ISSN No:-2456-2165

Big Mart Sales Analysis

[1]P. Yagneshwar Sai
B. Tech. CSE
DY Patil International University

[2]Lavan Satish Vyas
B. Tech. CSE
DY Patil International University

[3]Sulaxan J
Dept. of CSE
DY Patil International University

Abstract:- In today's industry, shopping outlets and Big Mart have adopted an practice of tracking sales data for every product. This helps them predict customer demand and effectively manage their inventory. This paper explores the case of Big Mart focusing on predicting sales, for types of items and understanding the factors that influence these sales. By analyzing features from a dataset collected for Big Mart and employing modeling techniques such as Xgboost, Linear Regression, Gradient Boosting, AdaBoost and Random Forest accurate results are obtained. These findings can then be utilized to make decisions aimed at improving sales performance.

Keywords:- *Big Mart, XGBoost, AdaBoost, Linear Regression.*

I. INTRODUCTION

Sales play a role in the success of any business. The ability to accurately predict sales has an impact, on companies helping them maintain standards and improve their performance through effective strategies. Forecasting sales has always been an area of focus for marketing organizations[1]. To ensure efficiency it is essential for all suppliers to employ an optimal forecasting method than relying on manual handling, which can lead to errors and hinder organizational management in todays fast paced environment. The primary objective of businesses is to attract their target audience. Therefore, it is vital for companies to have a prediction model, in place that enables them to achieve this goal. Big Mart, a network of stores places importance on analyzing trends at both product and location levels with the help of data scientists who identify potential growth centers.

A. Problem Statement

The objective of this research is to develop regression models that can effectively forecast sales using data and determine the efficient algorithm for predicting sales, in the context of Big Mart.

II. RELATED WORK

To predict sales we used a linear method, a decision tree approach and a reliable gradient approach. The initial dataset we evaluated had entries. However for the analysis we narrowed it down significantly by removing data duplicate entries and irrelevant sales information. This study demonstrates that many vendors can benefit from predicting transaction rates. The insights gathered here could be valuable, in designing a system that can predict trends.[2]

This research focuses on analyzing findings. The valuable insights gained through data visualization. It utilized data mining techniques with the Gradient Boost method proving to be the most accurate, in predicting transactions.[3] The objective of this study is to provide insights for predicting a companys sales or requirements by utilizing methods like Clustering Models and sales forecasting metrics. The potential of approaches is Further explored in subsequent studies.[4] In this research an examination of data collected from a store is conducted, along with projecting store management strategies. The study also, Evaluates the impact of factors such as weather conditions, holidays and other events that can significantly influence different departments and their effect, on sales.[5]

III. METHODOLOGIES

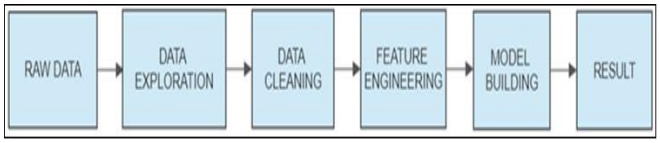

Fig 1 Methods used

A. Data Preprocessing

The Big Mart dataset was compiled by a group of data scientists who collected sales data for the year 2013. They gathered information on 1559 products sold across 10 stores situated in cities. Moreover they also identified the characteristics of each product and store. The objective is to create a model that can predict product sales, for every outlet. By using this model Big Mart aims to identify the factors that have an impact, on sales considering both product attributes and store features.6]

There are kinds of patterns that can be found in the data. These patterns can give us an understanding of the subject we're interested, in and provide insights, into the problem.

In order to make the data compatible, with a machine learning model and improve its efficiency it is necessary to preprocess the data by filling in the missing values. To fill the Item Weight values we used the weight of that specific item. Similarly for the missing Outlet Size values we filled them in using the size, for that particular type of outlet.

```
In [6]: df_train.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8523 entries, 0 to 8522
Data columns (total 12 columns):
 #   Column                     Non-Null Count  Dtype
---  ------                     --------------  -----
 0   Item_Identifier            8523 non-null   object
 1   Item_Weight                7060 non-null   float64
 2   Item_Fat_Content           8523 non-null   object
 3   Item_Visibility            8523 non-null   float64
 4   Item_Type                  8523 non-null   object
 5   Item_MRP                   8523 non-null   float64
 6   Outlet_Identifier          8523 non-null   object
 7   Outlet_Establishment_Year  8523 non-null   int64
 8   Outlet_Size                6113 non-null   object
 9   Outlet_Location_Type       8523 non-null   object
 10  Outlet_Type                8523 non-null   object
 11  Item_Outlet_Sales          8523 non-null   float64
dtypes: float64(4), int64(1), object(7)
memory usage: 799.2+ KB
```

Fig 2 Numerical Variables of the Dataset

➢ *The Description of the Dataset is as follows:*

Product Identifier is a unique ID, for each product. Product Weight is the weight of the product. Fat Content of the Product indicates the concentration of fat in the product. Product Visibility shows the percentage of display area occupied by products in a store. Product Category shows the category to which the product belongs. Store Identifier shows An ID assigned to each store. Maximum Retail Price (MRP) of the Product tells the price at which the product is sold to customers. Year of Store Establishment is the year when the store was established. Store Size is Categorized based on the area size of the store. Location Type of Store is categorized based on city tiers or sizes. Type of Store shows whether it is a grocery store or a specific type of supermarket. Sales of the Product, in a Specific Outlet.

B. EDA

EDA[7] stands for "Exploratory Data Analysis." It is an essential step in the data analysis process where analysts or data scientists examine and visualize the data to gain insights, identify patterns, spot anomalies, and understand the underlying structure of the dataset.

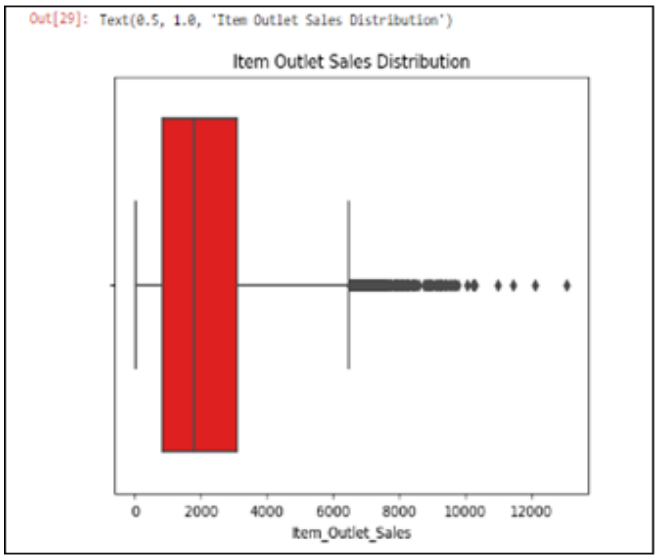

Fig 3 Item_Outlet_Sales_Distribution

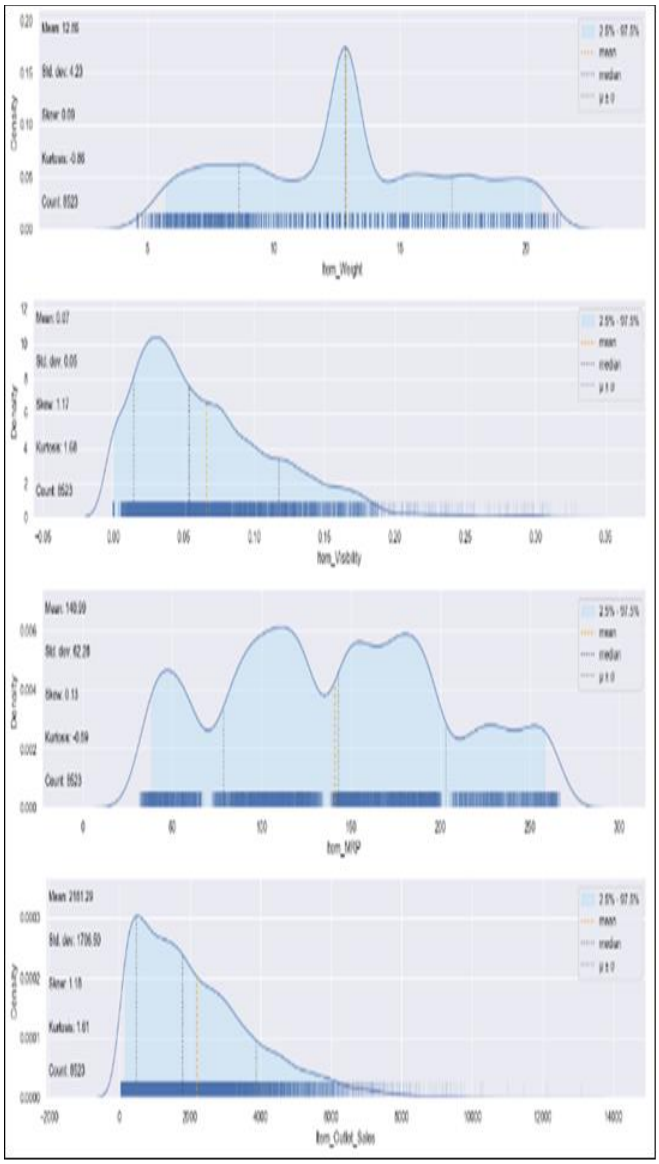

Fig 4 Density-Distribution Plot

ISSN No:-2456-2165

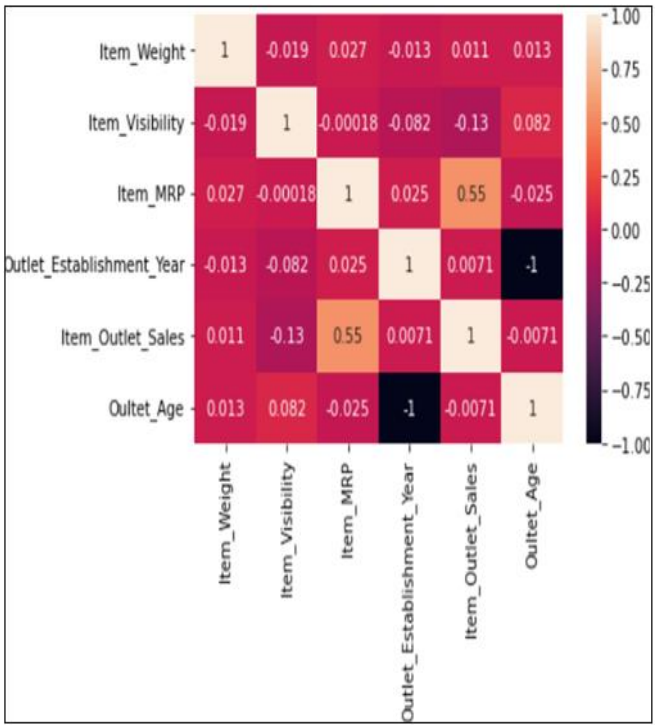

Fig 5 Correlation Matrix

➢ *Observations:*

- **IF content** shows that the most selling items are low fat content based.
- **Item_type shows that** is distinctly **fruits & veg**, **food snacks** are popular.
- **Item_Type_Combined shows that** most sold Item cateogory is food.
- **Outlet_Identifier shows most** sold items are ditributed evenly in all stores, execpt **OUT010** and **OUT019**.
- **Outlet Size** shows that Stores are mostly in **medium size** in this data.
- **Outlet location_Type**s shows that most common type of location is **Tier3**
- **Outlet_Type** shows that by a large margin Most Store Types are **SuperMarket Type1**.
- **Item_Visibility**: Looks like it has **negative correlation**.
- **Item_Weight** shows that there is noot any particular Pattern, Data is very spreaded.
- **Item_mrp** shows Items with high **MRP** Sales tends to sell better.

These Observation were made using the distribution and density plots of each category respectively.

C. Metrics for Evaluation

Validating the model plays a role, in the development of a machine learning model. Therefore it is important to construct a model and obtain recommendations, from it. Achieving precision takes time and effort as we continuously improve metrics based on the results. Evaluation metrics are used to describe the outcomes of a model with one key characteristic being their ability to distinguish between results. In this study we utilized the root mean squared error (RMSE) metric and r_2 score metric.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2$$

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2}$$

$$R^2 = 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2}$$

Where,

$\hat{y}$ – *predicted value of y*

$\bar{y}$ – *mean value of y*

Fig 6 Formulae of MSE,RMSE,R^2

IV. ALGORITHMS DEPLOYED

Scikit-Learn is a tool used in ML.It helps in various classification and regression algorithms[8]. Algorithms used for forecasting sales for this Big Mart dataset are explained below:

Linear Regression: It is the basic regression technique used for predicting the value of variable(dependent) using independent variables.

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_r x_r + e.$$

Where Y - variable to be predicted
and x – variable required for making a prediction
β_0, $\beta 1 \ldots \beta r$ - Regression Coeffecients e - Random Error

Random Forest[9]: The random forest Regressor is a ml algorithm that extends the Random Forest algorithm to handle regression tasks. It is commonly used for predicting values rather, than categorical class labels.

Decision Tree: A Decision Tree Regressor is an algorithm, in machine learning that's useful for solving regression problems. It is specifically designed to predict values as the target variable making it particularly helpful, in tasks involving data.

AdaBoost[10]: AdaBoost also known as Adaptive boosting is a popular ML ensemble method required for improving the performance of weak learners (usually decision trees) and creating a strong predictive model.

XGBoost[11]: XGB (Extreme Gradient Boosting) belongs to the family of boosting methods. Is a highly effective and widely used machine learning algorithm. It builds upon the Gradient Boosting Machines (GBM) algorithm offering efficiency, scalability and exceptional performance, across machine learning tasks.

V. RESULT

Different machine learning models such, as Linear Regression, KNN, Random Forest Regressor, Decision Tree Algorithm, Adaptive Boosting and XGBoost have been to forecast the sales revenue of Big Mart. Through analysis it has been determined that the Gradient Boosted XGBoost and Random Forest models yield the results in terms of predicting sales revenue, for Big Mart. These models exhibit the RMSE value and a high r2_score compared to algorithms.

Table 1 Accuracy Report

Models	r2_score	RMSE_score
Linear Regression	0.61	1134.65
KNN	0.63	1098.00
Random Forest	0.65	1068.48
Decision Tree	0.62	1104.34
AdaBoost	0.64	1078.65
XGBoost	0.68	1034.75

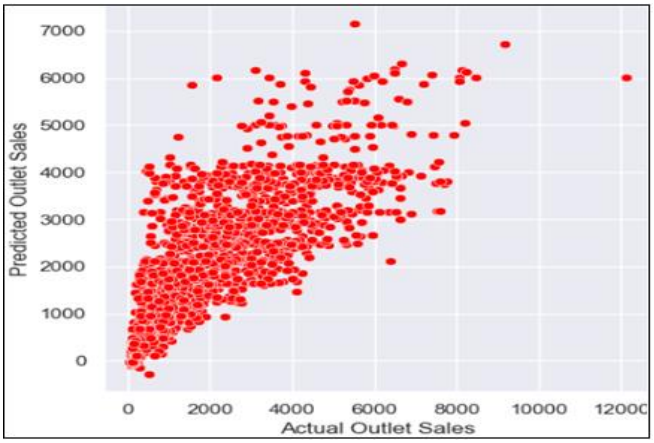

Fig 7 Scatter Plot Indicating the Predicted/Actual Outlet Sales

VI. CONCLUSION

In our research we examined the sales data from Big Mart. Utilized different regression algorithms to analyze it. We evaluated the performance of each algorithm. We found that the XGBoost Algorithm demonstrated the performance. These findings carry implications, for retailers as they strive to make sales predictions and optimize inventory management. Moving forward we suggest implementing ensemble techniques. Incorporating additional features to further improve the accuracy of predictions.

REFERENCES

[1]. Mascle, Christian, and Julien Gosse. "Inventory management maximization based on sales forecast: case study." Production Planning & Control 25.12 (2014): 1039-1057.

[2]. "Applied Linear Statistical Model", Fifth Edition by Ragg, Thomas, Wolfram Menzel, Walter Baum and Michael Wigbers. "Bayesian learning for sales rate prediction for thousands of retailers." Neurocomputing 43, no. 1-4 (2002):127-144.

[3]. Cheriyan, Sunitha, Shaniba Ibrahim, Sanju Mohanan and Susan Treesa. " Intelligent Sales Prediction Using Machine Learning Techniques." In 2018 International Conference on Computing, Electronics and Communication Engineering (iCCECE), pp. 53 – 58, IEEE, 2018.

[4]. Panjwani Mansi, Rahul Ramrakhiani, Hitesh Jumnani, Krishna Zanwar and Rupali Hande. "SalesPrediction System Using Machine Learning." No. 3243. EasyChair, 2020.

[5]. Singh Manpreet, Bhawick Ghutla, Reuben Lilo Jnr, Aesaan FS Mohammed and Mohmood A. Rashid "Walmart Sales Data Analysis – A Big Data Analytics Perspective." In 2017 4th Asia-Pacific World Conference on Computer Science and Engineering (APWC on CSE), pp. 114 – 119, IEEE, 2017

[6]. Behera, Gopal, and Neeta Nain. "A comparative study of big mart sales prediction." Computer Vision and Image Processing: 4th International Conference, CVIP 2019, Jaipur, India, September 27–29, 2019, Revised Selected Papers, Part I 4. Springer Singapore, 2020.

[7]. Aun, Yichiet, et al. "A Machine-Learning Ensemble Method for Temporal-aware Sales Forecasting." 2022 5th International Conference on Data Science and Information Technology (DSIT). IEEE, 2022.

[8]. Géron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. O'Reilly Media

[9]. Lin, Weiwei, et al. "An ensemble random forest algorithm for insurance big data analysis." Ieee access 5 (2017): 16568-16575.

[10]. Solomatine, Dimitri P., and Durga L. Shrestha. "AdaBoost. RT: a boosting algorithm for regression problems." 2004 IEEE International Joint Conference on Neural Networks (IEEE Cat. No. 04CH37541). Vol. 2. IEEE, 2004.

[11]. Shilong, Zhang. "Machine learning model for sales forecasting by using XGBoost." 2021 IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE). IEEE, 2021.