JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

BIG MART SALES PREDICTION USING XG BOOST

[1]Shital Dilpak, [2]Mansi Padwal, [3]Vaishnavi Patil and Prajkta Khaire
[1,2,3] Scholar, [4]Professor
Department of Information Technology,
SSJCOE, Dombivli, India

Abstract : Nowadays malls and Big Marts track their sales data of every single item to predict future customer demand and also update inventory management. These data warehouses essentially contain a large amount of customer data and attributes of individual items in the data warehouse. In addition, anomalies and frequent patterns are detected by data mining from the data warehouse. The resulting data can be used to predict future sales volume using various machine learning techniques for retailers like Big Mart. In this paper, we propose a predictive model using the XG boost Regressor technique to predict the sales of a company like Big Mart, and we find that this model provides better performance compared to existing models.

KEYWORDS: Machine Learning , Predictive Analytics , Sales Prediction

1. INTRODUCTION

Big Mart is a large supermarket chain with stores across the country and its current board of directors has set a challenge for all the Data Scientists out there to help them build a model that can predict sales per product for each store and provide accurate results. . Big Mart collected sales data from Kaggle for different products in different stores in different cities. The company hopes that we can use this information to identify the products and stores that play a key role in their sales and use this information to take the right actions to ensure the success of their business. Today, malls and Big Marts track sales data for every single item to predict future customer demand and update inventory management. In a data warehouse, these data stores essentially contain large amounts of customer data and individual item attributes. Generally, sales forecasting is critical for marketing, retail, wholesale, and manufacturing, and it plays an important role for marketing, retail, wholesale, and manufacturing in various companies. This proposed system will enable companies to plan a better strategy and achieve sales and provide companies with better growth in the future. Compared to other learning methods, this machine learning method provides accurate results. In recent years, the application of machine learning techniques has shown promising results in increasing the accuracy of sales forecasts. This research paper presents a comprehensive study on Big Mart sales prediction using machine learning algorithms. By using a diverse set of regression algorithms, including linear regression, decision trees, random forests, XG BOOST Regressor, we try to identify the most accurate ones for sales prediction. The competition between shopping malls and large supermarkets is increasing day by day due to rapid development. These devices carefully record sales data along with various dependent and independent factors. This data can be incredibly valuable in predicting future demand and managing inventory. The use of machine learning in data science is rapidly expanding due to its ability to process data using mathematical, statistical, and econometric techniques quickly, accurately, and precisely. It offers valuable business insights that facilitate the development and implementation of effective business strategies.

1.1 OBJECTIVES

1) The objective of Big Mart sales prediction is to predict the sales of different products in different stores operated by Big Mart using historical data and relevant features. This forecast helps in optimizing inventory management, designing effective marketing strategies and maximizing company revenue.
2) Improved pricing strategies: This can lead to increased revenue and better profit margins.
3) Improved inventory management: Big , mart can ensure that the right products are in stock at the right time, reducing the risk of overstocking and understocking. This minimizes costs and maximizes sales opportunities

4) Optimized Marketing and Promotion: Sales forecasting informs Big Mart's marketing and promotional strategies by identifying opportunities to boost sales through targeted campaigns and discounts. By understanding customer preferences and buying patterns, they can personalize marketing messages and promotions to increase effectiveness and ROI.

2. RELATED WORK

In today's competitive market, every company wants to stand out from the competition. Sales forecasting is a good idea for a company to analyze product sales.

Nikita Malik discussed sales prediction using machine learning. She used a machine learning algorithm (linear regression, Random Forest, etc.). She analyzed several products and established some correlation between the product and the business. The accuracy is between 70% and 80%. [1]

Aditi Narkhede collected the Big Mart dataset and used the ML algorithm to find the RMSE value. It did some calculation to find the RMSE value and it's pretty easy to use. She made the calculation look easy to use. [2]

Rajendra Pamula discussed the flowchart to understand things easily. Here he also used machine learning and data mining. The accuracy is between 50 and 60%. He used some complex calculation to get an output that is not easy to understand. [3]

Saju Mohanan used data mining techniques and machine learning algorithm for sales prediction. He used a decision tree and a generalized linear model for prediction. The accuracy of the model varies between 60 and 70%. He also drew the architecture of the system to keep it simple, but the output is in a very complex form. [4]

Pavan Chatradi discussed sales prediction using ML and XG boosting technique. Here he performed steps like data cleaning, data transformation, data reduction. The accuracy of this method is over 80%. But the method shown and the result are in complex form. Here the steps involved in forecasting are Dataset -> Data Exploration -> Data Cleaning -> Feature Engineering -> Model Building -> Model Testing -> Result. [5]

A machine learning technique to predict how much shoppers will spend on the next Black Friday deal. A decision was made to use exploratory data analysis to find relevant patterns in the data set. According to this study, when a user tries to predict which product a client is more likely to buy [9]

According to the author, the weather affects the regular sales of restaurants. The accuracy of two machine learning techniques, XG Boost and neural network, was investigated. and found that the XG Boost technique outperformed the neural network technique. They also found that taking meteorological variables into account improved the performance of their model by 2–4 percentage points. They took into account various criteria to increase accuracy, including date characteristics, sales history and weather conditions. [10]

3. LITERATURE REVIEW

Large Market Sales Forecasting Based on Random Forests and Multiple Linear Regression (2018) Kadam, H., Shevade, R., Ketkar, P. and Rajguru. A forecast for a large sales market based on random forest and multiple linear regression used random forest and linear regression for prediction analysis, which provides less accuracy. To overcome this problem, we can use XG boost Algorithm, which will provide more accuracy and be more efficient. [1]

Prognostic Methods and Applications (2008) Makridakis, S., Wheelwrigh. S.C., Hyndman. R. J. Prognostic methods and applications contain a lack of data and short life cycles. Thus, some data, such as historical data and consumer-oriented markets, face uncertain requirements and can be predicted for accurate results. [2]

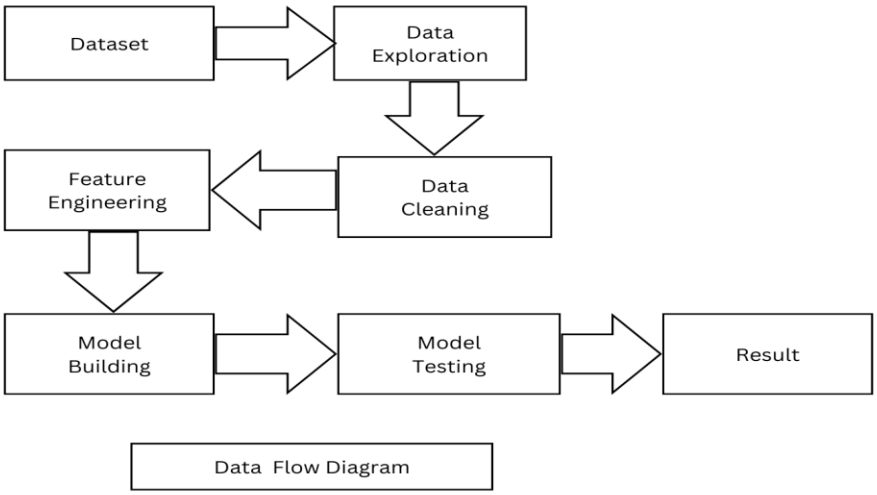
Comparing Different Machine Learning Algorithms for Multiple Regression on Black Friday Sales Data (2018) C. M. Wu, P. Patil, and S. Gunaseelan. Comparing Different Machine Learning Algorithms for Multiple Regression on Black Friday Sales Data used a neural network to compare different algorithms. To overcome this Complex model like neural networks is used for comparison between different algorithms which is not effective so we can use simpler algorithms for prediction. Prediction of Footwear Retail Sales Using Feedforward and Recurrent Neural Networks (2018) by Das, P., Chaudhury. Prediction of shoe retail sales using feedforward and recurrent neural networks used neural networks to predict sales. Using a neural network to forecast weekly retail sales is not efficient, so XG boost can work effectively. [3]

Complex models like neural networks are overkill for simple problems like regression. And even simpler models proper data cleaning works well for regression[4].
Using Random Forest, revenue prediction is easier and attention is paid to determining the optimal number of trees[5].

Random Forest is a tree algorithm in which a certain number of decision trees are combined to make it powerful prediction model. A general linear model was found using principal component analysis and random forest techniques provide better results, which are decided by the RMSE values[6].

Sales forecasts provide insight into how a business should manage its workforce, cash flow and assets. That's important a prerequisite for business planning and decision-making. It enables companies to formulate their business plans effectively[7].

4. PROPOSED SYSTEM



(FIG 1 SYSTEM ARCHITECTURE)

Big Mart sales data set to create a model to predict accurate results it goes through several sequences of steps as shown in Figure 1 and in this work we design a model using the Xgboost technique. Each step plays an important role in building in the proposed model. In our model, we used the 2013 Big mart dataset . After preprocessing and filling in missing values, we used a file classifier Decision Trees, Linear Regression, Ridge Regression, Random Forest and Xgboost. Both MAE and RSME are used as accuracy metrics for forecasting sales v
Big Mart.

5. METHODOLOGY

1 Data collection: We have collected data securely in accordance with agreed methodology. The procedure for the data collected may vary from client to client and depends on the type, amount, availability and need for the data.

2 Data Cleansing and Pre-processing: Collected data goes through a “cleansing” process to ensure that the data is properly segregated and any identified data gaps are filled with appropriate information to ensure data compatibility as well as to correct errors in the repository. systems that may cause data redundancy.

3 Data Modeling: This is primarily the process of analyzing a given dataset and the objects within it to gain a clear understanding of the requirements that can help us support our business model. Based on pattern analysis models are then created in the data based on the established flow of the project. This flow offers better assistance in using a previously agreed semi-formal model that shows the properties of the project. It also provides guidelines for tracking the relationship between data objects and other objects.

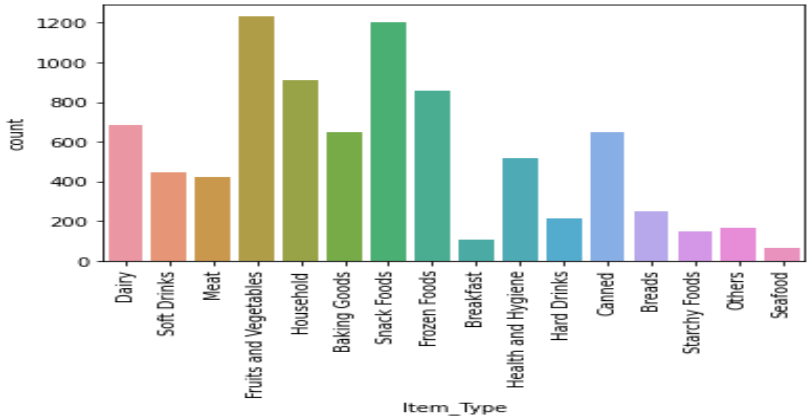
4 Data Prediction: Predictive machine learning models are trained in this process and later evaluated using the data. This is then applied to the preprocessed dataset. Some of the models to be used for prediction are:
E. Linear regression
F. Random forest
G. Decision tree
h. XG Boost Regressor

5 Data visualization: The analyzed data is then further displayed for customers and administrators to draw conclusio ns and take effective decisions.

```
big_mart_data.describe()
```

	Item_Weight	Item_Visibility	Item_MRP	Outlet_Establishment_Year	Item_Outlet_Sales
count	8523.000000	8523.000000	8523.000000	8523.000000	8523.000000
mean	12.857645	0.066132	140.992782	1997.831867	2181.288914
std	4.226124	0.051598	62.275067	8.371760	1706.499616
min	4.555000	0.000000	31.290000	1985.000000	33.290000
25%	9.310000	0.026989	93.826500	1987.000000	834.247400
50%	12.857645	0.053931	143.012800	1999.000000	1794.331000
75%	16.000000	0.094585	185.643700	2004.000000	3101.296400
max	21.350000	0.328391	266.888400	2009.000000	13086.964800

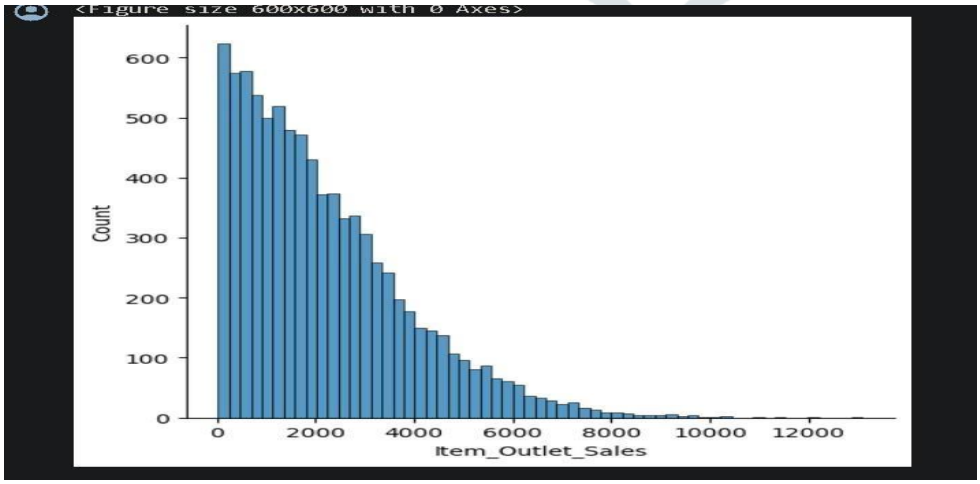
(FIG : Screenshot of dataset)



(FIG: Result of Big Mart prediction testing for different items)

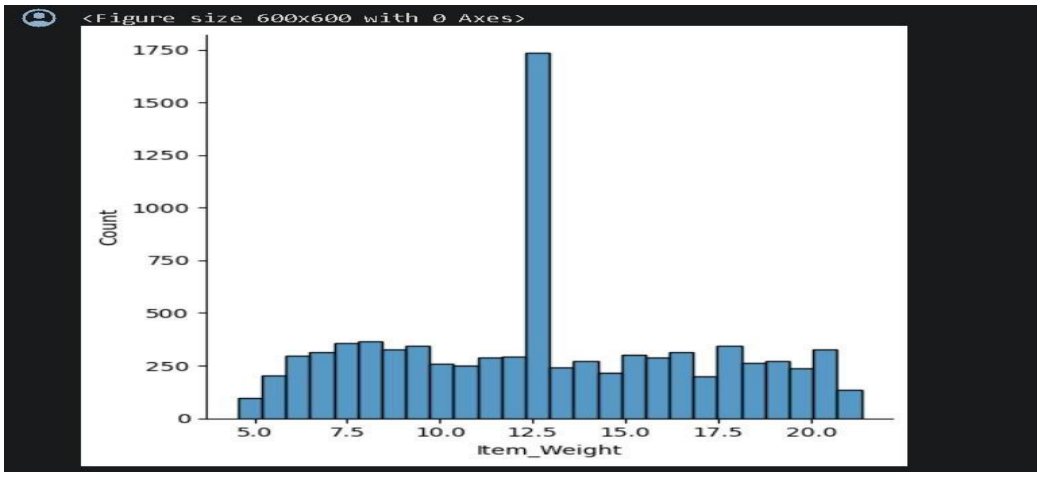
6. RESULT

1 Distribution of data for item _outlet_sales



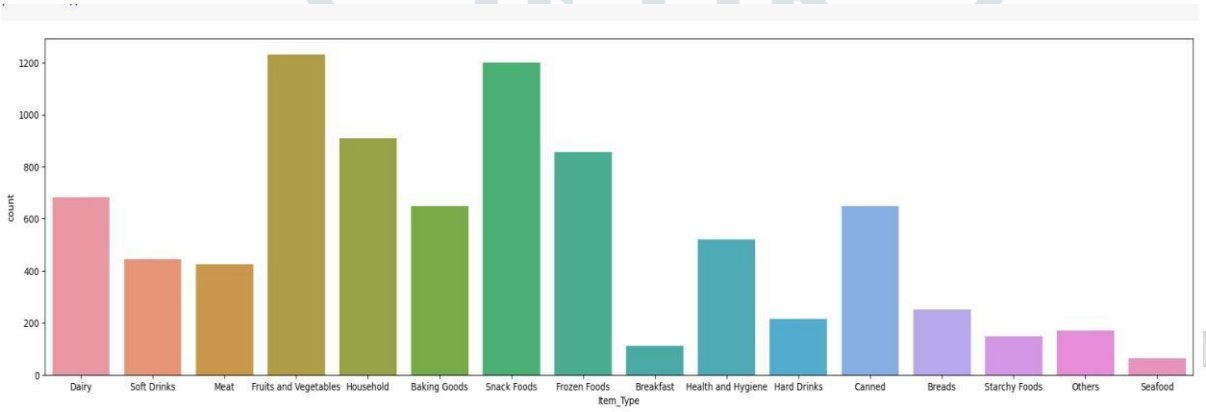
(fig 1 Distribution of data for item _outlet_sales)

2. Data distribution for the Item Weight column



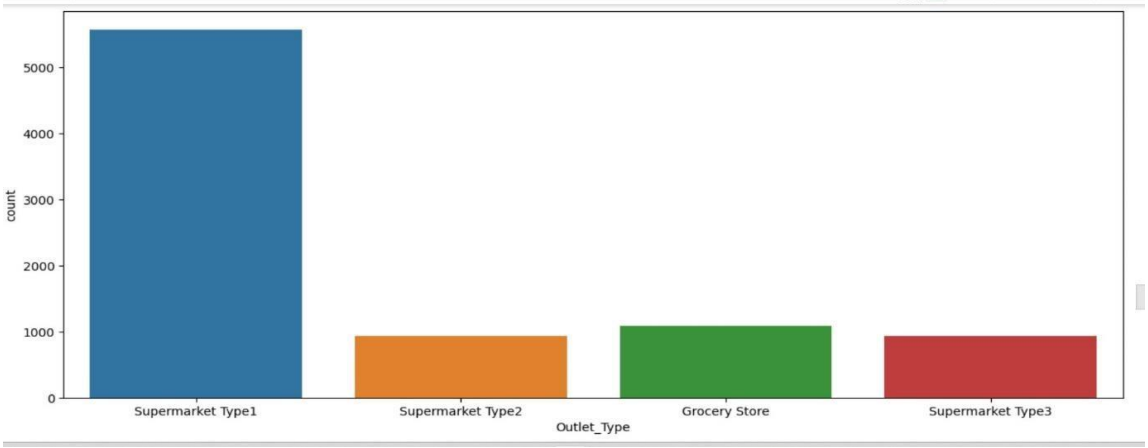
(fig 2 Data distribution for the Item Weight column)

3 Count chart for Item_type column



(fig 3 Count chart for Item_type column)

4 Count chart for Item_type column



(fig 4 Count chart for Item_type column)

7. CONCLUSION

This "BIG MART SALES PREDICTION" project offers a number of benefits such as improved inventory management, improved marketing strategies, increased revenue, cost savings, customer satisfaction and competitive advantage. Big Mart can optimize operations, respond effectively to market trends and ultimately drive business success. In today's digitally connected world, every shopping center wants to know the customer's requirements in advance to avoid sales shortages in all seasons. Companies or trading centers are forecasting more accurately day by day. In this work, the efficiency of decision tree regression on revenue and review data, the best performance-algorithm, here propose a software to use the regression approach for sales prediction focusing on past sales data, the prediction accuracy of linear regression can be improved by this method, and the regression can be determined decision tree. So we can conclude that decision tree regression provides better prediction with respect to accuracy

8. FUTURE WORK

Big Mart can use advanced predictive analytics techniques such as machine learning algorithms and deep learning models to increase the accuracy of sales forecasts. These models can analyze vast amounts of data, including historical sales data, customer behavior patterns, weather forecasts, and economic indicators to produce more accurate predictions. With real-time data streams and Internet of Things (IoT) devices, Big Mart can achieve real-time sales forecasting capabilities.

Real-Time Prediction and Dynamic Pricing: The future of Big Mart's sales prediction may include real-time prediction capabilities that enable dynamic pricing strategies. By constantly analyzing data and adjusting prices based on demand and market conditions, Big Mart can optimize revenue and profitability in real time.

ACKNOWLEDGMENTS

We would like to express our gratitude to our school principal, Dr. Pramod R. Rodge, for providing us with laboratory facilities and allowing us to continue our project. We also sincerely thank our H.O.D., Dr. Savita S. Sangam, who provided us with the necessary computer equipment in our laboratory and played a significant role in the success of our project. Without their cooperation, our project would have stopped. We would also like to thank our project guide, Dr. Savita S. Sangam, for her unwavering support throughout the project. Her professional guidance, kind advice and timely motivation were invaluable in the development of our project "Big Sales Prediction Machine Learning".

We would like to thank our colleagues who directly or indirectly supported us during the project. Finally, we sincerely thank everyone who participated in our project and helped us achieve success.

REFERENCES

[1] Ching Wu Chu and Guoqiang Peter Zhang, “A Comparative Study of Linear and Nonlinear Models for Forecasting Aggregate Retail Sales,” Int. Journal Production Economics, vol. 86, pp. 217-231, 2003

[2] Zone-Ching Lin, Wen-Jang Wu, “Multiple Linear Regression Analysis of Overlap Accuracy Model Zone”, IEEE Trans. On Semiconductor Manufacturing, vol. 12, No. 2, pp. 229–237, May 1999.

[3] O. Ajao Isaac, A. Abdullahi Adedeji, I. Raji Ismail, “Polynomial Regression Model of Cost Forecasting in Mixed Cost Analysis”, Int. Journal of Mathematical Theory and Modeling, Vol. 2, No. 2, pp. 14 – 23, 2012.

[4] Shashua, A. (2009). Introduction to Machine Learning: Class notes 67577. arXiv preprint arXiv:0904.3664.

[5]Zone-Ching Lin, Wen-Jang Wu, "Multiple Linear Regression Analysis of Zone Overlap Accuracy Model", IEEE Trans. on Semiconductor Manufacturing, Vol. 12, No. 2, pp. 229–237, May 1999.

[6]D. Fantazzini, Z. Toktamysová, Forecast of German car sales using Google data and multivariate models, Int. J. Production Economics 170 (2015) 97-135.

[7] MacKay, D.J., MacKay, D.J. (2003). Information theory, inference and learning algorithms. Cambridge University Press.