Big Mart Sales Prediction

Mini Project Report submitted in partial fulfilment. of the requirement for the degree of **B. E. (Information Technology)**

Submitted By

PRASHIK NIKUMBE (Roll No. 18101A0040)
SOHAM LOHAR (Roll No. 18101A0042)
YASH WAGHMARE (Roll No. 18101A0049)

Under the Guidance of

Prof. Shruti Agrawal

Department of Information Technology

Vidyalankar Institute of Technology

Wadala(E), Mumbai 400 037

University of Mumbai

2021-22

CERTIFICATE OF APPROVAL

For
Mini Project Report
On
R Programming Lab

This is to Certify that

PRASHIK NIKUMBE (Roll No. 18101A0040)

SOHAM LOHAR (Roll No. 18101A0042)

YASH WAGHMARE (Roll No. 18101A0049)

Have successfully carried out Mini Project entitled

"**Big Mart Sales Prediction**"

In partial fulfilment of degree course in

Information Technology

As laid down by University of Mumbai during the academic year 2021-22

Under the Guidance of
Prof. Shruti Agrawal

Signature of Guide

Head of Department

Examiner 1

Examiner 2

Principal
Dr. S. A. Patekar

ACKNOWLEDGEMENT

We would like to express our deepest appreciation to all those who provided us with the possibility to complete this report. We express our profound gratitude we give to our **Prof. Shruti Agrawal** Ma'am, our respectable project guide, for her gigantic support and guidance. Without her counselling, our project would not have seen the light of the day.

We extend our sincere thanks to **Dr. Vipul Dalal**, Head of the Department of Information Technology for offering valuable advice at every stage of this undertaking. We would like to thank all the staff members who willingly helped us. We are grateful to VIDYALANKAR INSTITUTE OF TECHNOLOGY for giving us this opportunity.

The days we have spent in the institute will always be remembered and also be reckoned as guiding in our career.

1. **Prashik Nikumbe**
2. **Soham Lohar**
3. **Yash Waghmare**

Abstract

Nowadays shopping malls and Big Marts keep 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 project, we have carefully observed the data and performed various analyses before using it for training the model. We have trained and tested various predictive models like Linear Regression, Random Forest, etc for predicting the sales of a company like Big Mart and found out which model produces better performance as compared to different models. A comparative analysis of the model with others in terms of root mean square error is also explained in detail.

Table of Contents

Introduction

Day by day competition among different shopping malls as well as big marts is getting more serious and aggressive only due to the rapid growth of the global malls and on-line shopping. Every mall or mart is trying to provide personalised and short-time offers for attracting more customers depending upon the day, such that the volume of sales for each item can be predicted for inventory management of the organisation, logistics and transport service, etc. Present machine learning algorithms are very sophisticated and provide techniques to predict or forecast the future demand of sales for an organisation, which also helps in overcoming the cheap availability of computing and storage systems. In this project, we are addressing the problem of big mart sales prediction items on customer's future demand in different big mart stores across various locations and products based on the previous record. Different machine learning algorithms like linear regression analysis, random forest, etc are used for prediction sales volume. As good sales are the life of every organisation so the forecasting of sales plays an important role in any shopping complex. Always a better prediction is helpful, to develop as well as to enhance the strategies marketplace of business about the marketplace which is also helpful to improve the knowledge of . The basic and foremost technique used in predicting sale is the statistical methods, which is also known as the traditional method, but these methods take much more time for predicting a sale. These methods could not handle non-linear data so to overcome these problems in traditional methods machine learning techniques are deployed. Machine learning techniques can not only handle non-linear data but also huge data-set efficiently.

Problem Definition

Traditional sales techniques are limited to only the available demand history, have less accuracy, focus on few demand factors and have less scope. while Machine Learning prediction models can take advantage of unlimited data, defining what is important, then line up available customer insights to stimulate future demand.

Our Aim is to find out what role certain properties of an item play and how they affect their sales by understanding Big Mart sales. In order to help Big Mart achieve this goal, a predictive model can be built to find out for every store, the key factors that can increase their sales and what changes could be made to the product or store's characteristics. We are also trying to find out which model is better for predicting the sales.

Components

4.1 Hardware Components

- A PC or Laptop with a minimum of 4 GB Ram and 500 GB Hard Disk.

4.2 Software Components

- R Programming Language
- R Studio

Methodology

We have found the big mart sales data on Kaggle and we have pre-processed the data before using it for training and testing the models. The steps we consider in our projects are Exploratory analysis of the data, Data Manipulation, Data Splitting, Training Models, Visualization and results.

Fig 4 : Process flow

The whole project is carried out in the R studio using the R language. We have various libraries in our projects such as dplyr, ggplot2, caret, corrplot, xgboost, cowplot etc. Key uses of libraries is given below,

Dplyr : Used for reading and manipulation and joining of data.

Ggplot2 : Used for plotting

Caret : Used for modelling

Corrplot : Used for making correlation plot

Xgboost : Used for building XGBoost model

Cowplot : Used for combining multiple plots

We have used several machine learning models in our projects. The five models we have trained and tested are Linear Regression, Lasso Regression, Ridge Regression, Random Forest, Xgboost. We have considered the RMSE for comparison of the models. We have also plotted the line graph showing the actual vs predicted sale values.

Result and Discussion

Code : https://github.com/PrashikNikumbe/R-project

Dataset : https://www.kaggle.com/brijbhushannanda1979/bigmart-sales-data

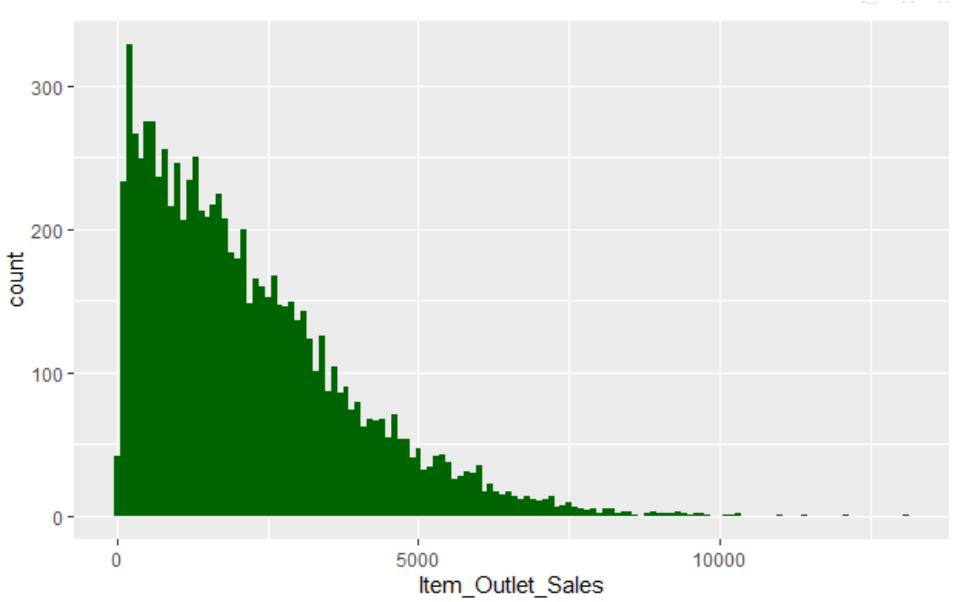

Fig 5.1 : Histogram of Item outlet sales

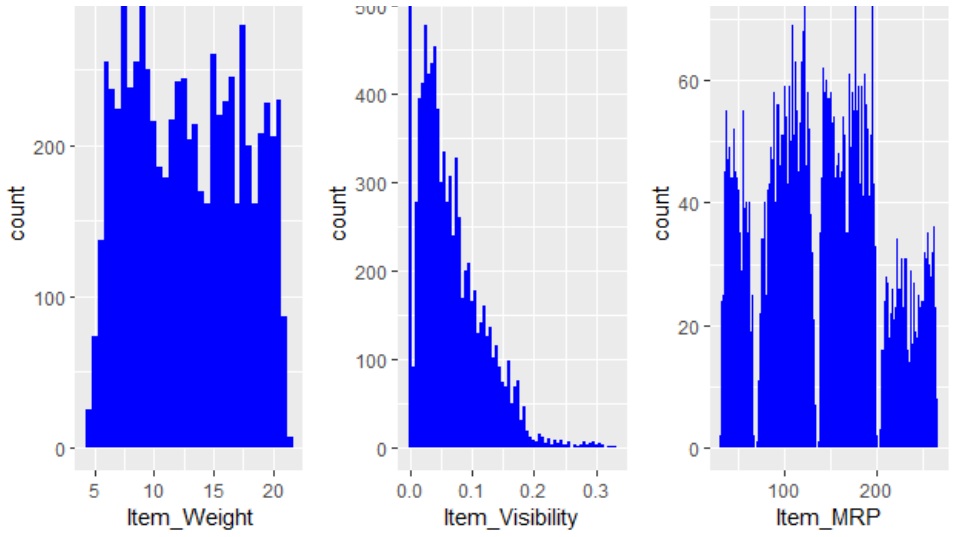

Fig 5.2 : Histograms of other numeric variable

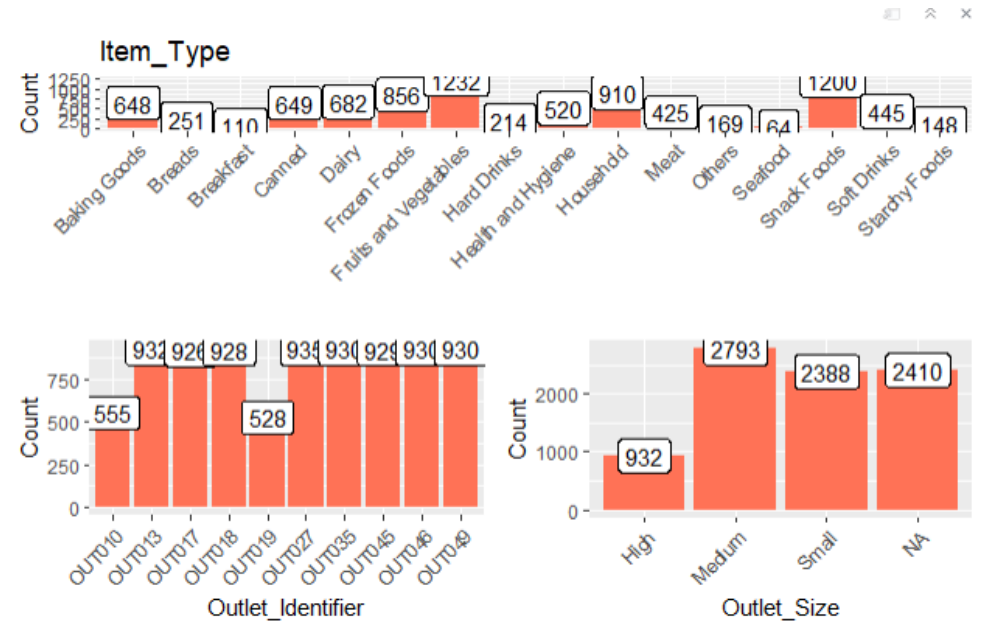

Fig 5.3 : Bar Graph other categorical variable

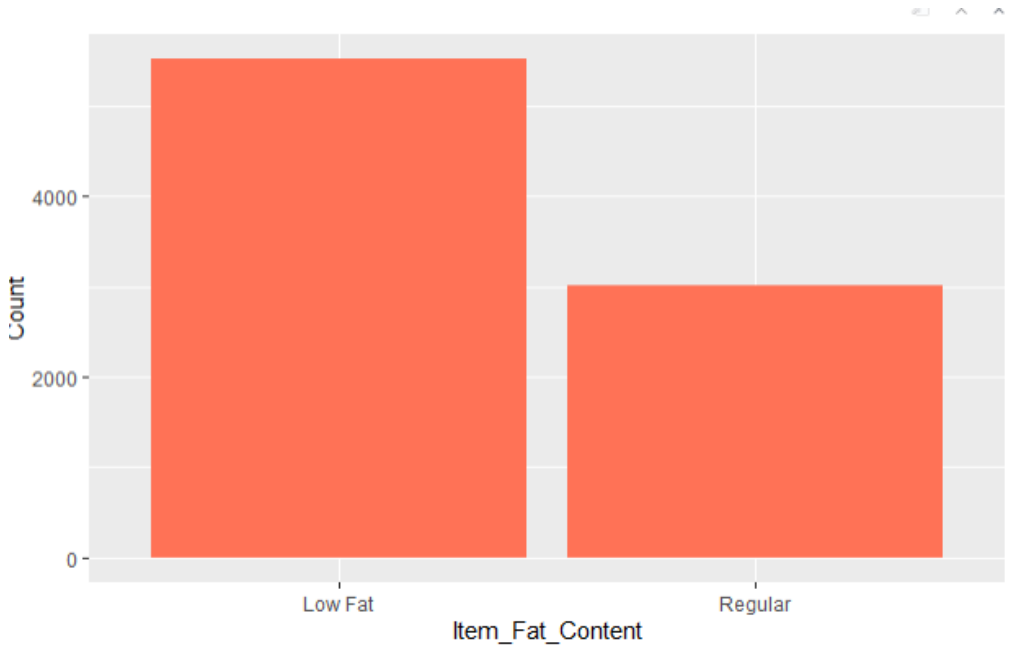

Fig 5.4 : Bar Graph of Item fat content

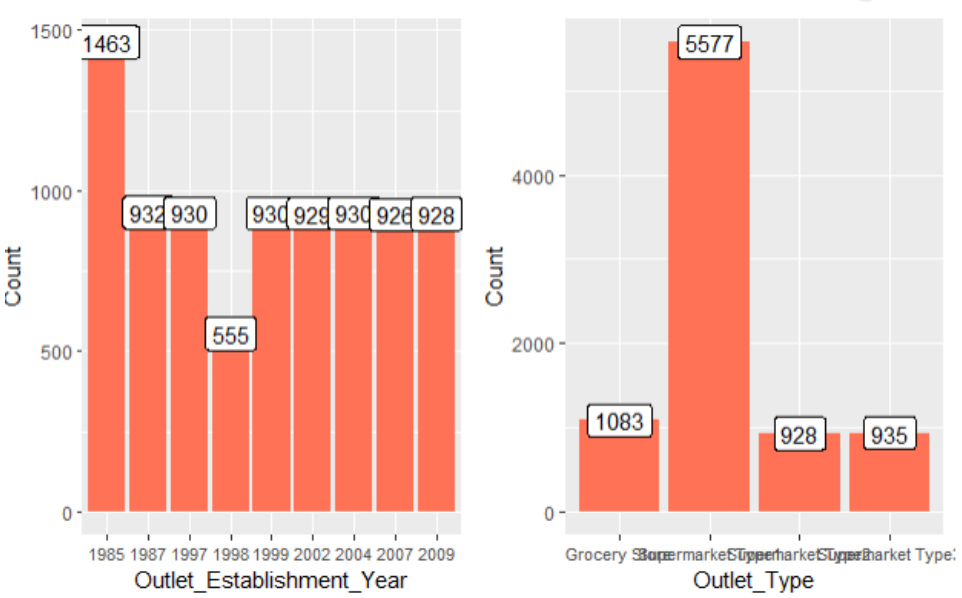

Fig 5.5 : Bar Graph of remaining variable

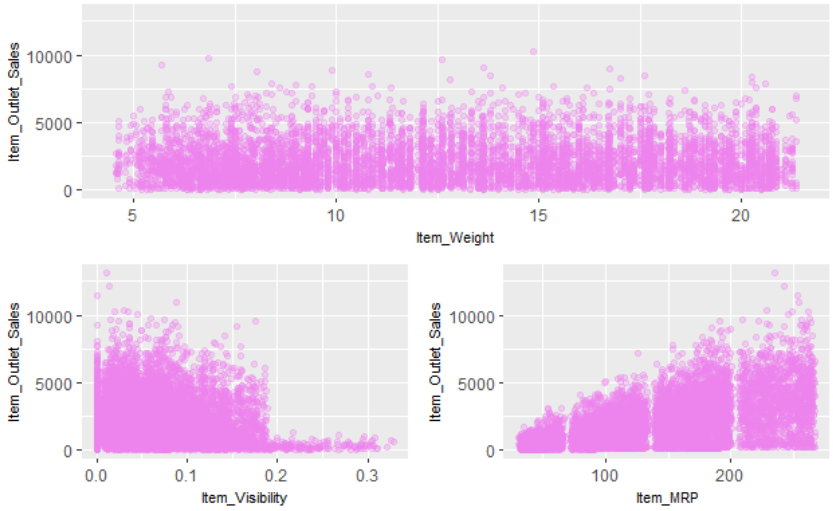

Fig 5.6 : Scatter plot of target vs numeric variable

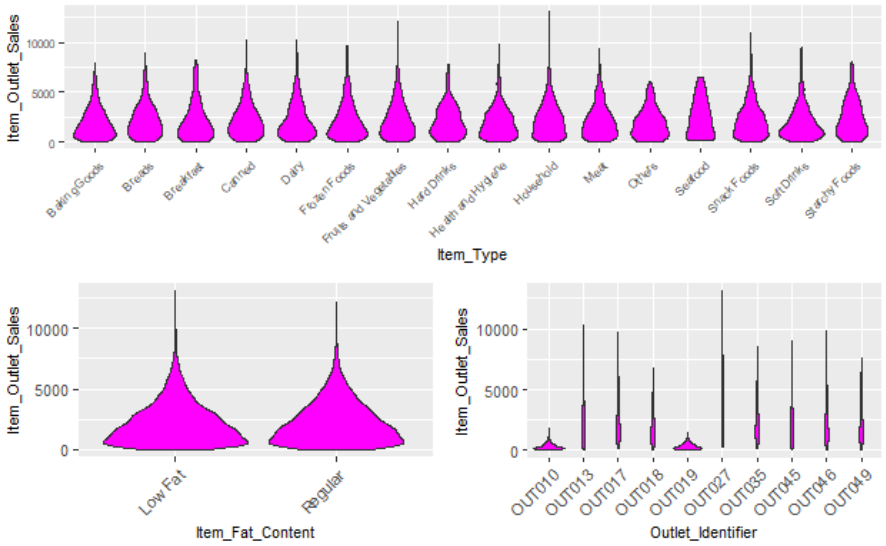

Fig 5.7 : Violin graph of target vs categorical variable

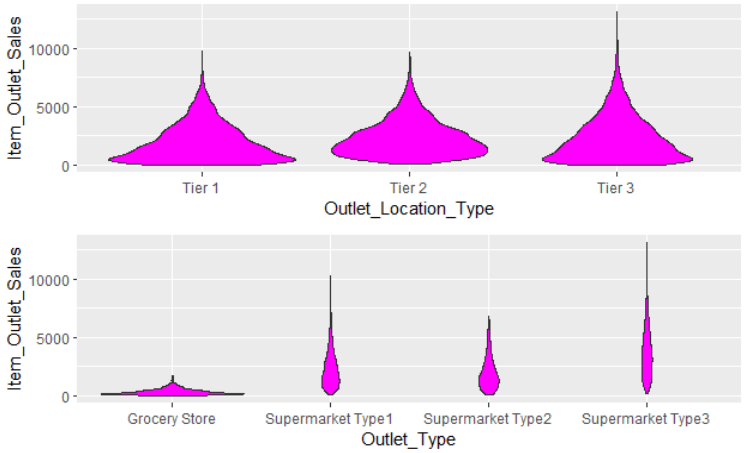

Fig 5.8 : Violin graph of target vs remaining variable

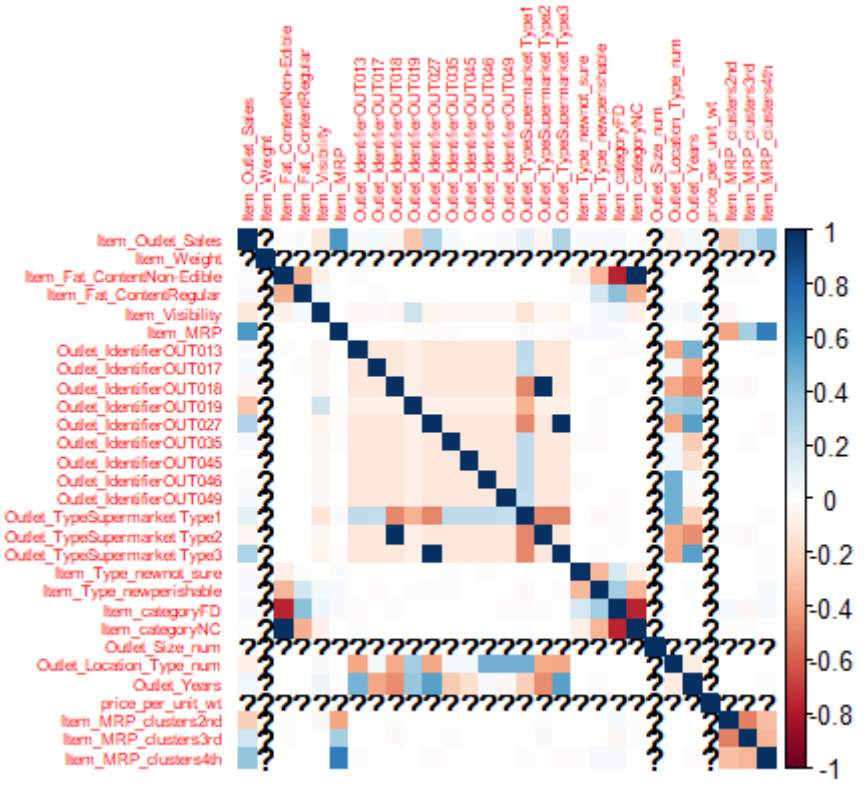

Fig 5.9 : Correlation graph

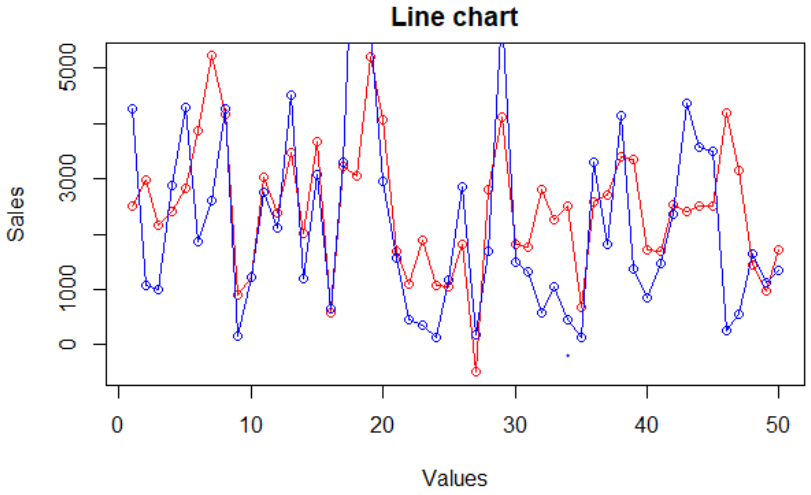

Fig 5.10 : Linear regression prediction

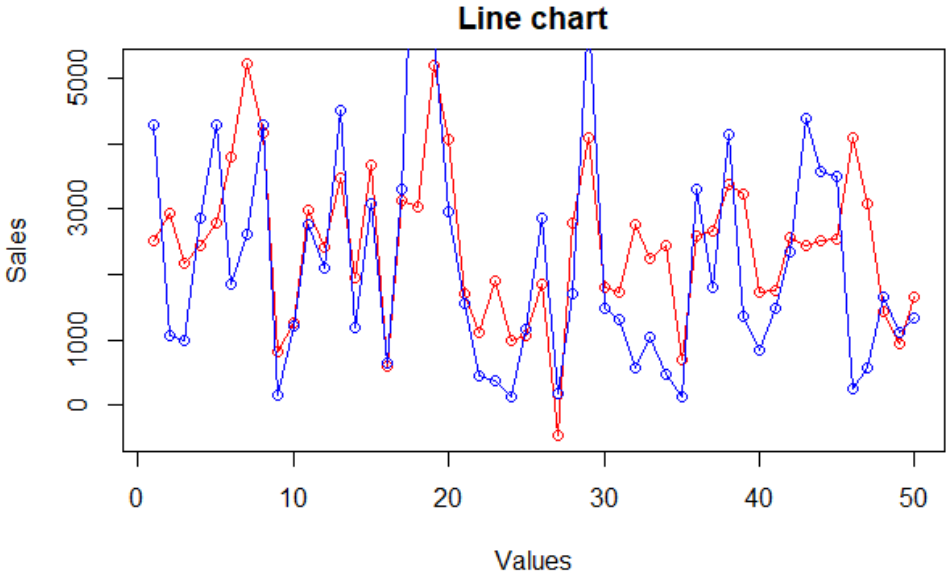

Fig 5.11 : Lasso regression prediction

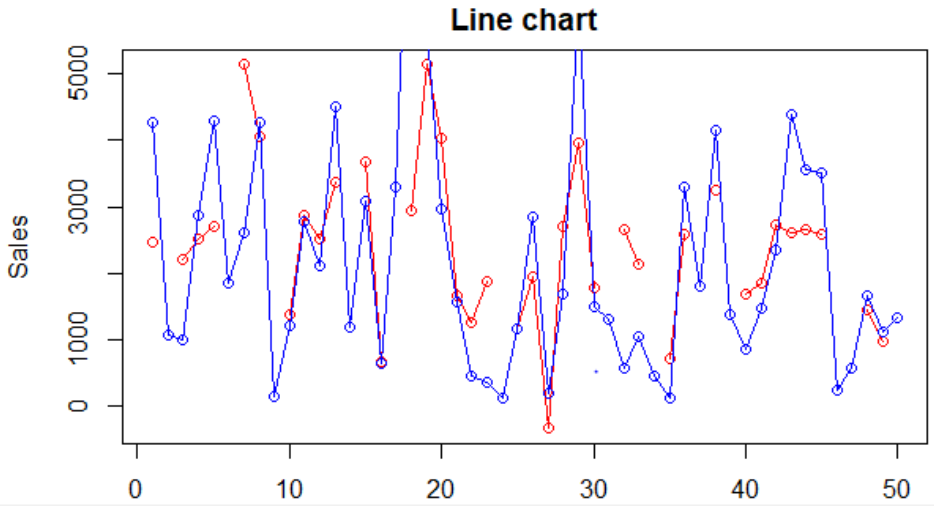

Fig 5.12 : Ridge regression prediction

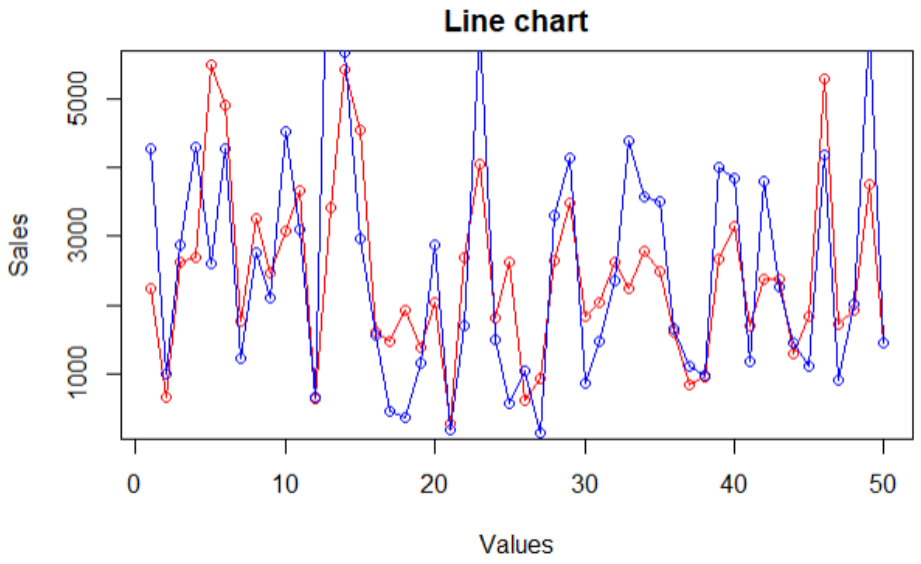

Fig 5.13 : Random Forest prediction

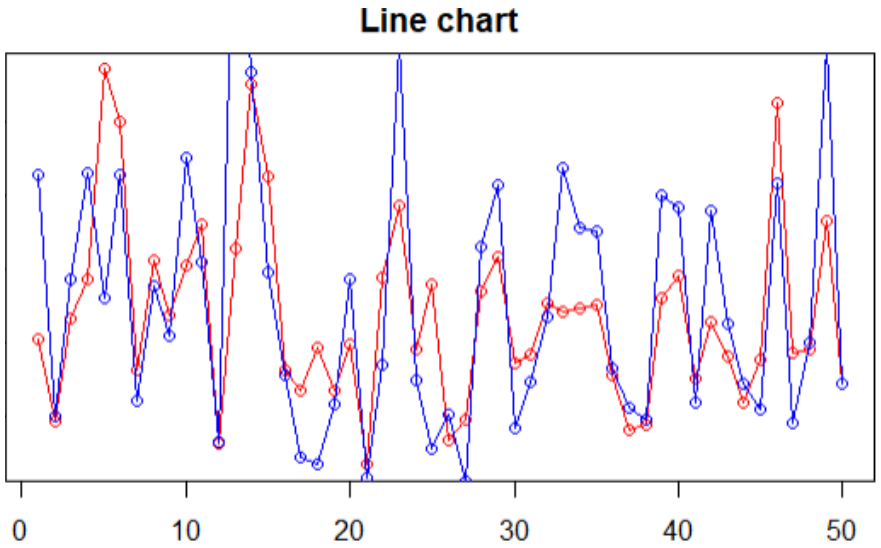

Fig 5.12 : XGboost prediction

Models	**RMSE Value**
Linear Regression	1257
Lasso Regression	1249
Ridge Regression	1188
Random Forest	1158
XGBoost	1169

Table 5 : RMSE comparison

Upon testing the above five models we found out that Random Forest attains lowest RMSE value means it is good in prediction as compared to others.

Conclusion

Day to day the companies or the malls are predicting more accurately the demand of product sales or user demands. Extensive research in this area at enterprise level is happening for accurate sales prediction. As the profit made by a company is directly proportional to the accurate predictions of sales, the Big marts are desiring a more accurate prediction algorithm so that the company will not suffer any losses. By reviewing all the models used in the project, we found out that the Random Forest gives better results which can be implemented in other systems for predicting the sales.

\

References

[1] Behera, Gopal & Nain, Neeta.” A Comparative Study of Big Mart Sales Prediction”,ResearchGate.

[2] Rohit Sav, Pratiksha Shinde, Saurabh Gaikwad , “Big Mart Sales Prediction using Machine learning ”, IJCRT

[3] Nikita Malik, Karan Singh, “Sales Prediction Model for Big Mart”, Maharaja Surajmal Institute Journal of Applied Research.

[4] https://courses.analyticsvidhya.com/courses/big-mart-sales-prediction-using-r

[5] https://www.geeksforgeeks.org/root-mean-square-error-in-r-programming/