Big Mart Sales forecasting using Machine Learning Model in Python

Master Degree in Artificial Intelligence

Course: Machine Learning

Academic Year: 2024–2025

Author Name: Gebre Haftom Desbele
Matriculation number:VR526525

Abstract

This project presents a machine learning-based approach to forecast product sales using the Big Mart sales dataset. Multiple regression algorithms were implemented, including Linear Regression, Polynomial Regression, Random Forest Regressor, Gradient Boosting Regressor, and XGBoost Regressor. The models were evaluated using R^2 score and Mean Absolute Error (MAE). Among all, the Polynomial Regression with degree 2 achieved the highest performance with an R^2 score of 0.6157 and an MAE of 0.4248, indicating better predictive accuracy. Despite challenges such as data variability and missing values, the results demonstrate the effectiveness of machine learning in retail sales forecasting. The model can support data-driven decisions in inventory and sales planning across Big Mart outlets.

Contents

List of Tables

List of Figures

Chapter 1

Motivation and Rationale

Introduction

In today's highly competitive and constantly changing business environment, the accurate and timely estimation of future sales, also known as sales prediction or sales forecasting, can offer critical knowledge to companies involved in wholesale,manufacturing or retial of products. Short term peridctions mainly help in production planning and stock management, while long term can help business development decision making[3].

Large shopping malls and massive retail outlets are increasingly common. Accurate tracking of the exchange of products, each with various dependent and independent features, is essential to forecast future demand and effectively manage inventory.

The current machine learning algorithm is very advanced and provides methods for Predicting or forecasting sales any kind of organization, extremely beneficial to overcome low–priced used for prediction. Always better predictions is helpful, both in developing and improving marketing strategies for the marketplace, which is also particularly helpful. [9]This project utilizes Big Mart, a one-stop retail center, to predict sales using machine learning.

1.1 Motivation

Sales forecasting is typically done arbitrarily by managers.Howver skilled managers are hard to find they are hard to find and they are not always available. Therefore, sales forecasting should be supported by computer systems that can play the role of skilled managers when she/he is not there and/or helt her take the right decision by providing estimates for future sales.One way to build such system would be to try and model the expert knowledge if skilled managers within computer system.Alternatively, one can exploit the wealth of sales data and related information to automaticaly constuct sales prediction models via machine learning techniques.The machine learning tecquince is much simpler process, it is not biased form particluaties of specific sales manager and it is dynamic, meaning it can adapt to changes in the data. Furthermore, it has the potential to outweigh the prediction accuracy of a human expert, who typically is imperfect[12].

1.2 Problem Statement

The data scientists at BigMart have collected sales data for 1559 products across 10 stores in different cities for the year 2013. Now each product has certain attributes that sets it apart from other products.

- **Type of Problem:** Supervised Machine Learning.
- **Target Variable:** `Item_Outlet_Sales`.

1.3 Objective

The purpose of this project is on "Big Mart Sales Forecasting Using Machine Learning" is to develop *accurate forecasting models for sales in retail environments, specifically focusing on Big Mart supermarkets.*This issue is of critical crucial for inventory management, transportation logistics, and overall business planning[9].Accurate forecasts help reduce overstock and stockouts, optimize supply chains, and ultimately improve customer satisfaction and profitability.

1.3.1 Specific Objectives

- To develop a predictive model for estimating future sales volumes using a variety of machine learning algorithms, including Linear Regression, Polynomial Regression, Ridge Regression, Decision Tree Regressor, Random Forest Regressor, and XGBoost Regressor.
- To compare the performance of different regression techniques and identify the most suitable model for accurate sales forecasting.

The project aims to effectively predict future sales volumes by applying a range of machine learning algorithms, including Linear Regression, Polynomial Regression, Ridge Regression, Decision Tree Regressor, Random Forest Regressor, and XGBoost Regressor. Explores these regression techniques to identify the most suitable model for accurate sales forecasting. Key features considered in the prediction include item weight, item visibility, item type, item pricing, and various outlet characteristics. By analyzing these factors, the project seeks to understand their influence on sales and enhance forecasting accuracy in the retail sector.Ultimately, the goal is to enhance the decision-making process for retailers like Big Mart, enabling them to optimize stock levels, streamline operations, and improve overall business efficiency.

1.4 Scope

The scope of this project, *Big Mart Sales Forecasting Using Machine Learning*, lies in building and evaluating predictive models that can forecast sales for retail chains like Big Mart. Using the publicly available dataset from Kaggle (https://www.kaggle.com/datasets/shivan118/big-mart-sales-prediction-datasets), this project focuses on developing machine learning models that analyze key features such as item weight, item type, visibility, MRP, and outlet characteristics.

The scope is limited to the features and records provided in the Kaggle dataset, which includes historical sales data across multiple outlets. However, this dataset still offers a strong foundation for experimenting with various regression techniques and understanding the factors influencing sales in a retail setting. The insights gained through this study can be generalized to similar retail environments, paving the way for more refined and scalable predictive solutions in the future.

1.5 Significance of the Project

This project aims to enhance the decision-making process for retailers such as Big Mart by providing an accurate sales forecasting tool. By understanding how various product and outlet features affect sales, the model enables retailers to:

- Optimize stock levels to prevent overstocking or understocking.
- Streamline operational planning based on more reliable sales estimates.
- Improve overall business efficiency by adopting a data-driven strategy.

The integration of machine learning into retail forecasting supports more informed and strategic business decisions, contributing to better resource management and customer satisfaction.

1.6 Paper Organization

The rest part of the paper is organized as follows. Section 2 presents the state of the art of the existing technologies, Section 3 introduces the Methodology includes proposed methodology,dataset and models. Section 4 introduces metrics, experimental setup, and analyzes the results. Eventually, Section 5, conclusion, and final recommendation.

Chapter 2

State of the Art

Sales forecasting has become increasingly important in the retail and Big Mart sectors, enabling more accurate planning, resource allocation, and reduction of financial loss. The state of the art reflects a shift from traditional statistical techniques to a more advanced application of machine learning (ML), including ensemble models and deep learning architectures.

2.1 Traditional and Statistical Models

Early approaches to sales prediction relied heavily on statistical models such as ARIMA and exponential smoothing. These methods, although effective in capturing trends and seasonality, are often limited in capturing complex nonlinear relationships and incorporating external variables like weather or promotions. Some studies have proposed hybrid models that combine ARIMA with neural networks to overcome these limitations. [13].

2.2 Machine Learning Models

a. Linear Regression

Linear regression remains a baseline for comparison due to its simplicity and interpretability. However, in small datasets or those with a majority of categorical features, such as in the Big Mart dataset, linear regression suffers from underfitting and poor predictive performance.[6, 2]

b. Random Forest and Gradient Boosting

Ensemble learning methods such as Random Forest and Gradient Boosting significantly outperform linear models in predictive tasks. These models are capable of capturing nonlinear interactions among features and effectively handle both categorical and numerical variables.

In the Big Mart sales prediction problem, Gradient Boosting slightly outperformed Random Forest across key evaluation metrics such as RMSE, MAE, and R^2, while both models substantially outperformed Linear Regression [1]. Random Forest also demonstrated strong performance in predicting house and book sales when spatial and sentiment-based features were incorporated into the dataset [1].

c. XGBoost and Its Variants

XGBoost, an optimized implementation of gradient-boosted decision trees, is widely recognized for its efficiency and high predictive accuracy in sales forecasting applications. In a restaurant sales study, XGBoost surpassed traditional uplift models by 10–15 percentage points in prediction accuracy [4]. Its ability to handle missing data, perform model regularization, and exploit parallel computation makes it a preferred model across many domains.

2.3 Deep Learning Models

a. LSTM (Long Short-Term Memory)

Long Short-Term Memory (LSTM) networks, a type of Recurrent Neural Network (RNN), are especially effective in modeling time-dependent sequences such as daily or weekly sales. Holmberg and Halldén (2018) applied LSTM to restaurant sales prediction and found that it consistently outperformed heuristic-based methods, particularly when external variables like weather data were incorporated [4].

In another study on Japanese supermarket sales, LSTM networks combined with stacked denoising autoencoders achieved better predictive accuracy than traditional machine learning models [11, 7].

b. Hybrid Architectures

Recent approaches increasingly combine deep learning models for feature extraction (e.g., autoencoders) with temporal sequence models (e.g., LSTM) to improve representation learning. These hybrid architectures leverage the strengths of both deep learning and traditional time-series modeling techniques [7].

Chapter 3

Methodology

3.1 Flow Diagram of the Proposed System

The proposed system follows a structured pipeline to perform predictive modeling on the BigMart dataset. The system is composed of several sequential steps, as illustrated in Figure 3.1.

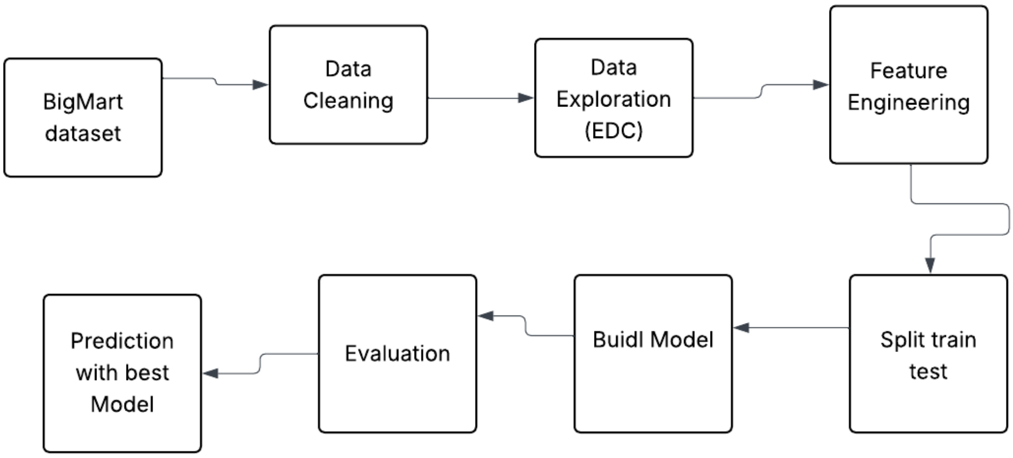


Figure 3.1: Data Flow Diagram of the Proposed System

The flow of the system is summarized as follows:

1. **BigMart Dataset:** The process begins with the collection of raw sales data from the BigMart dataset.

2. **Data Cleaning:** This step handles missing values, incorrect data types, and outlier removal to ensure data quality.

3. **Data Exploration (EDC):** Exploratory Data Analysis is performed to understand the underlying structure and relationships in the data.

4. **Feature Engineering:** Important features are created or transformed to improve model performance.

5. **Split Train/Test:** The dataset is divided into training and testing sets to evaluate the generalization of the models.

6. **Build Model:** Various machine learning models are trained using the training data.

7. **Evaluation:** The trained models are evaluated using performance metrics to determine their effectiveness.

8. **Prediction with Best Model:** The best-performing model is selected to make final predictions on unseen data.

This workflow ensures a systematic approach to solving the predictive task with optimal model selection and reliable results.

3.2 Dataset

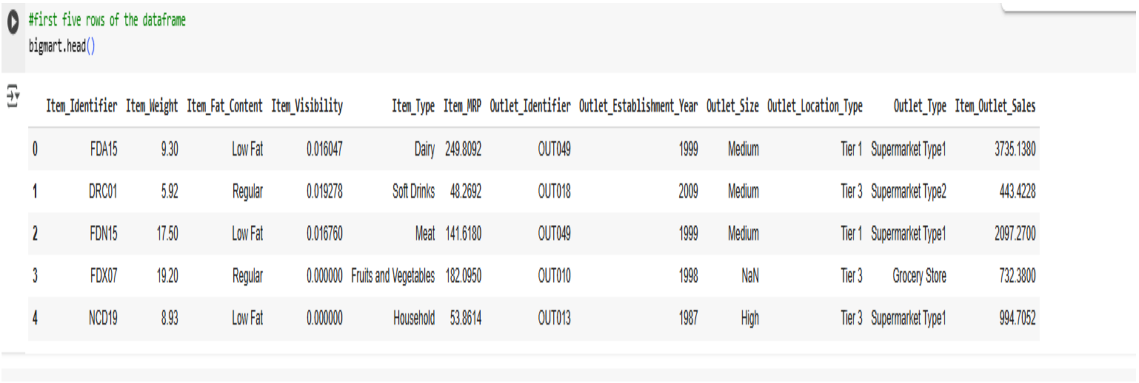
The dataset used in this project was obtained from *Kaggle* and contains sales data related to *Big Mart*, a supermarket brand with more than 60 stores operating across different locations. It includes detailed sales records for 1,559 products sold across 10 stores in various areas during the year 2013. The dataset provides information about each product, the outlet where it was sold, and the corresponding sales value.

The training set comprised 8,523 observations and was used to build and test the predictive model. The objective was to forecast the sales of each product at a given outlet based on product and outlet.

Table 3.1: Attributes Information of the Big Mart Sales Dataset. Adapted from the study [9] on predictive analysis using machine learning algorithms.

Attribute	**Description**
Item_Identifier	Unique product ID number.
Item_Weight	Weight of the product.
Item_Fat_Content	Indicates whether the item is low in fat or not.
Item_Visibility	Percentage of the overall viewing area assigned to the item in the store.
Item_Type	The category or group to which the product belongs.
Item_MRP	Maximum Retail Price of the product.
Outlet_Identifier	Unique ID for each outlet/store.
Outlet_Establishment_Year	The year when the outlet first opened.
Outlet_Size	Total area occupied by the supermarket.
Outlet_Type	Type of outlet (e.g., supermarket, grocery store).
Item_Outlet_Sales	Sales of the item at the particular outlet (target variable).

In the context of predicting Big Mart sales using machine learning, model selection plays a crucial role in determining the accuracy and effectiveness of the predictive system. Various algorithms such as Linear Regression, Polynomial Regression, Ridge Regression, and XGBoost Regression are compared based on their performance metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE).



```
#first five rows of the dataframe
bigmart.head()
```

	Item_Identifier	Item_Weight	Item_Fat_Content	Item_Visibility	Item_Type	Item_MRP	Outlet_Identifier	Outlet_Establishment_Year	Outlet_Size	Outlet_Location_Type	Outlet_Type	Item_Outlet_Sales
0	FDA15	9.30	Low Fat	0.016047	Dairy	249.8092	OUT049	1999	Medium	Tier 1	Supermarket Type1	3735.1380
1	DRC01	5.92	Regular	0.019278	Soft Drinks	48.2692	OUT018	2009	Medium	Tier 3	Supermarket Type2	443.4228
2	FDN15	17.50	Low Fat	0.016760	Meat	141.6180	OUT049	1999	Medium	Tier 1	Supermarket Type1	2097.2700
3	FDX07	19.20	Regular	0.000000	Fruits and Vegetables	182.0950	OUT010	1998	NaN	Tier 3	Grocery Store	732.3800
4	NCD19	8.93	Low Fat	0.000000	Household	53.8614	OUT013	1987	High	Tier 3	Supermarket Type1	994.7052

Figure 3.2: Sample data visualization from the Big Mart Sales dataset.

3.3 Data Cleaning

Data cleaning is the process of improving our dataset by refining the data [10]. We have two essential attributes in our dataset that contain missing values and need to be filled. The codes used give a comparison of missing values before and after. For example, `Item_Weight` has 1460 missing values before and zeros after implementing the above code. Another attribute with missing values, `Outlet_Size`, is handled using the mode in `aggfunc`. Table 3.2 shows the imputation of missing `Outlet_Size` using the mode based on each `Outlet_Type`.Now our data is ready for feature engineering.

Column Name	**Missing Values**
Item_Identifier	0
Item_Weight	1463
Item_Fat_Content	0
Item_Visibility	0
Item_Type	0
Item_MRP	0
Outlet_Identifier	0
Outlet_Establishment_Year	0
Outlet_Size	2410
Outlet_Location_Type	0
Outlet_Type	0
Item_Outlet_Sales	0

Table 3.2: Missing values in Big Mart dataset columns

Outlet_Type	Grocery Store	Supermarket Type1	Supermarket Type2
Outlet_Size (Mode)	Small	Small	Medium

Table 3.3: Imputation of missing `Outlet_Size` using mode per `Outlet_Type`

3.4 Feature Engineering

Feature engineering helps us understand the data for better analysis. Here, we will create some new variables from the original data.[10] After data cleaning and feature engineering, we were going to build our predictive model. This model plays a vital role in predicting sales.

3.5 Models

3.5.1 Linear Regression

Linear Regression is one of the most popular and widely used machine learning algorithms. It is primarily used to model the relationship between a dependent (target) variable and one or more independent (predictor) variables.[5]

The main objective of linear regression is to determine the optimal linear relationship that best fits the data. This involves estimating the coefficients of the independent variables such that the predicted values are as close as possible to the actual observed values.

Mathematically, the linear regression model can be expressed as:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

where:

- $\hat{y}$ is the predicted (dependent) value,
- β_0 is the intercept,
- $\beta_1, \beta_2, \ldots, \beta_n$ are the coefficients (weights),
- $x_1, x_2, \ldots, x_n$ are the independent variables.

The best-fitting line minimizes the sum of squared errors (residuals), where the error is defined as the difference between the observed value y_i and the predicted value $\hat{y}_i$:

$$\text{Error (residual)} = y_i - \hat{y}_i$$

The optimal line, often referred to as the line of best fit, is determined using the *least squares* method, which minimizes the total squared differences between the observed and predicted values:

$$\min \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

This approach ensures that the model achieves the lowest possible error and represents the underlying data distribution as accurately as possible.

3.5.2 Polynomial Regression Algorithm

Polynomial Regression is a regression algorithm that models the relationship between a dependent variable y and an independent variable x as an n^{th}-degree polynomial. It extends simple linear regression by introducing additional polynomial terms to capture non-linear relationships.

The general form of the polynomial regression equation is:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \cdots + \beta_n x^n + \epsilon$$

where:

- y is the dependent variable,
- x is the independent variable,
- $\beta_0, \beta_1, \ldots, \beta_n$ are the coefficients,
- ϵ is the error term.

Polynomial regression is often considered a special case of multiple linear regression. This is because by transforming the original input x into polynomial features (such as $x^2, x^3, \ldots$), the problem can still be solved using a linear model.

- Polynomial regression is suitable for datasets with a non-linear relationship between the variables.
- Although it models non-linear data, the model is linear in the coefficients, and thus it can be solved using standard linear regression techniques.
- It enhances model accuracy by fitting more flexible curves to the data, especially when linear regression fails to capture complex patterns.

This method is particularly useful when a simple linear model underfits the data. However, care must be taken to avoid overfitting by choosing an appropriate polynomial degree.

3.5.3 Random Forest Regressor

In the domain of Big Mart sales prediction using machine learning, the **Random Forest Regressor** emerges as a potent tool. By leveraging the ensemble learning technique, it aggregates the predictions of multiple decision trees to capture complex, non-linear relationships within the dataset. This model is highly effective in dealing with both numerical and categorical data.[8]

Its inherent ability to perform feature selection enables the identification of significant predictors such as item weight, item visibility, and outlet attributes. Moreover, its ensemble nature mitigates the problem of overfitting, which is often encountered in decision tree models, thus improving generalization to unseen data.

The Random Forest Regressor operates by training multiple trees on bootstrapped samples of the dataset and averaging their predictions for regression tasks.

3.5.4 XGBRF Regressor

The **XGBRF Regressor** (Extreme Gradient Boosting Random Forest) represents a hybrid model that combines the strengths of Random Forests with the accuracy enhancements provided by gradient boosting frameworks. Within the context of Big Mart sales prediction, this model offers significant advantages in predictive performance and computational efficiency.

XGBRF enhances traditional random forests by incorporating gradient-based optimization, which allows for iterative improvement of the model during training. Unlike traditional boosting methods that build trees sequentially to minimize loss, XGBRF applies random sampling and regularization, blending both boosting and bagging principles

Chapter 4

Experiments and Results

Evaluation Metrics

Mean Absolute Error (MAE):

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

Root Mean Squared Error (RMSE):

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

Coefficient of Determination ($\mathbf{R^2}$):

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

Result

The table below shows the performance of the proposed model.

Model	MAE	RMSE	R^2
Linear Regression	0.4835	0.6481	0.5501
Polynomial Regression	0.4248	0.5989	0.6157
Random Forest	0.4815	0.6926	0.5475
Gradient Boosting	0.4637	0.6642	0.5839
XGBoost	0.5050	0.7228	0.5072

Table 4.1: Performance comparison of regression models on MAE, RMSE, and R^2.

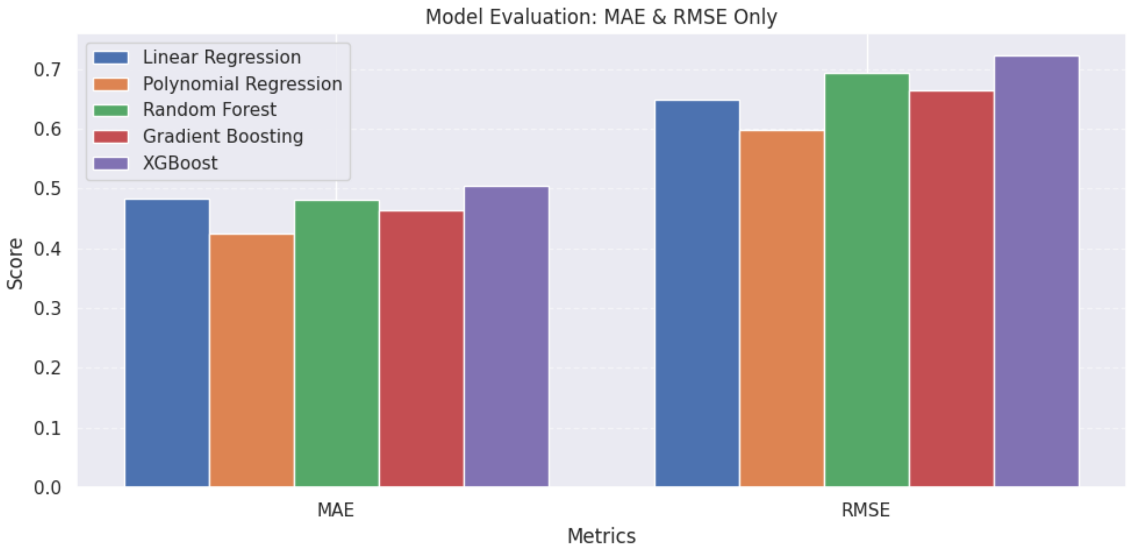


Figure 4.1: Model comparison based on Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE).

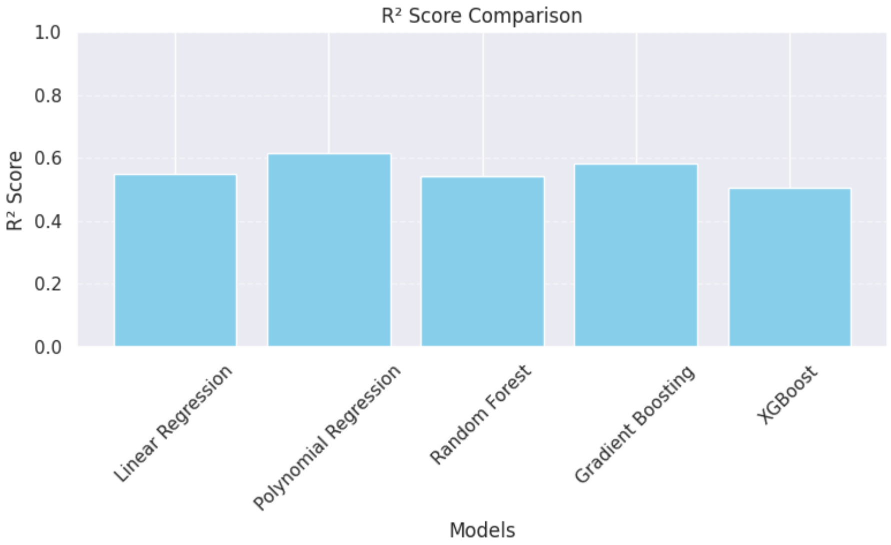


Figure 4.2: Comparison of models based on R^2 score.

Model Performance Analysis

Model Performance Summery

Figure 4.1 and Figure 4.2 summarize the predictive performance of five regression models: Linear Regression, Polynomial Regression, Random Forest, Gradient Boosting, and XGBoost.

Polynomial Regression outperformed all other models in terms of both MAE and RMSE, indicating more accurate and better predictions. It also achieved the highest R^2 score ($R^2 = 0.6157$), as shown in Figure 4.2.

While ensemble models like Random Forest and Gradient Boosting showed competitive results, they slightly underperformed compared to Polynomial Regression. XGBoost, although known for high performance, yielded the least favorable metrics in this specific task.

Polynomial Regression Analysis

According to the results, Polynomial Regression is the model. To evaluate the impact of model complexity on prediction performance, we tested Polynomial Regression with degrees 1, 2, and 3. Degree 1 corresponds to standard Linear Regression, while higher degrees introduce nonlinear terms to better capture complex patterns in the data. The evaluation revealed that the second-degree polynomial model achieved the best overall performance, with the lowest Mean Absolute Error (MAE =.4248), the lowest Root Mean Squared Error (RMSE = 0.5989), and the highest coefficient of determination ($R^2 = 0.6157$). Increasing the degree to 3 slightly worsened performance, indicating signs of overfitting. Therefore, a second-degree polynomial provided the optimal trade-off between model complexity and predictive accuracy for this dataset.

Chapter 5

Conclusion

Conclusion

In this project, we used different machine learning models to predict sales for Big Mart. We tested Linear Regression, Polynomial Regression, Random Forest, Gradient Boosting, and XGBoost. Among them, Polynomial Regression with degree 2 gave the best results with the lowest error and highest accuracy.

Ensemble models like Random Forest and Gradient Boosting also performed well, while Linear Regression had the weakest performance. The results show that machine learning can help retailers make better decisions by predicting future sales more accurately.

Recommendations

For future work, we recommend:

- Testing more advanced models like LSTM or hybrid deep learning methods.
- Including external features such as holidays, promotions, and weather.
- Deploying the best model in a real-time sales prediction system for Big Mart.

.

Bibliography

[1] Kaggle Community. *Big Mart Sales Prediction*. https://www.kaggle.com/competitions/big-mart-sales-prediction. 2020.

[2] Aurélien Géron. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*. 2nd ed. O'Reilly Media, 2019.

[3] Mikael Holmberg and Pontus Halldén. *Machine learning for restaurant sales forecast*. 2018.

[4] P. Holmberg and C. Halldén. "Sales Forecasting for Restaurants Using XGBoost and External Data". In: *Caspeco White Paper* (2018). Available at: https://caspeco.se/whitepapers/sales-prediction.

[5] Ashish Jain et al. "Forecasting Future Sales Using Linear Regression Approach". In: *2024 International Conference on Cybernation and Computation (CYBERCOM)*. IEEE. 2024, pp. 269–272.

[6] Kaggle Community. *Big Mart Sales Prediction*. https://www.kaggle.com/competitions/big-mart-sales-prediction. n.d.

[7] Z. Li and Y. Wang. "Hybrid Deep Learning Models for Sales Prediction Using Product and External Features". In: *International Journal of Forecasting* 36.1 (2020), pp. 45–60.

[8] Sagar Mondal, Anindita Debbarma, and B Prakash. "Big mart sales prediction using machine learning". In: *2024 10th International Conference on Communication and Signal Processing (ICCSP)*. IEEE. 2024, pp. 742–747.

[9] P Ranjitha and M Spandana. "Predictive analysis for big mart sales using machine learning algorithms". In: *2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS)*. IEEE. 2021, pp. 1416–1421.

[10] Faizan Ali Rao et al. "BMSP-ML: big mart sales prediction using different machine learning techniques". In: *IAES International Journal of Artificial Intelligence* 12.2 (2023), p. 874.

[11] K. Shin and K. Yamashita. "Deep Learning for Sales Forecasting in Japanese Retail". In: *Journal of Applied Data Science* 5.2 (2019), pp. 112–125.

[12] Grigorios Tsoumakas. "A survey of machine learning techniques for food sales prediction". In: *Artificial Intelligence Review* 52.1 (2019), pp. 441–447.

[13] G Peter Zhang. "Time series forecasting using a hybrid ARIMA and neural network model". In: *Neurocomputing* 50 (2003), pp. 159–175.