IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Vol. 12, Issue 4, April 2023

DOI: 10.17148/IJARCCE.2023.124112

BIG MART SALES PREDICTION USING MACHINE LEARNING

Sathyanarayana S[1], Apeksha C[2], Chethana S[3], Chinmayee H C[4], Abhishree G L[5]

Research Scholar, Department of Computer Science and Engineering, JNNCE, Shimoga, India[2,3,4,5]

Assistant Prof., Department of Computer Science and Engineering, JNNCE, Shimoga, India[1]

Abstract: Machine Learning has several category of algorithms that allows software applications to be more accurate in predicting results without being explicitly programmed. In this paper, the case of Big Mart, a one-stop-shopping center, has been discussed to predict the sales of different attributes, item varieties and for understanding the effects of different factors on the items' sales. Considering various aspects of a dataset collected for Big Mart, methodology followed for constructing a predictive model, highly accurate results are generated, and these observations can be employed to take decisions to improve sales.

Keywords*:* Machine Learning, Sales Prediction, XGBoost Algorithm, Random Forest Algorithm, Linear Regression algorithm.

I. INTRODUCTION

The competition among shopping malls and large supermarkets is intensifying with each passing day, owing to rapid development. These establishments are diligently recording sales data along with various dependent and independent factors. This data can be incredibly valuable in forecasting future demand and managing inventory.The use of machine learning in data science is rapidly expanding because of its ability to process data using mathematical, statistical, and econometric techniques swiftly, precisely, and accurately. It offers valuable business intelligence that facilitates the development and implementation of effective business strategies.

This paper aims to predict retail sales efficiently using machine learning algorithms. Its primary objective is to forecast sales and compare the results of implemented models based on RMSE, R2_score, and MAE. The structure of the paper is as follows: Section II discusses previous research papers, Section III covers the methodology and algorithms used, and Section IV compares the results of different models. The study is expected to be beneficial for retailers and decision-makers in the retail industry.

II. RELATED WORK

Ayesha Syed et al [1] experimented big mart sales using machine learning with data analysis on 2013 Big mart dataset using XGBoost algorithm and additional hyperparameter tuning was conducted on XGBoost with Bayesian Optimization Technique to get accurate results. RMSE is the only evaluation metrics used. An advanced sales forecasting using machine learning algorithms is conducted by K. Vedavathi et al [2] in which they made use of ARIMA model, XGBoost, Random forest algorithms for Rossmann shop dataset with 1017209 entries. The points used ranged from keep data to purchaser data as properly associated geographical information. The gradient boosting algorithm is very sensitive to the outliers. And the main limitation is it is almost impossible to scale up. Supermarket sales prediction using regression was suggested by Melvin Tom et al [5] to forecast sales using data with multiple instances parameters and various other factors can be used for predicting the sales more innovatively and successfully, but it achieved quite low accuracy.

During the experiment in T K Thivakaran et al [4] implemented A comparative study of statistical analysis on big mart using data mining techniques. Five different regression algorithms such as linear, ridge regression, decision tree, XGBoost, ARIMA were used. They found using XGBoost along with ARIMA model showed better results. But the gradient descent algorithm can suddenly change in wrong direction due to frequent updates. In Nayana R et al[9] presents Predictive Analysis for Big mart sales using machine learning algorithms like linear regression, polynomial regression, ridge regression, XGBoost regression in which they have shown that based on the data collected, the accuracy can be enhanced. The presence of one or two outliers in the data can seriously affect the result of the polynomial regression model. Navaneeth Kumar et al [7] sets out sales prediction, suggested to use hyperparameters that make algorithm shine and produce high accuracy. The accuracy of this method is above 80%. But the method and the result shown are in the complex form. In prediction of big mart sales by Naveenraj R et al [8] comparision of all popular algorithms is done. Performance of Random Forest and XGBoost over other algorithms are shown for big mart sales data 2013.

DOI: 10.17148/IJARCCE.2023.124112

In Varshini S et al [10] implemented an analysis of machine learning algorithms to predict sales using XGBoost regressor, ANN, Random Forest, SVR, on comparision they found Random Forest performs well. But accuracy can still be improved to reduce the gap between predicted sales and actual sales value.

III. METHODOLOGY

The primary objective of this research project is to analyse and predict future sales using various machine learning techniques that can produce comprehensive and reliable models.

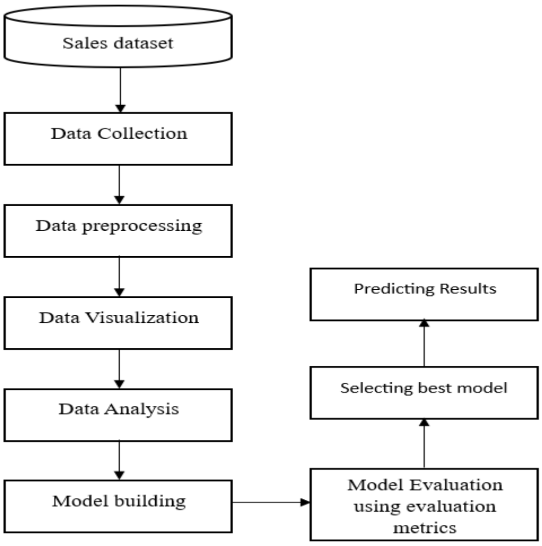

Fig. 1 System Framework

A. Data Collection

The data plays a vital role for any model to predict the future sales in a retailer's environment accurately. The dataset was collected from sample data that appears in the December Tableau User Group presentation which has 17 attributes.

TABLE 1 ATTRIBUTES INFORMATION

Invoice ID	A unique record number assigned to each invoice issued.
Branch	Branches of the store
City	name of the city where the branches of the store are located
Customer Type	Category of a customer like normal or member
Gender	Customer gender
Product Line	Describes the category to which the product belongs to
Unit price	Price of a product
Quantity	How many units the customer have purchased
Tax	An indirect tax imposed on goods and services
Total	Total amount paid on the product purchased
Date	Date at which the product is sold
Time	Time at which the product is sold
Payment	Describes how the payment is made
Cogs	Sum of all direct costs associated with making a product
Gross margin percentage	Store's profit percentage
Gross income	Store's profit income
Rating	The variable that has to be predicted; it contains the product's sales in the specific store

DOI: 10.17148/IJARCCE.2023.124112

B. Data Preprocessing

Data preprocessing is an essential step in machine learning where raw data is cleaned, missing values are removed, transformed, and organized into a format that can be easily understood and processed by machine learning models. The quality of the input data is critical to the success of any machine learning model, and data preprocessing helps to ensure that the data is suitable for the model to learn from.

C. Data Visualization

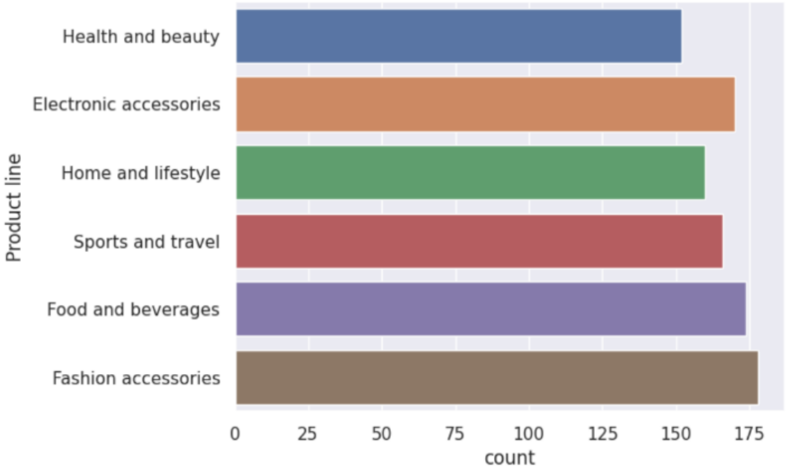

Fig. 2 Bar plot for product line and its count in given data.

In fig.2 there are six types of product line, out of which fashion accessories and electronic accessories sales count is higher, so the retailer can keep stock of this product line so that sales can be boosted.

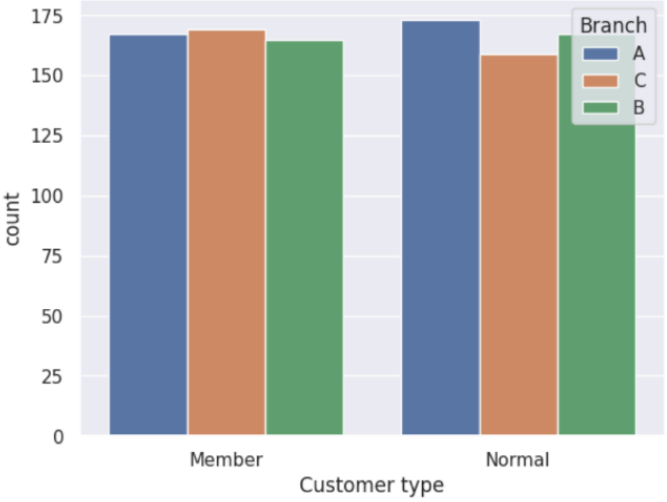

Fig. 3 Shows the count of sales for two customer types for each branch

In fig.3 it shows the count out of total sales for two different customer types for each branch.

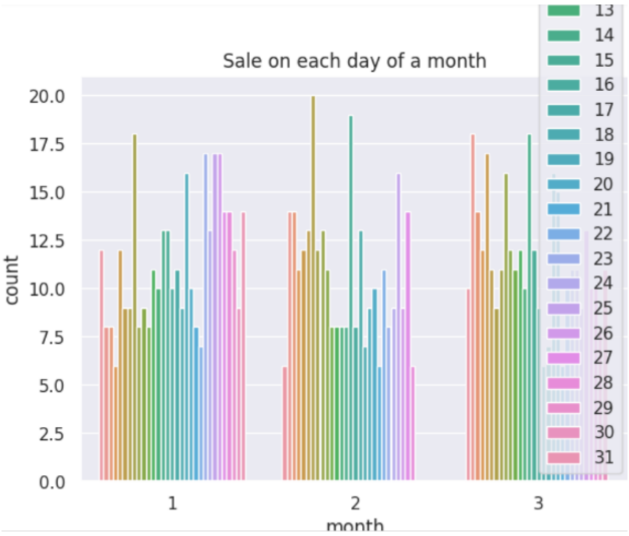

Fig. 4 Plot showing count of sales on each day of a month for three months

Fig.4 shows the count of sales on each day of a month for first three months of year 2019. Such plots will be helpful to visualize on which type of day sales will be more like, weekdays, weekends, holidays, working days.

D. Data Analysis

It is done after data preprocessing and visualization. Analysis helps to understand the nature of data.

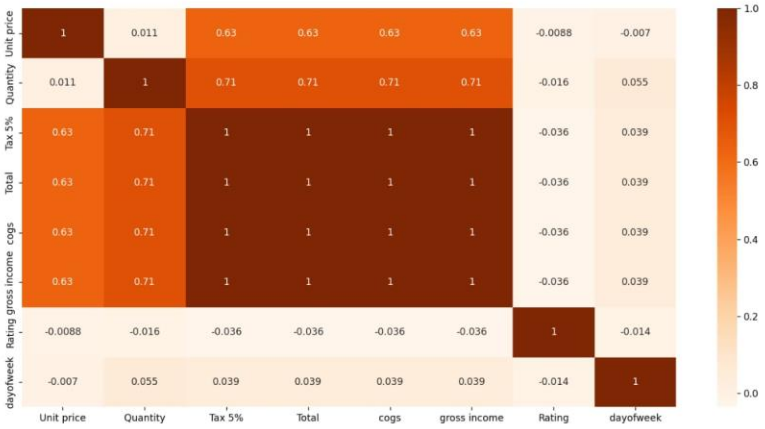

Fig. 6 Correlations between numerical variables and the target variable

Fig.6 shows correlation between numerical variables and the target variable. Rating has strong negative correlation with almost all features. Gross income is perfect positively correlated to tax, total and cogs. Similarly, cogs is perfectly correlated with other three features. Quantity has strong positive correlation with total, cogs, tax 5%, gross income.

E. Model Building

After completion of the preceding phases, the dataset is now ready to use to create a predictive model to forecast Big Mart's sales. We reviewed three machine learning algorithms that can be used to solve prediction problems in this research, including XGBoost, Random Forest, Linear regression.

- *XGBoost:* XGBoost (Extreme Gradient Boosting) is a popular machine learning algorithm used for regression and classification tasks. It is a boosted decision tree algorithm that combines the strengths of both gradient boosting and random forest methods. In XGBoost regression, the goal is to predict a continuous output variable based on a set of input features. The algorithm works by building a sequence of decision trees, where each tree is trained to correct the errors of the previous tree. The process continues until the error cannot be further reduced or a predefined stopping criterion is met.

- *Random Forest:* Random Forest is an ensemble method that combines the predictions of multiple decision trees, resulting in a more robust and accurate model. In Random Forest regression, the goal is to predict a continuous output variable based on a set of input features. The algorithm works by building many decision trees on random subsets of the training data and input features. Each tree is trained to predict the output variable based on a subset of the input features, and the final prediction is made by averaging the predictions of all trees.

- *Linear Regression:* Linear regression is used for predicting a continuous output variable based on one or more input features. It assumes that there is a linear relationship between the input features and the output variable. In linear regression, the goal is to find a linear function that best fits the data, by minimizing the difference between the predicted output and the actual output. This function is represented as:
y=b0 + b1x1 + b2x2 + … + bnxn
Where y is the output variable, b0 is the intercept, b1, b2, … bn are the coefficient of the input features x1, x2, …, xn.

F. Model Evaluation

Evaluation Metrics

- *RMSE:* RMSE (Root Mean Squared Error) is a commonly used evaluation metric in machine learning for regression tasks. It measures the average deviation of the predicted values from the actual values, and is defined as the square root of the average of the squared differences between the predicted and actual values.

The formula for calculating RMSE is:

RMSE = sqrt (1/n * sum((y_pred - y_actual)^2))
where n is the number of observations, y_pred is the predicted value, and y_actual is the actual value.

- *R2_Score:* It measures the proportion of variance in the output variable that is explained by the input features. R2_score takes values between 0 and 1, where 1 indicates a perfect fit of the model to the data and 0 indicates a model that performs no better than a model that predicts the mean of the output variable. The formula for calculating R2_score is:

R2_score = 1 - (sum ((y_actual - y_pred) ^2) / sum ((y_actual - y_mean) ^2))
where y_actual is the actual value, y_pred is the predicted value, and y_mean is the mean value of the output variable.

- *MAE:* The Mean Absolute Error (MAE) measures the average absolute magnitude between the actual values and the predicted values by regression model. The MAE can be written mathematically as
MAE = (1/n) * Σ|yi — xi|
Where, n = total numbers of observation
yi = actual value for the ith observation
xi = predicted value for the ith observation

IV. RESULTS AND DISCUSSION

For the given dataset, the various models mentioned before were employed to accomplish the prediction. The models were assessed using the RMSE, R - squared, and MAE metrics.

TABLE 2 COMPARISION TABLE

Algorithm	RMSE	R2_score	MAE	Accuracy
XGBoost	0.2188	0.9938	0.1329	99.386
Random Forest	0.2437	0.9932	0.1521	99.319
Linear regression	0.1482	0.9975	0.0694	81.218

From the table above, XGBoost performed well compared to other algorithms with better accuracy and lesser RMSE, R2_score, MAE value.

DOI: 10.17148/IJARCCE.2023.124112

V. CONCLUSION

In the modern world, retailers and businesspeople seek to anticipate customer preferences in advance to ensure that they have adequate stock available. Accurate day-to-day forecasts allow retailers to identify peak sales days and periods of the month, enabling companies to achieve a higher return on investment. By making better predictions, businesses can avoid losses.

Our research involved the use of three regression algorithms, among which XGBoost outperformed the other two with an accuracy of 99.386%. When predicting the quantity of the first product line sold, XGBoost produced a result of 7.005, closely resembling the actual value of 7. This information can be valuable for retailers in making informed decisions. Moving forward, we aim to improve the results further by employing more efficient techniques for larger datasets.

REFERENCES

[1]. Ayesha Syed, Asha Jyothi Kalluri, Venkateswara Reddy Pocha, Venkata Arun Kumar Dasari, B. Ramasubbaiah, BIGMART SALES USING MACHINE LEARNING WITH DATA ANALYSIS," Journal of Engineering Science. Vol 11, Issue 2, FEB/2020 ISSN NO:0377-9254

[2]. B.Sri Sai Ramya , K. Vedavathi," An Advanced Sales Forecasting Using Machine Learning Algorithm**",** International Journal of Innovative Science and Research Technology. Volume 5, Issue 5, May – 2020 ISSN No: -2456-2165

[3]. Purvika Bajaj, Renesa Ray, Shivani Shedge, Shravani Vidhate, Prof. Dr. Nikhilkumar Shardoor," SALES PREDICTION USING MACHINE LEARNING ALGORITHMS", International Research Journal of Engineering and Technology (IRJET), Volume: 07 Issue: 06 | June 2020

[4]. T K Thivakaran , Dr M Ramesh, **"**A Comparative Study of Statistical Analysis on Big Mart using Data Mining Techniques**"**, International Journal of Advanced Trends in Computer Science and Engineering. Volume 9, No.5, September-October 2020 ISSN 2278-3091

[5]. Melvin Tom, Nayana Raju, Asha Issac, Jeswin James, Rani Saritha R, "Supermarket Sales Prediction Using Regression", International Journal of Advanced Trends in Computer Science and Engineering. Volume 10, No.2, March - April 2021 ISSN 2278-3091

[6]. Donti Reddy Sai Rakesh Reddy, Katanguru Shreya Reddy, S. Namrata Ravindra B. Sai Sahith," Prediction and Forecasting of Sales Using Machine Learning Approach", International Research Journal of Engineering and Technology (IRJET), Volume: 08 Issue: 09 | Sep 2021

[7]. Asst. Prof. Keyaben patel, Navneet Kumar, Suraj Choudhari," Big Mart Sale Prediction using Machine Learning", International Journal of Innovative Science and Research Technology, Volume 6, Issue 9, September – 2021, ISSN No: -2456-2165

[8]. Naveenraj R, Vinayaga Sundharam R," PREDICTION OF BIG MART SALES USING MACHINE LEARNING", International Research Journal of Modernization in Engineering Technology and Science,Volume:03/Issue:09/September-2021 e-ISSN: 2582-5208

[9]. Nayana R, Chaithanya G , Meghana T , Narahari K S , Sushma M," Predictive Analysis for Big Mart Sales using Machine Learning Algorithms",2022, International Journal of Engineering Research & Technology (IJERT). Volume 10, Issue 12 ISSN: 2278-0181

[10]. Ranjitha P, Spandana.M "Predictive Analysis for Big Mart Sales Using Machine Learning Algorithms" Proceedings of the Fifth International Conference on Intelligent Computing and Control Systems (ICICCS 2021) IEEE Xplore Part Number: CFP21K74-ART; ISBN: 978-0-7381-1327-2.

[11]. Dr. Bandaru Srinivasa Rao, Dr. Kamepalli Sujatha, Dr. Nannpaneni Chandra Sekhara Rao, Mr. T.Nagendra Kumar," Retail Sales Prediction Using Machine Learning Algorithms", Turkish Online Journal of Qualitative Inquiry (TOJQI), Volume 12, Issue 1, Janurary 2021: 315-322

[12]. Sanjay. N Gunjal, D.B Kshirsagar, B.J Dange, H.E Khodke, C.S Kulkarni," Machine Learning Approach for Big-Mart Sales Prediction Framework", International Journal of Innovative Technology and Exploring Engineering (IJITEE), Volume-11 Issue-6, May 2022, ISSN: 2278-3075

[13]. Gopal Behera, Neeta Nain "Grid Search Optimization (GSO) Based Future Sales Prediction for Big Mart," 2019 15th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)

[14]. Manish Kumar Nishad, Sujata Kondekar "BIG MART SALES PREDICTION", International Research Journal of Modernization in Engineering Technology and Science, Volume:04/Issue:05/May-2022, e-ISSN: 2582-5208

[15]. Abhay Mishra, Mohd Hamd, Anubhav Yadav, Shubham Tiwari "Sales Component Analysis & Prediction Using Linear Regression" IJIRT Volume 8 Issue 2 ISSN: 2349-6002, July 2021

[16]. Varshini S, Dr. D. Preethi," An Analysis of Machine Learning Algorithms to Predict Sales," International Journal of Science and Research (IJSR). Volume 11 Issue 6, June 2022