IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

BIG MART SALES PREDICTION USING MACHINE LEARNING

Rohit Sav, Pratiksha Shinde, Saurabh Gaikwad

Student, Department of Computer Engineering, Student, Department of Computer Engineering, Student, Department of Computer Engineering

JSPM's Imperial College Of Engineering and Research, Wagholi, Pune, Maharashtra

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.

I. INTRODUCTION

Big Mart is a big supermarket chain, with stores all around the country and its current board set out a challenge to all Data Scientist out there to help them create a model that can predict the sales, per product, for each store to give accurate results. Big Mart has collected sales data from the year 2013, for 1559 products across 10 stores in different cities.

With this information the corporation hopes we can identify the products and stores which play a key role in their sales and use that information to take the correct measures to ensure success of their business.

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 used Random Forest and Linear Regression 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.

III. OBJECTIVES

- The objective of this framework is to predict the future sales from given data of the previous year's using Machine Learning Techniques.
- Another objective is to conclude the best model which is more efficient and gives fast and accurate result by using XG Boost Regressor.
- 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

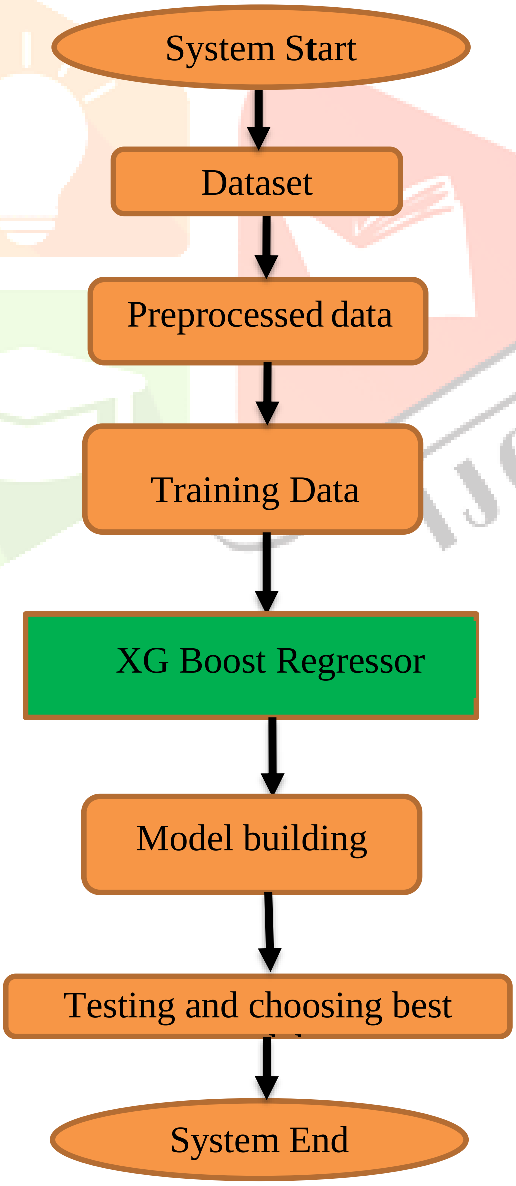


Fig:-Flow of XG Boost Regressor

V. PROPOSED MODEL

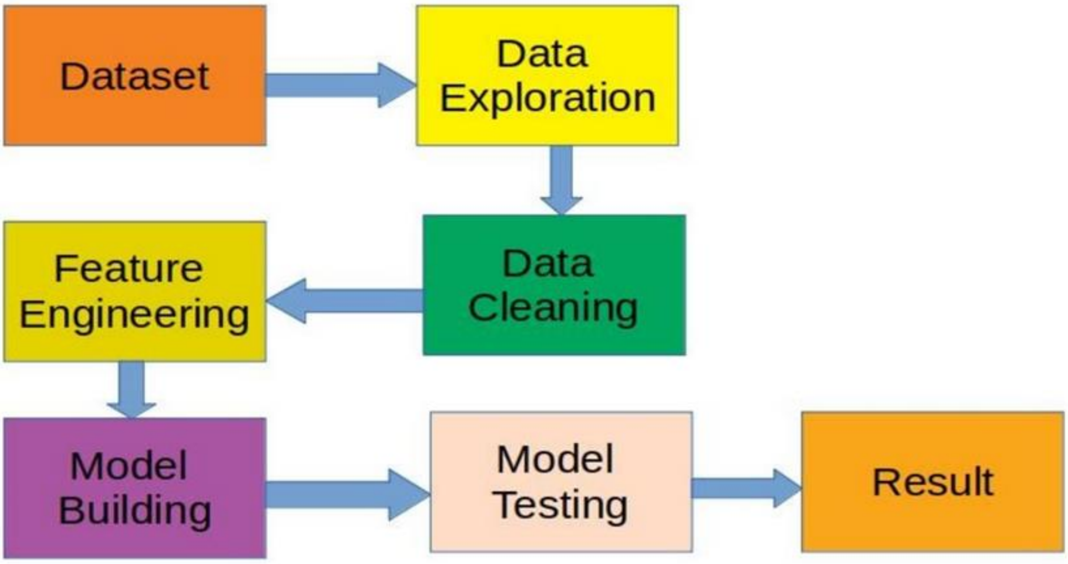
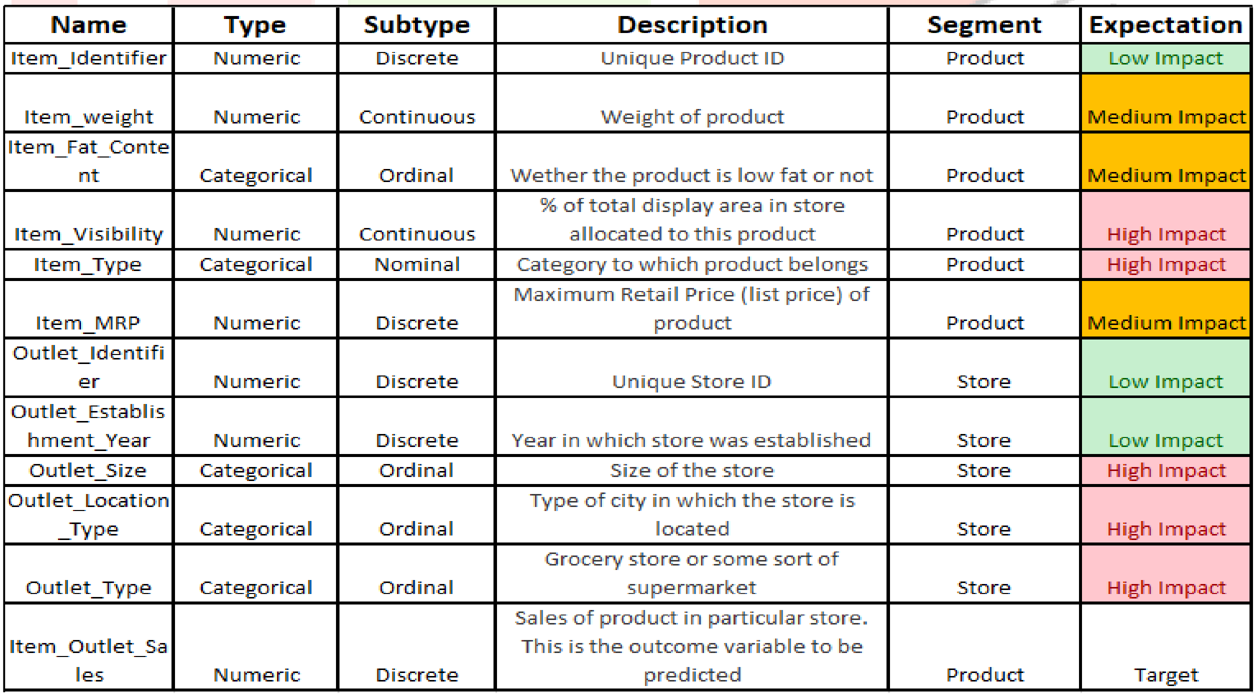


Fig:-Working procedure of proposed model

VI. PHASES IN MODEL

1. DATASET



Name	Type	Subtype	Description	Segment	Expectation
Item_Identifier	Numeric	Discrete	Unique Product ID	Product	Low Impact
Item_weight	Numeric	Continuous	Weight of product	Product	Medium Impact
Item_Fat_Content	Categorical	Ordinal	Wether the product is low fat or not	Product	Medium Impact
Item_Visibility	Numeric	Continuous	% of total display area in store allocated to this product	Product	High Impact
Item_Type	Categorical	Nominal	Category to which product belongs	Product	High Impact
Item_MRP	Numeric	Discrete	Maximum Retail Price (list price) of product	Product	Medium Impact
Outlet_Identifier	Numeric	Discrete	Unique Store ID	Store	Low Impact
Outlet_Establishment_Year	Numeric	Discrete	Year in which store was established	Store	Low Impact
Outlet_Size	Categorical	Ordinal	Size of the store	Store	High Impact
Outlet_Location_Type	Categorical	Ordinal	Type of city in which the store is located	Store	High Impact
Outlet_Type	Categorical	Ordinal	Grocery store or some sort of supermarket	Store	High Impact
Item_Outlet_Sales	Numeric	Discrete	Sales of product in particular store. This is the outcome variable to be predicted	Product	Target

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. Establishment year of Outlet varies from 1985 to 2009. These values may not be appropriate in this form. So, we need to convert them into how old a particular outlet is. There are 1559 unique products, as well as 10 unique outlets, present in the dataset. The attribute Item type contains 16 unique values. Whereas two types of Item Fat Content are there but some of them are misspelled as regular instead of 'Regular' and low fat, LF instead of Low Fat. 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 our 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 our model we are assuming that there is no relationship between the measured attribute and imputed attribute.

Code

```
inp:-  #Check missing values:

       data.apply(lambda x: sum(x.isnull()))

out:-        Item_Fat_Content                0
            Item_Identifier                  0
            Item_MRP                       0
            Item_Outlet_Sales             5681
            Item_Type                        0
            Item_Visibility                0
            Item_Weight                   2439
            Outlet_Establishment_Year        0
            Outlet_Identifier                0
            Outlet_Location_Type             0
            Outlet_Size                   4016
            Outlet_Type                   0
            source                           0
            dtype: int64
```

4. FEATURE ENGINEERING

Some nuances were observed 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 completing the previous phases, 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.

```
inp:- # Reading modified data
        train2 = pd.read_csv("train_modified.csv")
        test2 = pd.read_csv("test_modified.csv")
```

```
inp:-train2.head()
```

In our work, we propose a model using XG boost Regressor algorithm

VI. CONCLUSION

We are predicting the accuracy for XG Boost Regressor. Our predictions help big marts to refine their methodologies and strategies which in turn helps them to increase their profit. The results predicted will be very useful for the executives of the company to know about their sales and profits. This will also give them the idea for their new locations or Centre's of Big-mart.

VII. 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)