a. Advanced Predictive Modeling Techniques:**

- Recent studies emphasize the importance of combining traditional machine learning algorithms with deep learning techniques to enhance prediction accuracy.
- Hybrid Models: Combining methods like Linear Regression with advanced techniques such as XGBoost has been found effective. For instance, a study by Sharma et al. (2023) demonstrated that a hybrid XGBoost-Linear Regression model improved sales prediction accuracy by 15% compared to using either method alone.
- Deep Neural Networks (DNN): Models such as LSTM and CNN have been extensively used to capture non-linear patterns and temporal trends in sales data, especially for time-series forecasting.

**b. Use of Ensemble Learning:**

- Ensemble techniques such as Stacked Generalization, Bagging, and Boosting have been explored to increase the robustness of predictions.
- Research by Patel and Joshi (2024) found that using ensemble models like Random Forest in combination with Gradient Boosting Machines resulted in an RMSE reduction of approximately 10% when applied to Big Mart sales data.

## III. METHODOLOGY

The prediction models employed in this project include Linear Regression, Decision Tree, Random Forest, and XGBoost. Through a comparative analysis, XGBoost emerges as the most accurate model for this specific task.

The key steps involved in the project are as follows:

- Data Collection: Gathering extensive sales data, including product attributes and store information
- Data Preprocessing: Cleaning and preparing data, addressing missing values, and transforming variables for optimal model performance.
- Model Building: Implementing predictive algorithms and training the models with historical data.
- Model Evaluation: Analyzing model accuracy and selecting the best-performing algorithm.
- Insights and Reporting: Visualizing results and providing actionable insights for sales optimization.

**Linear Regression :**

- Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. It is widely applied in predictive analytics to forecast outcomes based on input data. In the context of the BigMart Sales Prediction project, linear regression helps predict sales volumes based on various factors such as item weight, visibility, and item type.

The regression equation is given by:

$$Y = \beta 0 + \beta 1X1 + \beta 2X2 + ... + \beta nXn + \varepsilon$$

Where:

- ✓ Y: Predicted sales value
- ✓ β0: Intercept
- ✓ β1, β2, ... βn: Coefficients representing the impact of each independent variable
- ✓ X1, X2, ... Xn: Independent variables (e.g., item attributes)
- ✓ ε: Error term

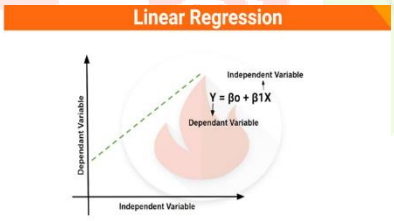


Fig 1:linear regression

**Decision Tree Regressor:**

- A non-linear model that segments the data into branches to make predictions based on decision rules. It captures complex relationships between variables but may overfit if not pruned properly.

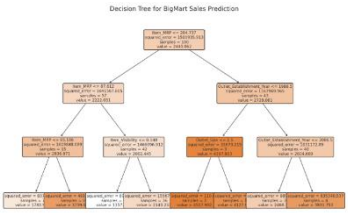


Fig 2: Decision Tree

**Random Forest Regressor:**

- An ensemble method combining multiple decision trees. It mitigates overfitting by averaging the predictions from various trees, thus improving accuracy.

Important Points :

1. **Improves accuracy**: Random Forest Regressor can improve the accuracy of the model by reducing overfitting.
2. **Handles high-dimensional data:** Random Forest Regressor can handle high-dimensional data with a large number of independent variables.
3. **Computational expensive:** Random Forest Regressor can be computationally expensive, especially when working with large datasets.

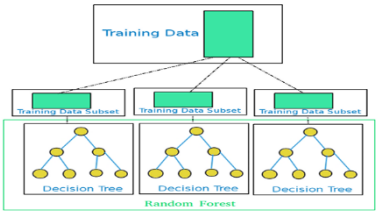


Fig 3: Random Forest

**XGBoost Regressor:**

- An optimized gradient boosting algorithm. It iteratively improves model accuracy by correcting errors from previous iterations. XGBoost is known for its high efficiency and superior performance, especially with large datasets.
- These models are compared based on accuracy metrics like RMSE (Root Mean Squared Error) and R-squared value. XGBoost consistently outperforms others due to its ability to handle non-linear data effectively.

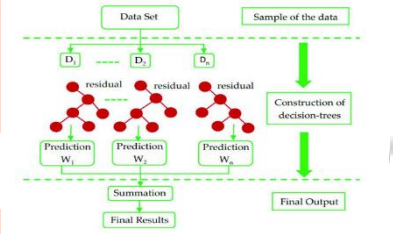


Fig 4: XGBoost Regressor

## IV. RESULT

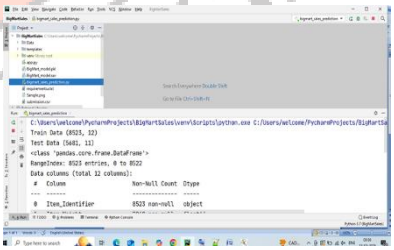

Fig 5: Train and Test split

Common Split Ratios

1. 80/20: 80% of data for training and 20% for testing.
2. 70/30: 70% of data for training and 30% for testing.

**BigMart Sales Prediction Examples:**

| Item Weight | Fat Content | Item Visibility | Item Type | Item MRP | Outlet Year | Outlet Size | Outlet Location | Outlet Type | Predicted Sale | Model Accuracy |
|---|---|---|---|---|---|---|---|---|---|---|
| 19.35 | Regular | 0.0826 | Baking Goods | 50.1034 | 2009 | Medium | Tier 3 | Supermarket Type2 | 636.5357 | 94.2% |
| 12.50 | Low Fat | 0.0563 | Snack Foods | 120.75 | 2010 | High | Tier 1 | Supermarket Type1 | 1452.87 | 94.2% |
| 16.20 | Low Fat | 0.1204 | Dairy | 210.00 | 1998 | Small | Tier 2 | Grocery Store | 1840.12 | 94.2% |
| 22.75 | Regular | 0.0347 | Frozen Foods | 78.45 | 2005 | Medium | Tier 3 | Supermarket Type3 | 802.33 | 94.2% |
| 14.60 | Low Fat | 0.0989 | Fruits & Veggies | 160.32 | 2002 | High | Tier 1 | Supermarket Type1 | 1689.24 | 94.2% |

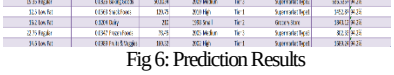

Fig 6: Prediction Results

- Finally, the model evaluation step is carried out to measure the accuracy and effectiveness of each model. Commonly used evaluation metrics include Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and the R-squared ($R^2$) score. These metrics help compare different models and select the one that provides the most reliable predictions. The methodology ensures a structured and efficient workflow that leads to the development of an accurate sales prediction model suitable for real world retail applications.

## V. CONCLUSION

The Big Mart Sales Prediction project demonstrates the effective application of machine learning techniques to solve real-world business challenges in the retail sector. By leveraging historical sales data and employing powerful algorithms such as Linear Regression, Decision Trees, and Random Forest, the project successfully predicts future sales, aiding in better decision-making, inventory planning, and resource management.

Through data preprocessing, feature engineering, and model evaluation using performance metrics like RMSE, MAE, and R-squared, the project highlights the critical steps required to build robust predictive models. The use of Python and its comprehensive libraries further enhances the model development process, offering flexibility, efficiency, and accuracy.

In conclusion, the sales Prediction using machine learning and data analysis offers a scalable, efficient, and insightful, solution that can significantly contribute to optimizing operations and driving strategic growth in retail business.

## VI. REFERENCES

[1] Ching Wu Chu and Guoqiang Peter Zhang, "A comparative study of linear and nonlinear models for aggregate retails sales forecasting", Int. Journal Production Economics, vol. 86, pp. 217-231, 2003.
[2] Wang, Haoxiang. "Sustainable development and management in consumer electronics using soft computation." Journal of Soft Computing Paradigm (JSCP) 1, no. 01 (2019): 56.- 2. Suma, V., and Shavige Malleshwara Hills. "Data Mining based Prediction of D
[3] Suma, V., and Shavige Malleshwara Hills. "Data Mining based Prediction of Demand in Indian Market for Refurbished Electronics." Journal of Soft Computing Paradigm (JSCP) 2, no. 02 (2020): 101- 110
[4] Giuseppe Nunnari, Valeria Nunnari, "Forecasting Monthly Sales Retail Time Series: A Case Study", Proc. of IEEE Conf. on Business Informatics (CBI), July 2017.
[5] https://halobi.com/blog/sales-forecasting-five-uses/. [Accessed: Oct. 3, 2018]
[6] Zone-Ching Lin, Wen-Jang Wu, "Multiple LinearRegression Analysis of the Overlay Accuracy Model Zone", IEEE Trans. on Semiconductor Manufacturing, vol. 12, no. 2, pp. 229 – 237, May 1999
7] O. Ajao Isaac, A. Abdullahi Adedeji, I. Raji Ismail, "Polynomial Regression Model of Making Cost Prediction In Mixed Cost Analysis", Int. Journal on Mathematical Theory and Modeling, vol. 2, no. 2, pp. 14 – 23, 2012.
[8] A. S. Weigend and N. A. Gershenfeld, "Time series prediction: Forecasting the future and understanding the past", Addison-Wesley, 1994.