e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:08/August-2022 Impact Factor- 6.752 www.irjmets.com

BIG MART SALES PREDICTION USING MACHINE LEARNING

Yashraj Bharambe*[1]

*[1]Information Technology, JSPM's Rajarshi Shahu College Of Engineering, India.

ABSTRACT

Nowadays shopping malls and Big Marts keep the track of their sales data of each and every individual item for predicting future demand of the customer and update the inventory management as well. These data stores basically contain a large number of customer data and individual item attributes in a data warehouse. Further, anomalies and frequent patterns are detected by mining the data store from the data warehouse. The resultant data can be used for predicting future sales volume with the help of different machine learning techniques for the retailers like Big Mart. In this paper, we propose a predictive model using XG boost Regressor technique for predicting the sales of a company like Big Mart and found that the model produces better performance as compared to existing models.

Keywords: Machine Learning, Sales Prediction, Big Mart, XG Boost Regressor.

I. INTRODUCTION

In today's modern world, huge shopping centres such as big malls and marts are recording data related to sales of items or products with their various dependent or independent factors as an important step to be helpful in prediction of future demands and inventory management. The dataset built with various dependent and independent variables is a composite form of item attributes, data gathered by means of customer, and also data related to inventory management in a data warehouse. The data is thereafter refined in order to get accurate predictions and gather new as well as interesting results that shed a new light on our knowledge with respect to the task's data. This can then further be used for forecasting future sales by means of employing machine learning algorithms such as the random forests and simple or multiple linear regression model.

II. LITERATURE SURVEY

1] **Title**: A Forecast for Big Mart Sales Based on Random Forests and Multiple Linear Regression (2018)

Author: Kadam, H., Shevade, R., Ketkar, P. and Rajguru

Description: A Forecast for Big Mart Sales Based on Random Forests and Multiple Linear Regression. Random Forest and Linear Regression were the algorithms used for prediction analysis which gives less accuracy. To overcome this we can use XG boost Algorithm which will give more accuracy and will be more efficient.

2] **Title**: Forecasting methods and applications (2008)

Author: Makridakis, S., Wheelwrigh.S.C., Hyndman. R.J

Description: Forecasting methods and applications contains Lack of Data and short life cycles. So some of the data like historical data, consumer-oriented markets face uncertain demands, can be prediction for accurate result.

3] **Title**: -Comparison of Different Machine Learning Algorithms for Multiple Regression on Black Friday Sales Data (2018)

Author: C. M. Wu, P. Patil and S. Gunaseelan

Description: Comparison of Different Machine Learning Algorithms for Multiple Regression on Black Friday Sales Data Used Neural Network for comparison of different algorithms. To overcome this Complex models like neural networks is used for comparison between different algorithms which is not efficient so we can use more simpler algorithm for prediction.

4] **Title**: Prediction of retail sales of footwear using feed forward and recurrent Neural Networks (2018)

Author: Das, P., Chaudhury

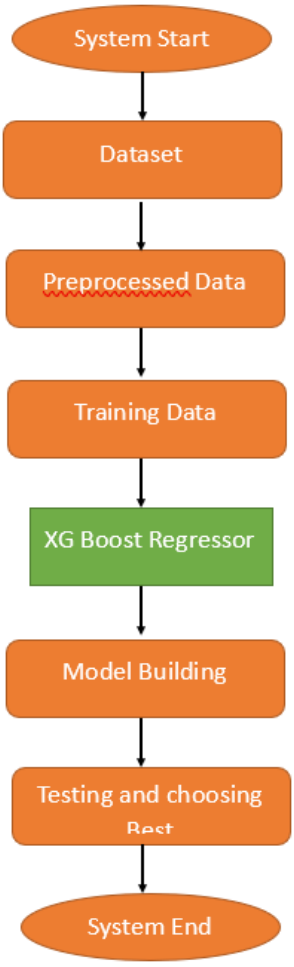
Description: Prediction of retail sales of footwear using feed forward and recurrent neural networks used neural networks for prediction of sales. Using neural network for predicting of weekly retail sales, which is not efficient, So XG boost can work efficiently.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:08/August-2022 Impact Factor- 6.752 www.irjmets.com

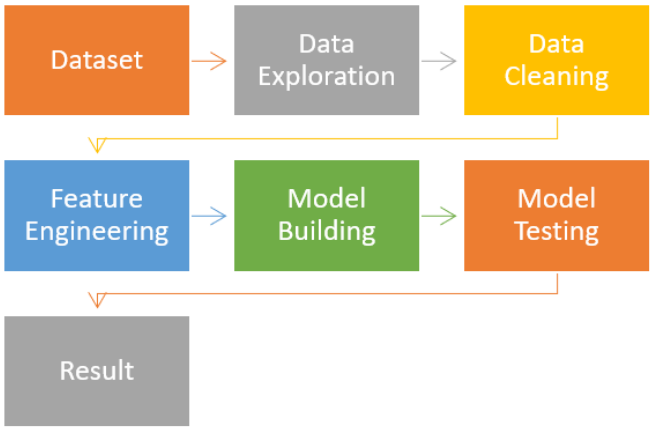
III. AIMS AND OBJECTIVES

1. The objective of this framework is to predict the future sales from given data of the previous year's using Machine Learning Techniques.
2. Another objective is to conclude the best model which is more efficient and gives fast and accurate result by using XG Boost Regressor.
3. To find out key factors that can increase their sales and what changes could be made to the product or store's characteristics.

IV. SYSTEM ARCHITECTURE



V. PROPOSED MODEL



e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:08/August-2022 Impact Factor- 6.752 www.irjmets.com

VI. PHASES IN MODEL

1. Dataset :

A data set is a collection of data. In case of a tabular data, a data set is in correspondence to one or more data base tables, wherein a particular variable is represented by each column, and the given record of the data set in question is represented by each row.

	Item_Iden	Item_Weig	Item_Fat_	Item_Visib	Item_Type	Item_MRP	Outlet_Ide	Outlet_Est	Outlet_Siz	Outlet_Lo	Outlet_Ty	Item_Outlet_Sales
2	FDA15	9.3	Low Fat	0.016047	Dairy	249.8092	OUT049	1999	Medium	Tier 1	Supermark	3735.138
3	DRC01	5.92	Regular	0.019278	Soft Drinks	48.2692	OUT018	2009	Medium	Tier 3	Supermark	443.4228
4	FDN15	17.5	Low Fat	0.01676	Meat	141.618	OUT049	1999	Medium	Tier 1	Supermark	2097.27
5	FDX07	19.2	Regular	0	Fruits and	182.095	OUT010	1998		Tier 3	Grocery St	732.38
6	NCD19	8.93	Low Fat	0	Household	53.8614	OUT013	1987	High	Tier 3	Supermark	994.7052
7	FDP36	10.395	Regular	0	Baking Goo	51.4008	OUT018	2009	Medium	Tier 3	Supermark	556.6088
8	FDO10	13.65	Regular	0.012741	Snack Foo	57.6588	OUT013	1987	High	Tier 3	Supermark	343.5528
9	FDP10		Low Fat	0.12747	Snack Foo	107.7622	OUT027	1985	Medium	Tier 3	Supermark	4022.764
10	FDH17	16.2	Regular	0.016687	Frozen Foo	96.9726	OUT045	2002		Tier 2	Supermark	1076.599
11	FDU28	19.2	Regular	0.09445	Frozen Foo	187.8214	OUT017	2007		Tier 2	Supermark	4710.535
12	FDY07	11.8	Low Fat	0	Fruits and	45.5402	OUT049	1999	Medium	Tier 1	Supermark	1516.027
13	FDA03	18.5	Regular	0.045464	Dairy	144.1102	OUT046	1997	Small	Tier 1	Supermark	2187.153
14	FDX32	15.1	Regular	0.100014	Fruits and	145.4786	OUT049	1999	Medium	Tier 1	Supermark	1589.265
15	FDS46	17.6	Regular	0.047257	Snack Foo	119.6782	OUT046	1997	Small	Tier 1	Supermark	2145.208
16	FDF32	16.35	Low Fat	0.068024	Fruits and	196.4426	OUT013	1987	High	Tier 3	Supermark	1977.426
17	FDP49	9	Regular	0.069089	Breakfast	56.3614	OUT046	1997	Small	Tier 1	Supermark	1547.319

2. Data Exploration :

In this phase useful information about the data has been extracted from the dataset. That is trying to identify the information from hypotheses vs available data. Which shows that the attributes Outlet size and Item weight face the problem of missing values, also the minimum value of Item Visibility is zero which is not actually practically possible. The response variable i.e. Item Outlet Sales, was positively skewed. So, to remove the skewness of response variable a log operation was performed on Item Outlet Sales.

3. Data Cleaning :

It was observed from the previous section that the attributes Outlet Size and Item Weight has missing values. In my work in case of Outlet Size missing value we replace it by the mode of that attribute and for the Item Weight missing values we replace by mean of that particular attribute. The missing attributes are numerical where the replacement by mean and mode diminishes the correlation among imputed attributes. For my model I am assuming that there is no relationship between the measured attribute and the imputed attribute.

4. Feature Engineering :

There were some nuances in the data-set during data exploration phase. So this phase is used in resolving all nuances found from the dataset and make them ready for building the appropriate model. During this phase it was noticed that the Item visibility attribute had a zero value, practically which has no sense. So the mean value item visibility of that product will be used for zero values attribute. This makes all products likely to sell. All categorical attributes discrepancies are resolved by modifying all categorical attributes into appropriate ones. Finally, for determining how old a particular outlet is, we add an additional attribute Year to the dataset.

5. Model Building :

After the previous phases are completed, the dataset is now ready to build proposed model. Once the model is build it is used as predictive model to forecast sales of Big Mart.

6. Model Testing :

In machine learning, model testing is referred to as the process where the performance of a fully trained model is evaluated on a testing set.

After the Model is built using Training Data, it is tested with the help of Testing Data.

VII. CONCLUSION

I have successfully predicted the accuracy for XG Boost Regressor. These predictions help Big Mart to refine their methodologies and strategies which in turn helps them to increase their profit. The predicted results will be very useful for the executives of the company to know about their sales and profits. This will also give them idea for their new locations or Centre's of Big Mart.

VIII. REFERENCES

[1] Makridakis, S., Wheelwright, S.C., Hyndman, R.J.: Forecasting methods and applications. John wiley & sons (2008).

[2] Kadam, H., Shevade, R., Ketkar, P. and Rajguru.: "A Forecast for Big Mart Sales Based on Random Forests and Multiple Linear Regression." (2018).

[3] C. M. Wu, P. Patil and S. Gunaseelan: Comparison of Different Machine Learning Algorithms for Multiple Regression on Black Friday Sales Data (2018).

[4] Das, P., Chaudhury: Prediction of retail sales of footwear using feed forward and recurrent neural networks (2018)

[5] Das, P., Chaudhury, S.: Comparison of Different Machine Learning Algorithms for Multiple Regression on Black Friday Sales Data (2007)

[6] Tianqi Chen, Carlos Guestrin: XGBoost-A Scalable Tree Boosting System.

[7] Ruihua Liu, Yicen Liu: XGBoost-Based Algorithm Interpretation and Application.