# Big mart Sales Prediction


Manikya.Prakash.Sarashetty
B.Tech Student
Computer Science and Engineering
PES university,EC campus
manikyasarashetti@gmail.com

Ramyashree.J.R
B.Tech Student
Computer Science and Engineering
PES university,EC campus
ramyarnayak0511@gmail.com

Nandini.R.Sonth
B.Tech Student
Computer Science and Engineering
PES university,EC campus
nandini.r.sonath@gmail.com



**Abstract**- In day today life malls and Big Marts keep the track of their sales of every individual item for predicting future demand of the customer and to avoid shortage of sales in any season. In this paper, we tried to do a comparative analysis of the model with others for predicting the sales of a company like Big Mart and to find that model which produces better performance as compared to existing models. The final data will be useful to predict future sales with different machine learning techniques which will be useful for the retailers like Big Mart and other industries to increase their profit.



Keywords— Forecast Sales,Machine Learning Algorithms and techniques, Linear regression, random forest , Decision tree and XG booster.


## I. INTRODUCTION

These days, each and every organization wants to uplift its revenue as well as profit. At the end of the day, the main objectives of any organization is to increase sales which leads to increase in profit. In this paper, we are trying to address the problem of big mart sales prediction to satisfy the customer's future demand in different big mart stores across various places and products based on the past records. We are trying to use Different machine learning algorithms like linear regression , random forest, decision tree, XGBoost etc are used for forecasting of bigmart sales.

Machine learning techniques can handle non-linear data and also huge data-set efficiently. To measure the performance of all the models we used Root Mean Squared Error (RMSE) is as an evaluation metric. Here in this project metrics are used as the parameter for accuracy measure of a continuous variable where accuracy is the most important factor to find the best model.

In this paper we worked on some of the above algorithms and tried to come up with a better model to predict sales.

The first and foremost technique used in predicting sale is the statistical methods and techniques, which is also known as the traditional method, but these methods take more time in predicting sales and also these methods could not handle non linear data. So to overcome the problems in traditional methods machine learning techniques are introduced. Generally, in pre-processing of data raw data is converted into useful form of data to gain the better understanding of the data. Data preprocessing involves the following steps

- Data-cleaning.
- Data-transformation.
- Data-reduction.
- Feature Engineering

A standard bigmart sales prediction study can help in deeply analyzing the situations and the conditions previously occurred and then, the inference can be applied about customer purchase, and strengths before setting a budget and marketing plans for the coming years. In other words, sales prediction is based on the availability of resources from the past records. Depth knowledge of past is required for improving and enhancing the likelihood of marketplaces disregarding of any circumstances especially to the external circumstance, which is required to prepare for the upcoming needs for the business enhancement.

## II.Literature survey

To solve our problem of bigmart sales prediction we Referred many research papers from the google scholar source.some of the insights are summarized below:

In paper[1] as we can see they have used Random forest regression approach and Xg booster approach to predict the sales of bigmart. They have also plotted graph to compare different models to predict the accuracy of the model.This boxplot mainly refers to

the mean error produced by testing the algorithm of different models.

They have mainly focused on product level hypothesis.

In paper[2] they have used Xg boost technique to predict bigmart sales.The data-set is also based on hypothesis of store level and product level. After visualizing and exploring the data ,data-set is divided into two parts, train dataset and test dataset in the ratio 80 : 20.
We got to know that the Xgboost has exclusive features like Sparse Aware (that is the missing data values are automatic handled) from this paper.

In paper[3] they have used random forest approach and linear regression model to solve the problem of predicting bigmart sales.We can infer accuracy is the key factor in predictive systems from this paper. There is a diagram which shows correlation between different factors which is very important to know about the variables to proceed with the building of model. The largest size store type did not produce the highest sales, in fact the location that produced largest sales is medium in size.

In this paper, we use comparison methodology in which raw data obtained at large mart will be pre-processed for missing data and outliers . Then machine learning algorithms will be used to predict the final results. ETL stands for Extract, Transform and Load and finally we compare all the models and predict which model gives accurate result to predict the sales of each item in bigmart.

Other attempts-In this paper we are mainly trying to solve which method is best for forecasting sales of bigmart.Many others who have attempted to solve the same problem have tried to propose a model and then compare them with other models to get the correct model with the best accuracy.

# III.Problem Statement

Our data analysis involves around predicting Bigmart Sales based on the characteristics or properties of all attributes. Our dataset contain the previous year's data Using which we would be doing predictive analysis, to forecast the future sales.

"The next question in proceeding our project is to find out what are the characteristics of the items and how they will affect the sales of Bigmart which will be carried out by analyzing and understanding Big Mart sales.".To increase the sales of retailers we need to have a better understanding of the data about what attributes help in increasing sales.

Our next question definitely will be "What kind of variables should we use to predict sales and increase profit margin by precise model? " .Data visualization is helpful to find the useful information from the data through specific methodologies. Specifically, it looks for what kind of variables have impact on pastsales and if those influential variables change when the focus is shifted to sales of three specific regions.

Further we would like to model "The Sales Prediction based on how the attributes influence the sales". On analysis of the influence of each and every attribute we would be considering the most influential variables using data reduction and transformation techniques on the dataset.

## A.Overview of data

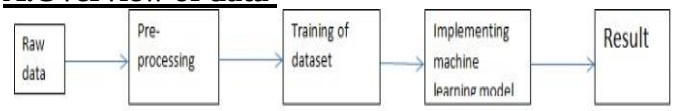



Fig. 1. Steps followed for obtaining results

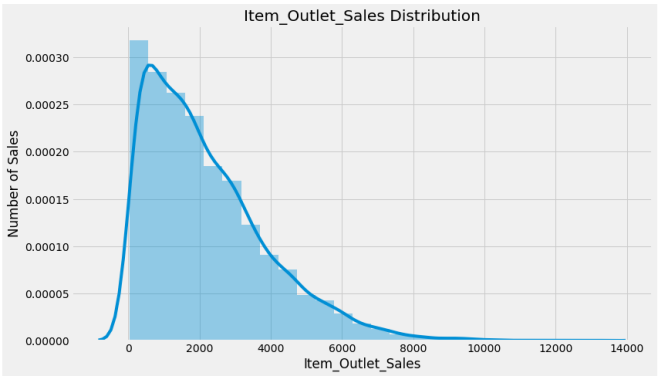
## . B.Data Description

We used bigmarket sales data as a dataset from the kaggle source in our work in which we have 13 attributes (when we combine both train and test dataset as one file). These 13 attributes define the basic features of the data which is being used for prediction. The dataset consists of 13 attributes like:
Item_MRP,Outlet_Identifier,Outlet_Establishment_Year,Outlet_Size,Item_Identifier,Item_Weight,Item_Fat_Content,Item_Visibility,Item_Type,Outlet_Location_Type,Outlet_Type,Item_Outlet_Sales. Out of all these attributes target variable is the Item Outlet Sales attribute.

## C. Data Exploration and Data VisualiZation

In the data exploration phase useful information in the data has been extracted from the dataset. That is trying to identify the information from available data. Which shows that the attributes Outlet size and Item weight and Item outlet Sales have missing values, also the minimum value of Item Visibility is zero which is practically not possible.

As you can infer from the diagram that Item outlet sales is a right skewd variable and it needs some data transformation to treat its skewness.




Data visualization is very important in data analysis As it is dependent on data that it can be done before or after data cleaning. In our case we visualized the data first to better understand the data and its characteristics. When we try to plot the correlation matrix we get to know that Item_Outlet_Sales is our target variable.

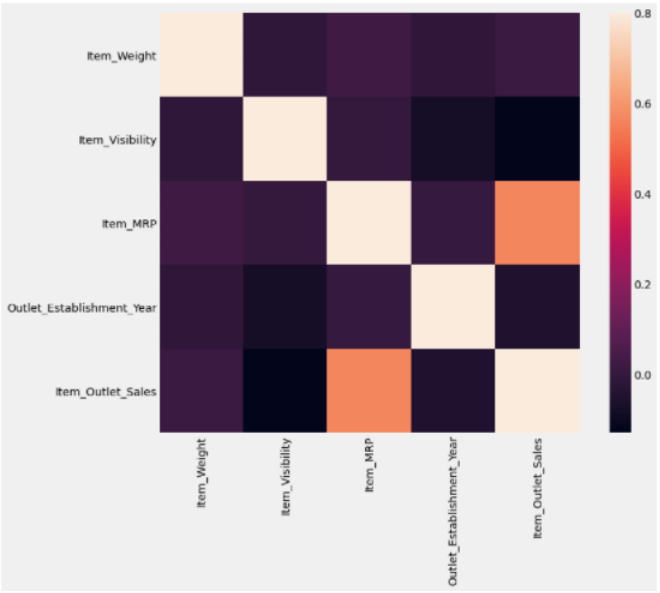
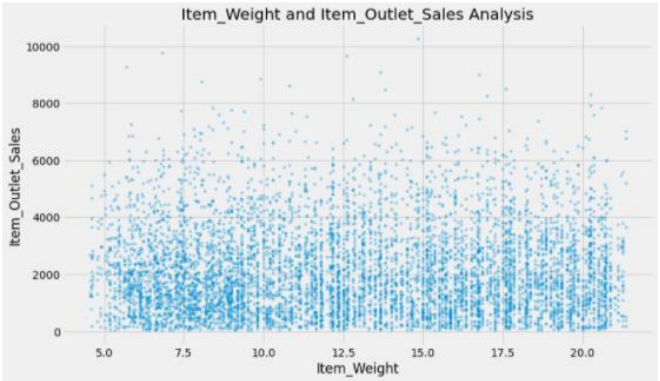



Fig. Correlation among attributes of a dataset. Brown squares show attributes are highly correlated among themselves and black squares show least correlation.

Next when we peek at the data we get to know about many insightful information that how other properties of the data affect the target variable.
Data visualization through graphs gives better knowledge about the data.




Item_Outlet_Sales and Item_Fat_content is spread across all regions.

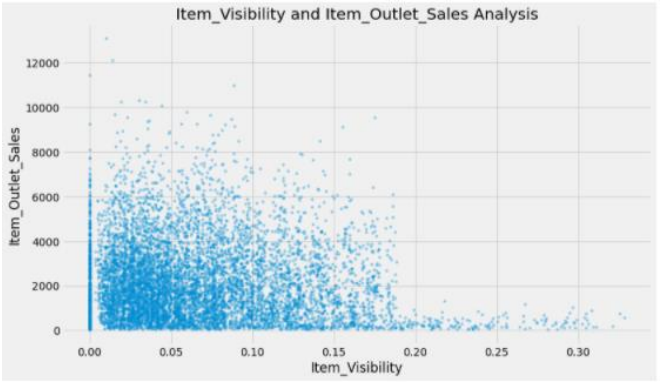



Fig.Impact of item visibility on variable item outlet sales. Which tells us that Less visible items are sold more compared to more visibility items.

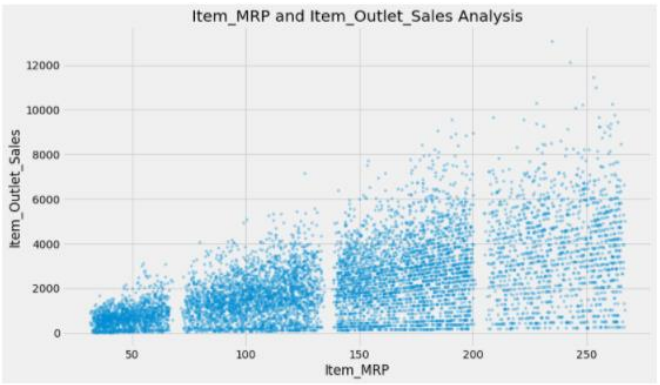
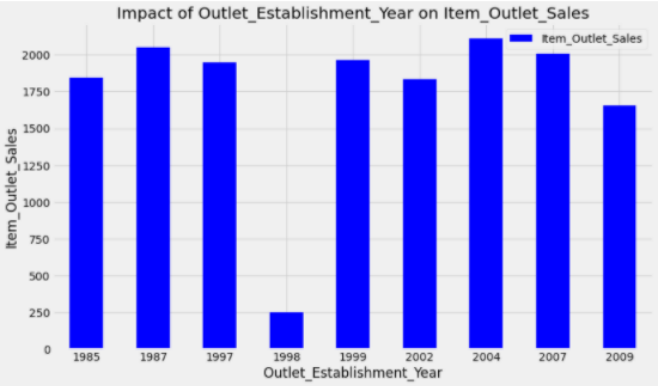
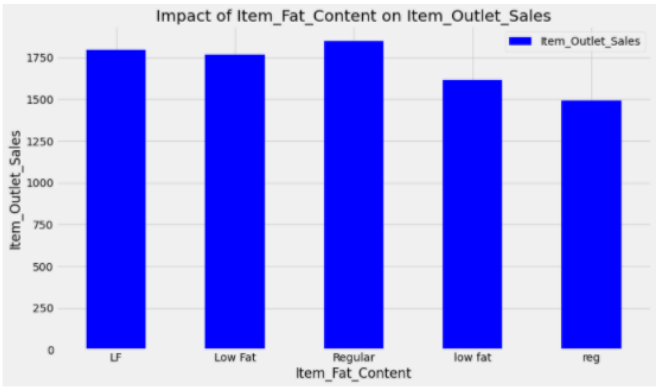
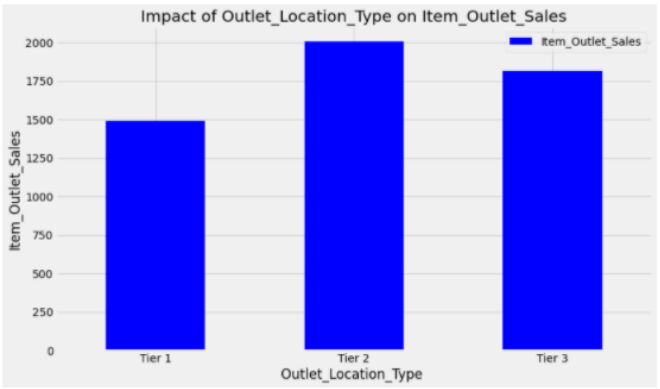


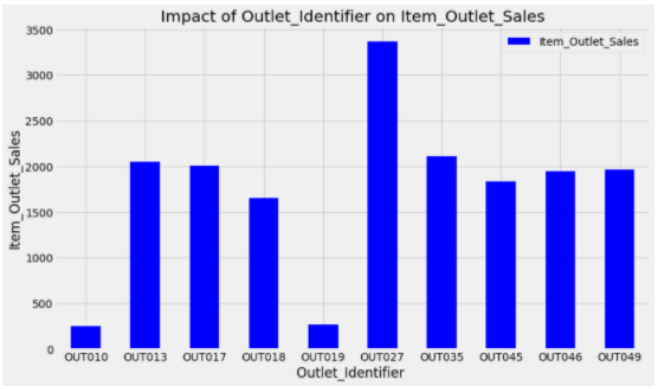

Fig. In the above plotwe can clearly see 4 segments of prices.

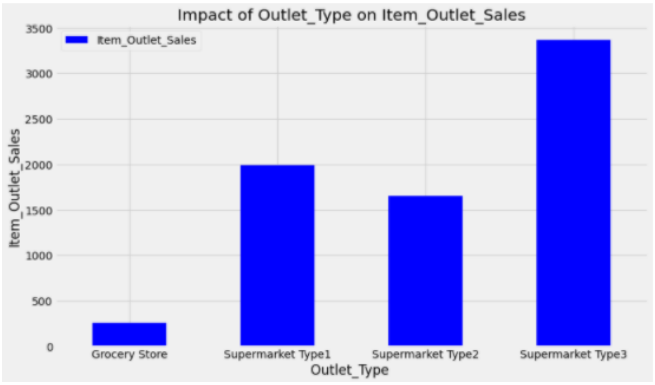


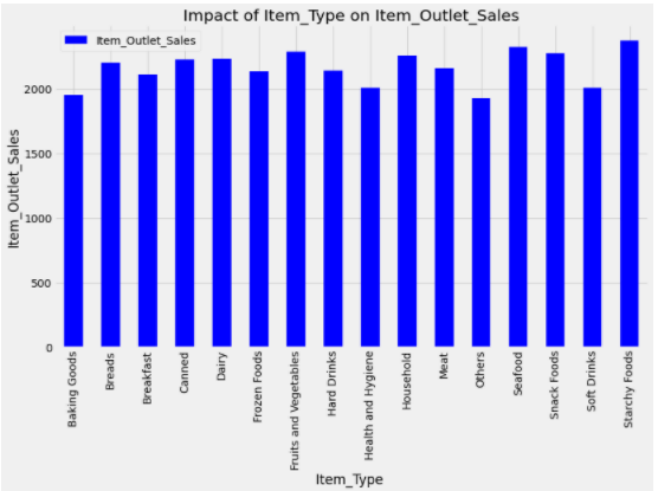

There is lesser number of sales in the year 1998.



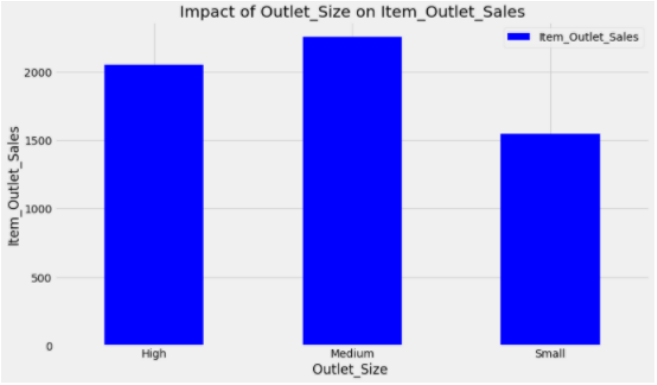




Distribution of target variable Item_Outlet_Sales across the categories of Item_Type and Item_Fat_Content is not very distinct. The distribution for Outlet identifier 010 and 019 are quite similar and very much different from the rest of the categories of Outlet_Identifier.




In the Outlet_Type, Grocery Store has most of its data points around the lower sales values as compared to the other categories. These are the kind of perception that we can draw by visualizing our data.Therefore Data visualization is vital step in data analytics process.

## D.Data Cleaning

It was observed from the data that the attributes Item Outlet Sales,Outlet Size and Item Weight has missing values. In our project in case of Outlet Size missing value we replace it by the mode of that attribute and for the Item Weight and Item Outlet Sales we replace missing values with the mean of that attribute. The missing attributes are numerical.

## E. Feature Engineering

Feature engineering is all about converting cleaned data into predictive models to present the available problem in a better way. During data exploration, some noise was observed. In this phase, this noise is resolved and the data is used for building appropriate model. New features are created to make the model work precisely and effectively. A few features can be combined for the better working of model. Feature engineering phase converts data into a form understandable by the algorithms.We drop the unnecessary columns and divide the data into train and test data.To continue our analysis we export our files to modified versions.

# IV.Implementation and Result

## A. Model Building

As mentioned earlier we use a comparative methodology like we try to build most of the models and compare them. We tried to use models like Linear Regression, Decision Tree model ,Random Forest and Xg booster approach to predict the sales of bigmart.By calculating RMSE(root mean squared error) we will come to conclusion about the best model to predict the sales.

### 1.Linear Regression model

Regression can be termed as a parametric technique which is used to predict a continuous or dependent variable on basis of a provided set of independent variables. This technique is said to be parametric as different assumptions are made on basis of data set.

In simple terms Linear regression algorithm tries to predict the results by plotting the graph between an independent variable and a dependent variable that are derived from the dataset. It is a general statistical analysis mechanism used to build machine learning models. The general equation for linear regression is Z = a + bE Where, Z is the dependent variable and E is independent variable.

After modifying data we filt multiple linear regression to the training set.After predicting the test set results we perform cross validation and predict accuracy.With the linear regression model we got accuracy of 56.36% and RMSE of 1127.

### 2. Decision Tree Model

Decision Tree algorithm technique pertain to the group of supervised learning algorithms. Among different supervised learning algorithms, the decision tree algorithm can be used for solving both type of problems like **regression and classification.**

The goal of the Decision Tree is to create a training model that can be used to predict the value of the target variable by **learning simple decision rules** deduced from the data.

In Decision Tree algorithm, to forecast a class label for a record we start from the **root** of the tree. We differentiate the values of the root attribute with the record's attribute. On comparing we follow the branch corresponding to that value and bounce to the next node. With the decision tree model we got accuracy of 59% and RMSE of 1095.

### 3.Random Forest Model

Random forest also belongs to the group of supervised learning algorithms. The "forest" it builds, is a group of decision trees, usually taught with the "bagging" method. The conclusion of the bagging method is that a combination of learning models will increase the overall effect.
To put it in simple words : random forest assemble multiple decision trees and combines them together to get a more exact and invariable prediction.
With the Random Forest model we got accuracy of 61% and RMSE of 1062.

### 4.Xgboost approach

The gradient boosting decision tree algorithm is taken from XGBoost library.
This algorithm has different names such as gradient boosting, multiple additive regression trees, stochastic gradient boosting or gradient boosting machines.

Boosting is an grouping technique where new models are added to correct the errors made by existing models. Models are added subsequently until no additional improvements can be made. A popular example named as the AdaBoost algorithm that weights data points that are hard to predict.With the Xgboost model we got RMSE of 1031.

Thus we conclude that Xgboost approach is the best model to predict the sales with the least root mean squared error.The RMSE is used as accuracy measure to find the best model.Lower values of **RMSE** indicate better fit.

# V.Conclusion

We started **with** making some hypothesis about the data without looking at it. Then we moved on to data exploration where we found out some nuances **in** the data which required remediation. Next, we performed data cleaning **and** feature engineering, where we imputed missing values **and** solved other irregularities, made new features **and** also made the data model-friendly by one-hot-coding. Finally we made linear regression, decision tree model ,random forest model **and** xgboost got a glimpse of how to tune them **for** better results.

As moving further towards data analysis we come to the point that understanding about the data is really vital to draw the conclusions from the information that the data contain.We got to know about the different

models to forecast sales. As boosting is a grouping model that comes with an easy to read and interpret algorithm, making its predictions and interpretations easy to handle. The prediction or forecast ability is efficient through the use of its clone methods, such as random forest, and decision trees. Boosting is a flexible model and method that restricts over-fitting easily.

One disadvantage of boosting that it makes the model to fail is that it is sensitive to outliers since every classifier is required to fix the errors in the predecessors. Thus, the method is too dependent on outliers. Another disadvantage of the boosting approach is that the method is almost impossible to scale up. This is because every estimator is based on its correctness on the previous predictor sand forecast, thus making the course of action difficult to streamline.

In present era of digitally connected world every shopping mall desires to know the customer demands beforehand to avoid the shortfall of sale items in all seasons. Day to day the companies or the malls are predicting more accurately the demand of product sales or user demands. Large scale research in this area at enterprise level is happening for getting accurate sales prediction. As the profit made by a company is directly proportional and dependent to the accurate predictions of sales. The Big marts are expecting more accurate prediction algorithms so that their company will not suffer any losses. Our predictions help big marts to refine their methodologies and strategies which in turn helps them to increase their profit.We can use the output to help retailers like Bigmart.

## References


[1] IMPROVIZING BIG MARKET SALES PREDICTION, N Meghana, P Chatradi, A Chakravarthy, SM Kalavala, GITAM School of Technology, Bengaluru campus, Karnataka, India (URL: http://xajzkjdx.cn/gallery/423-april2020.pdf)

[2] A Comparative Study of Big Mart Sales Prediction, Gopal Behera, N Nain ,September 2019 ,Conference: 4th International Conference on Computer Vision and Image Processing, At: MNIT Jaipur(url : https://www.researchgate.net/publication/336530068_A_Comparative_Study_of_Big_Mart_Sales_Prediction)

[3] Parichay: Maharaja Surajmal Institute Journal of Applied Research Volume 3, Issue 1; January-June 2020, Sales Prediction Model for Big Mart 22 Nikita Malik, Karan Singh (Url: https://msi-ggsip.org/msijr/pdf/MSIJAR_VOL03_ISSUE01.pdf#page=26)


## Authors


Manikya.Prakash.Sarashetty,
Student ,SRN:PES2201800301 Department of CSE
PES University ,EC Campus

Ramyashree.J.R,
Student ,SRN:PES2201800728, Department of CSE
PES University ,EC Campus

Nandini.R.Sonth,
Student ,SRN:PES2201800401 Department of CSE
PES University ,EC Campus


Contribution of each Team member

1.Manikya Prakash Sarashetty – Data Exploration and Model Building

2.Ramyashree.J.R – Data visualization and Model Building

3.Nandini.R.Sonth- Data cleaning and Model Building

By working on this project we got to know about many things like how to handle the data, the insights we see as consumers are the result of a great deal of work,scraping data and problem solving skills.We will try to take our project to the advanced level to make our proposed solution to be best fit.