BIG MART SALES PREDICTION USING MACHINE

LEARNING

## A PROJECT REPORT

*Submitted to*



# MAHARSHI DAYANAND UNIVERSITY, ROHTAK

***In partial fulfillment for the award of the degree of***

# BACHELOR OF TECHNOLOGY

***in***

**Computer Science & Engineering**

***Submitted by***

# DEEPAK

## ROLL NO: 21/610

***Under the supervision of***

**----------------- LECTURER**

**DEPARTMENT OF CSE**



**FACULTY OF ENGINEERING, J B KNOWLEDGE PARK, VILLAGE-MANJHAWALI, FARIDABAD**

JUNE 2024

**DEPARMENT OF COMPUTER SCIENCE & ENGINEERING FACULTY OF ENGINEERING,**

**J B KNOWLEDGE PARK, VILLAGE-MANJHAWALI, FARIDABAD**

# BONAFIDE CERTIFICATE

Certified that this project report **BIG MART SALES PREDICTION USING**

**MACHINE LEARNING** is the bonafide work of **Deepak** who carried out the project work undermy supervision.

**Signature:**

**Name of Supervisor: Designation: Department:**

**Signature:**

**Name of Project Coordinator:**

**Head of Department**

# TABLE OF CONTENTS

**CONTENT PAGE NO.**

Certificate (I)

Acknowledgement (II)

Table of Content (III)

List of Abbreviations (V)

Abstract (VI)

Candidate’s Declaration (VII)

**CHAPTER 1: INTRODUCTION**

* 1. [Introduction 01](#_bookmark0)
  2. [Problem Statement 02](#_bookmark1)
  3. [Objectives 03](#_bookmark2)
  4. Methodology Used 04

CHAPTER 2: LITERATURE SURVEY

* 1. [Literature survey 05](#_bookmark3)

[CHAPTER 3: SYSTEM DEVELOPMENT](#_bookmark4)

* 1. [Algorithms employed 07](#_bookmark5)
  2. [Phase of model 12](#_bookmark6)

CHAPTER 4: PERFORMANCE ANALYSIS

* 1. Performance comparison 40

CHAPTER 5: CONCLUSIONS

* 1. Conclusions 44
  2. Future Scope 44

**LIST OF ABBREVIATIONS**

|  |  |
| --- | --- |
| **ABBREVIATION** | **WORD** |
| Ml | MACHINE LEARNING |
| EDA | EXPLORATORY DATA ANALYSIS |
| FIG | FIGURE |
| DESCR | DESCRIPTION |
| RF | RANDOM FOREST |
| LR | LINEAR REGRESSION |
| CSE | COMPUTER SCIENCE AND ENGINEERING |

**LIST OF FIGURES**

### DESCRIPTION PAGE NO.

12

Fig 3.5 Depicting the feature of the dataset 12

13

Fig 3.6 How libraries, train and test datasets are imported 13

13

Fig 3.7 Head function representing first five dataset 13

14

Fig 3.8 Description of dataset using info() function

15

Fig 3.9 Description of dataset using describe() function

16

Fig 3.10 Depicts the number of missing values

17

Fig 3.11 Datatype of various features of dataset

18

Fig 3.12 Missing value in outlet\_size column=2410

18

Fig 3.13 Filling values in outlet\_size

19

Fig 3.14 No missing values in item\_weight and outer\_size

Columns. 20

Fig 3.15 Representing how to import dtale library 21

Fig 3.16 The Dtale window 25

Fig 3.21 Color-encoded correlation matrix 27

Fig 3.23 Correlation between different features 28

Fig 3.24 Cleaning the data using Klib library 29

Fig 3.26 Represents the 12 features of the dataset 29

Fig 3.27 Converting to more efficient data types. 31

Fig 3.28 Label encoding code 31

Fig 3.29 Splitting of data into train and test dataset 32

Fig 3.30 Standardization of dataset 32

Fig 3.31 X\_train\_std array and X\_test\_std array 33

Fig 3.32 Y\_train array and Y\_test array 35

Fig 3.35 Value of R2 value in LR 35

Fig 3.36 Value of R2 value in RF 37

Fig 3.38 Value of R2 value in XGBoost Regression 38

Fig 3.39 Value of R2 value in Decision Tree 39

Fig 3.40 Value of R2 value in Ridge Regression

Fig 4.1 Performance of LR 40

Fig 4.2 Performance of RF 41

Fig 4.3 Performance of Hyper Tuning Parameter 41

Fig 4.4 Performance of Decision Tree 42

Fig 4.5 Performance of XGBoost Regression 42

Fig 4.6 Performance of Ridge Regression 43

# LIST OF GRAPHS

### DESCRIPTION PAGE NO.

Fig 1.1 Process of building a model 03

Fig 1.2 Working procedure of proposed model 03

Fig 3.1 Figure represent line of regression 08

Fig 3.2 Flowchart of RF 09

Fig 3.3 Relationship between Feature Importance and their 10

F score in Hyper Parameter Tuning

Fig 3.4 Types of XGBoost Regression 11

Fig 3.17 Frequency of values in the columns 22

Fig 3.18 Item\_weight value range 23

Fig 3.19 Categorical dat plot using Klib library 24

Fig 3.20 Feature correlation using Klib library 25

Fig 3.22 Distribution plot for every numeric feature 26

Fig 4.7 Comparison of RMSE and MSE values 45

Fig 4.8 Comparison of R2 and MAE values 46

# LIST OF TABLE

### DESCRIPTION PAGE NO

Table 4.1 Algorithms Performance 44

# ABSTRACT

Nowadays many shopping malls keep track of individual item sales data in order to forecast future client demand and adjust inventory management. In order to be ahead of the competition and earn more profit one needs to create a model which will help to predict and find out the sales of the various product present in the particular store.So to predict out the sales for the big mart one need to use the very important tool i.e. Machine Learning (ML). ML is that field of computer science which gives machines ie computers the ability to learn without doing any type of programming.Using the concepts of machine and basics of data science one can build a model which can help to predict the sales of the big mart.Because of increasing competition among various shopping complex one needs to have some predictive model which could help to gain some useful insights so as to maximize the profit and be ahead of the competitors.

**Chapter 1 INTRODUCTION**

### INTRODUCTION

The daily competition between different malls as well as big malls is becoming more and more intense because of the rapid rise of international supermarkets and online shoppings. Every mall or mart tries to provide personal and short-term donations or benefits to attract more and more customers on a daily basis, such as the sales price of everything which is usually predicted to be managed through different ways such as corporate asset management, logistics, and transportation service, etc. Current machine learning algorithms that are very complex and provide strategies for predicting or predicting long-term demand for a company's sales, which now also help in overcoming budget and computer programs.

In this report, we basically discuss the subject of specifying a large mart sale or predicting an item for a customer’s future need in a few supermarkets in various locations and products that support the previous record. Various ML algorithms such as linear regression, random forest, etc. are used to predict sales volume. As we know, good marketing is probably the lifeblood of all organizations, so sales forecasting now plays an important role in any shopping mall. It is always helpful to predict the best, and develop business strategies about useful markets and to improve market knowledge. Regular sales forecasting research can help in-depth analysis of pre-existing conditions and conditions and then, assumptions are often used in terms of customer acquisition, lack of funding, and strength before setting budgets and marketing plans for the coming year.

In other words, sales forecasts are predicted on existing services of the past. In-depth knowledge of the past is required to develop and enhance market opportunities no matter what the circumstances, especially the external environment, which allows to prepare for the future

needs of the business. Extensive research is ongoing in the retailer’s domain to predict long- term sales demand. An important and effective method used to predict the sale of a mathematical method, also called the conventional method, but these methods take more time to predict sales. And these methods could not manage indirect data so to overcome these problems in traditional methods the machine learning techniques used. ML methods can handle not only indirect data but also large data sets well.

### PROBLEM STATEMENT

Due to increasing competition many malls and bigmart are trying their best to stay ahead in competition.In order to find out what are the various factors which affect the sales of bigmart and what strategies one needs to employ in order to gain more profit one need to have some model on which they can rely .So a predictive model can be made which could help to gain useful information and increase profit.

### OBJECTIVES

Objectives of these project are:

1. Predicting future sales from a given dataset.
2. To understand the key features that are responsible for the sale of a particular product.
3. Find the best algorithm that will predict sales with the greatest accuracy.

### METHODOLOGY

Figure 1.1 represents the steps of building a model. Following are the steps which one needs to follow while creating a model.

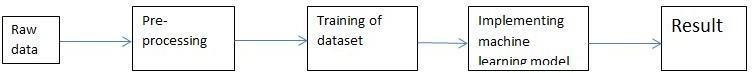


Fig 1.1: Process of building a model.

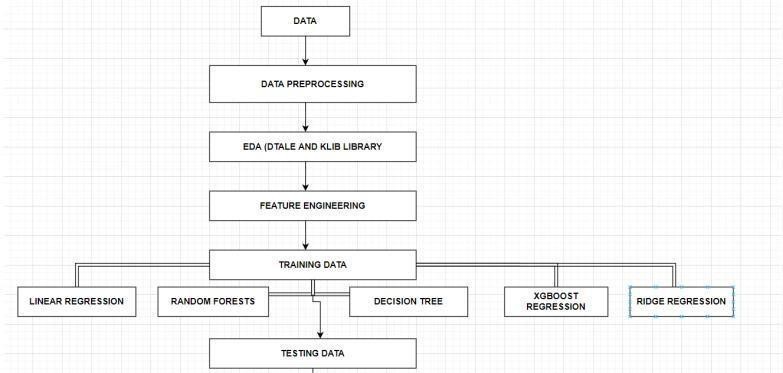


Fig 1.2:Working procedure of proposed model

* + 1. Data collection- The step of every project is to collect the data.

We collected our data from the Kaggle whose link is given below- <https://www.kaggle.com/brijbhushannanda1979/bigmart-sales-data/code>

* + 1. Data preprocessing-In this step we basically clean our dataset for example check for any missing value in the dataset , if present then handle the missing values. In our dataset attributes like Item Weight and Outlet Size had the missing value.
    2. EDA-This part is considered as one of the most important parts when it comes to data analysis.To gain important insights of our data one must need to do exploratory data analysis.Here in our project we used two libraries i.e. klib and dtale library.
    3. Tested various algorithms-Then various algorithms like simple LR, xgboost algorithm were applied in order to find out which algorithm can be used to predict the sales.
    4. Building the model -After completing all the previous phases which are mentioned above, now our dataset is ready for further phases that is to build the model.

Once we built the model now it is ready to be used as a predictive model to forecast sales of Big Mart.

* + 1. Web deployment-Finally once the prediction can be made for making it more user friendly we have used web development.

## . CHAPTER 2

**LITERATURE SURVEY**

### LITERATURE SURVEY

Kadam,et.al [1] have suggested when the prediction for the sales for bigmart was done using the algorithm like random forest and LR for prediction analysis it gave lesser accuracy.So to overcome this problem we can use another algorithm which is XG boost algorithm which not only gives better accuracy but also is more efficient.

Makridakis, et.al [2] have suggested predicting methods and applications containing Data Lack and short life cycles. So some data like historical data, consumer-focused markets face uncertain needs, which can be an accurate predictor of outcome.

C. M. Wu , et.al [3] have suggested comparison of Different ML Algorithms for Multiple Regression on Black Friday Sales Data used the concept of neural network to compare the various different algorithms.Using neural network as the concept which is very complex and less efficient concluded that we should use much simpler algorithm for the prediction purpose.

Das, et.al [4] have suggested in the prediction of retail sales of footwear which used recurrent Neural Networks and feed forward used the neural network to predict the sales.Using neural network for predicting the sales which is not an efficient method so XGboost algorithm can be used.

S. Cheriyan, et.al [5] have suggested in the study they implemented three ML algorithms on the given dataset and the models for evaluating the performance. Based upon the testing the algorithm which gave maximum accuracy was chosen for the prediction which was found to be a gradient boosting algorithm.

A. Krishna, et.al[6] have suggested that both the normal regression and boosting algorithms were implemented and found out that boosting algorithms have better results than the regular algorithms.

## CHAPTER 3 SYSTEM DEVELOPMENT

### ALGORITHMS EMPLOYED

* + 1. **LINEAR REGRESSION (LR)**

As we know Regression can be termed as a parametric technique which means we can predict a continuous or dependent variable on the basis of a provided datasets of independent variables.

The Equation of simple LR is:

Y = βo + β1X + ∈ (1)

where,

Y : It is basically the variable which we used as a predicted value. X : It is a variable(s) which is used for making a prediction.

βo : It is said to be a prediction value when X=0.

β1 : when there is a change in X value by 1 unit then Y value is also changed. It can also be said as slope term ∈

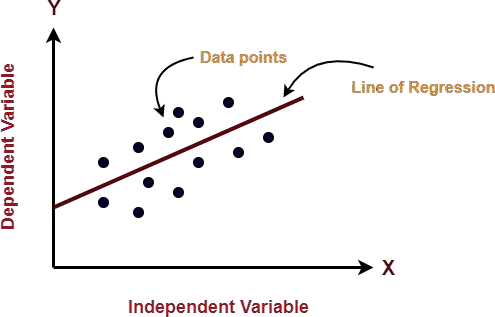


Fig 3.1 Given figure represent line of regression

### RANDOM FOREST REGRESSION

Random Forest is a tree-based bootstrapping algorithm based on that tree that includes a certain number of decision trees to build a powerful predictive model. Individual learners, a set of random lines and a randomly selected few variables often create a tree of choice. The final prediction may be the function of all predictions made by each learner. In the event of a regression. The final prediction may be the meaning of all the predictions.

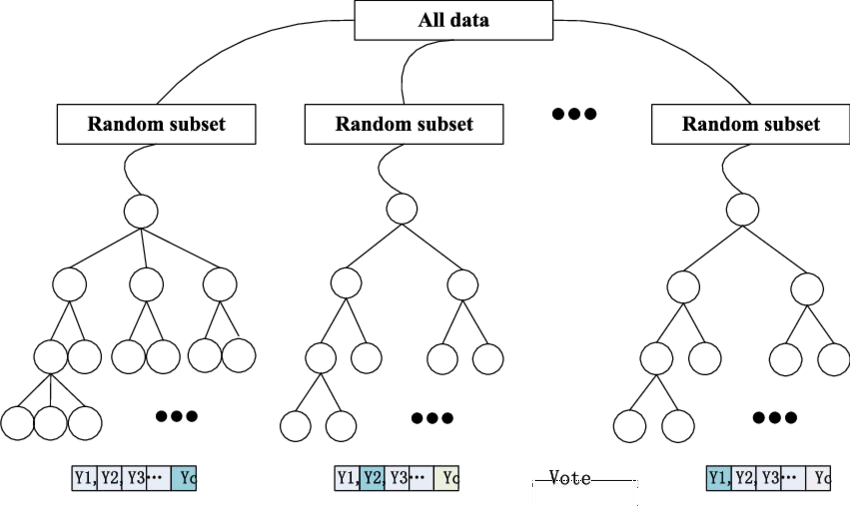


Fig 3.2 : Flowchart of Random Forest Regression

### HYPER PARAMETER TUNING

In ML, optimization of the hyperparameter or problem solving by selecting the correct set of parameters for the learning algorithm. To control the learning process a hyperparameter parameter value is used.In contrast, the values of some parameters are calculated.

The same type of ML model may require different types of weights, learning scales or constraints in order to make different data and information patterns more general. The steps are also called hyperparameters and must be used for the model to solve the ML problem.

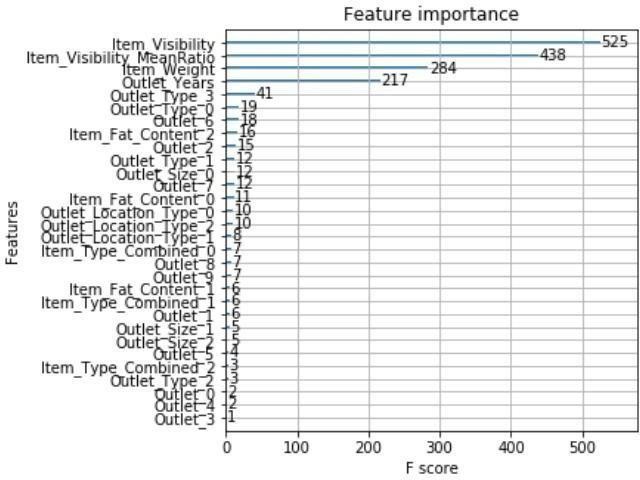


Fig 3.3: Relationship between Feature Importance and their F score in Hyper parameter tuning

### XGBOOST REGRESSION

XGBoost stands for eXtreme Gradient Boosting. The implementation of an algorithm designed for the efficient operation of computer time and memory resources. Boosting is a sequential process based on the principle of the ensemble. This includes a collection of lower learners as well improves the accuracy of forecasts.No model prices n heavy for any minute t, based on the results of the previous t-speed. Well-calculated results are given less weight, and the wrong ones are weighed down. With this algorithm system

The XGBoost model uses stepwise, ridge regression internally, automatically selecting features as well as deleting multicollinearity.

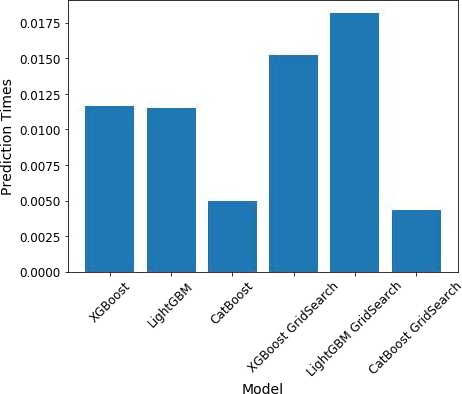


Fig 3.4 : Represents the types of XGBoost regression

### PHASE OF MODEL

* + 1. **DATA AND ITS PREPROCESSING**

In our work, we have used the 2013 Big Mart sales data as a database. Where the data set contains 12 features such as Item Fat, Item Type, MRP Item, Output Type, Object Appearance, Object Weight, Outlet Indicator, Outlet Size, Outlet Year of Establishment, Type of Exit, Exit Identity, and Sales. In these different aspects of responding to the Item Outlet Sales features as well, the other features are also used as the predictive variables. Our dataset has in total 8523 products in various regions and cities. The data set is also based on product level and store- level considerations . Where store level includes features such as city, population density, store capacity, location, etc. and product-level speculation involves factors such as product, ad, etc. After all considerations, a data set is finally created, then the data set is split into two parts that are tested and trained in a ratio of 80:20.



Fig 3.5: Depicting the features of the dataset

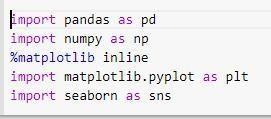


Fig 3.6 : How libraries, train and test datasets are imported.

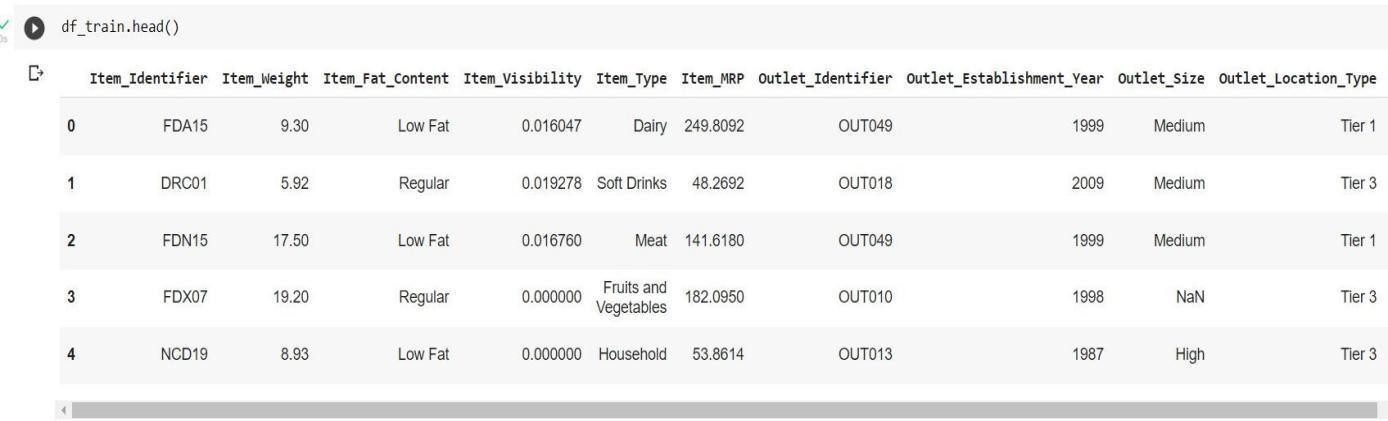


Fig 3.7 Head function representing first five dataset

Item\_Visibility has a value = 0 as values which have no meaning, Item\_Identifier is a character string with some specific code used by the bigmart and Outlet\_Size contains some missing values as well.

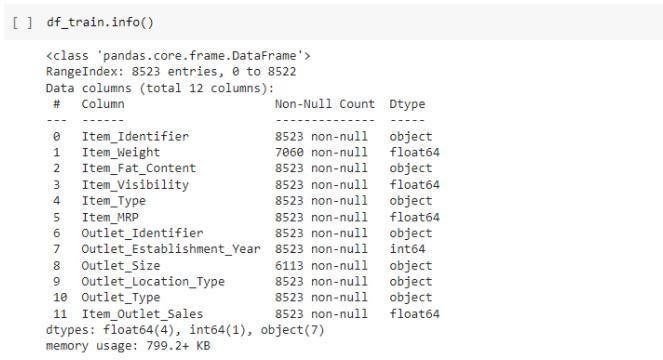


Fig 3.8 : Description of dataset using info() method

In figure 3.8 we can clearly see that there are in total 12 features out of which Numeric data count is 5 and Categorical data count is 7.



Fig 3.9 : Description of dataset using describe() method

In figure 3. Item\_Visibility feature has a minimum value of 0.00 and Item\_weight has count of 7060.

### HANDLING MISSING VALUES

While analyzing the dataset we come across some missing values in the dataset.In order to check for the missing value we have the following code-

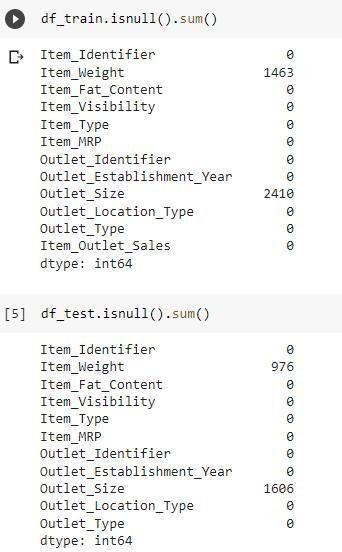


Fig 3.10 Depicts the number of missing value

From the above Fig 3.10 we can clearly see that column names item\_weight and outlet\_size have 976 and 1606 missing values respectively.

In order to handle these missing values we have different approaches for e.g. dropping the rows having missing value or filling the missing value with suitable values using different methods. Looking at our dataset we have 8523 rows so dropping would not be a better option as it would lead to decrease the prediction accuracy.

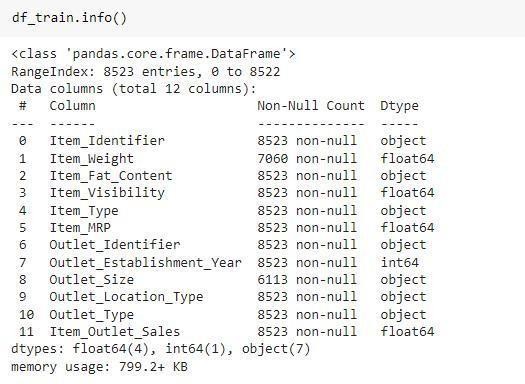


Fig 3.11 Datatype of various features of dataset

Since item\_weight is a numerical feature, filling its missing value using the average imputation method.

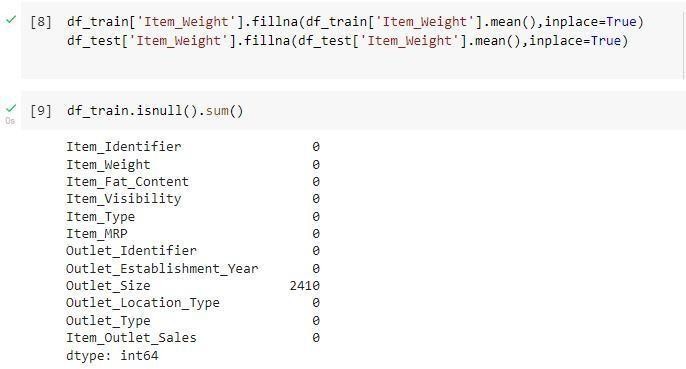


Fig 3.12 Missing value in outlet\_size column = 2410



Fig 3.13: Filling Values in Outlet\_Size.

Outlet size is a categorical feature so filling the value using the mode imputation method

So finally -

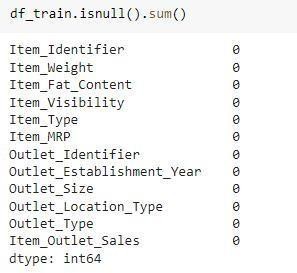


Fig 3.14 : Now there are no missing values in the item\_weight and Outer\_size columns.

### EDA

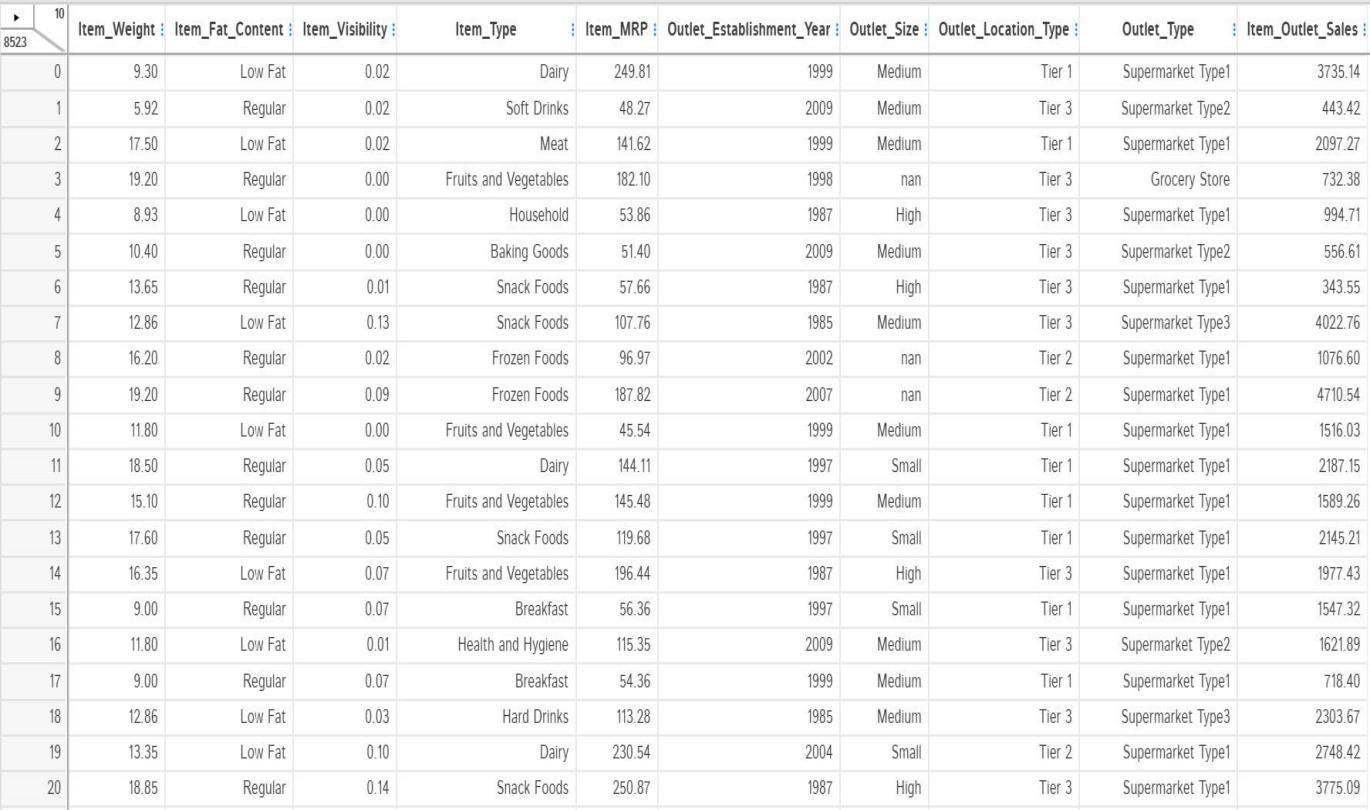
* + - 1. **EDA WITH DTALE LIBRARY**

D-Tale is a Flask and React-based powerful tool which is used to analyze and visualize pandas' data structure seamlessly.

D-Tale also supports objects like Data Frame, Series, etc.



Fig 3.15: Represents how to import dtale library and display the table



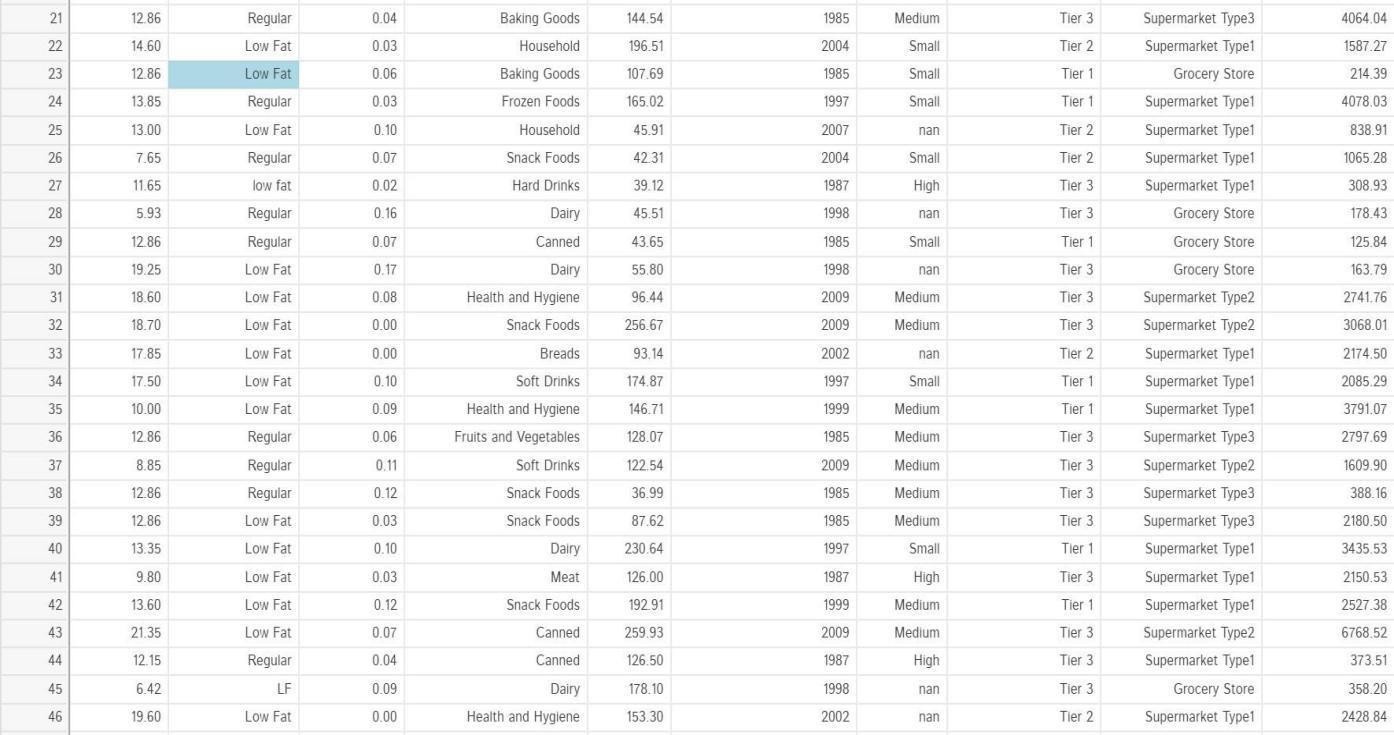


Fig 3.16: The Dtale Window

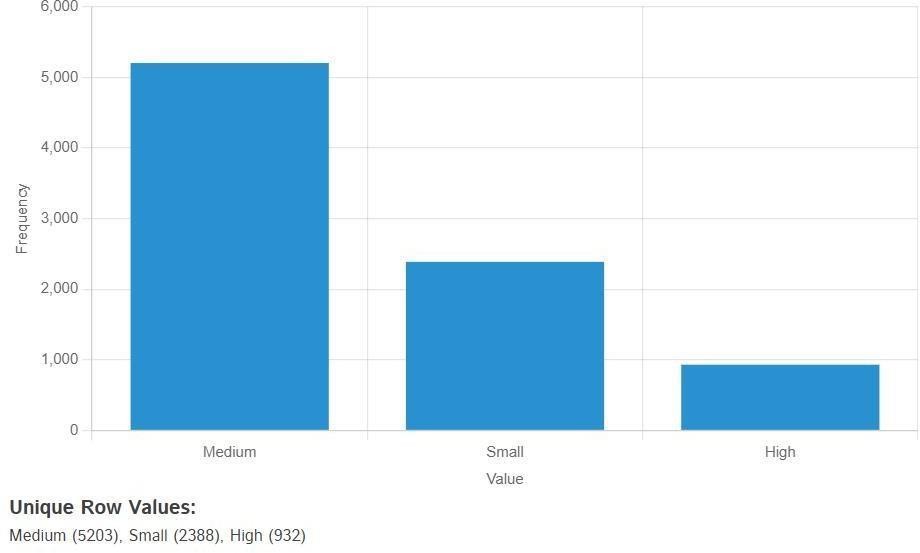


Fig 3.17: Frequency of values in the column name Outlet\_Size

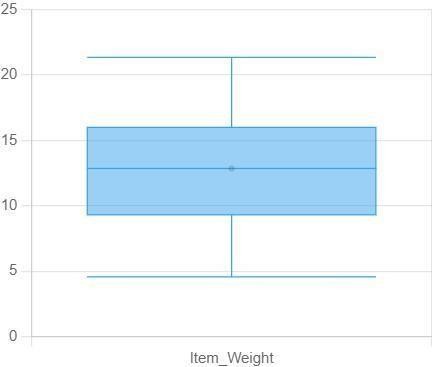
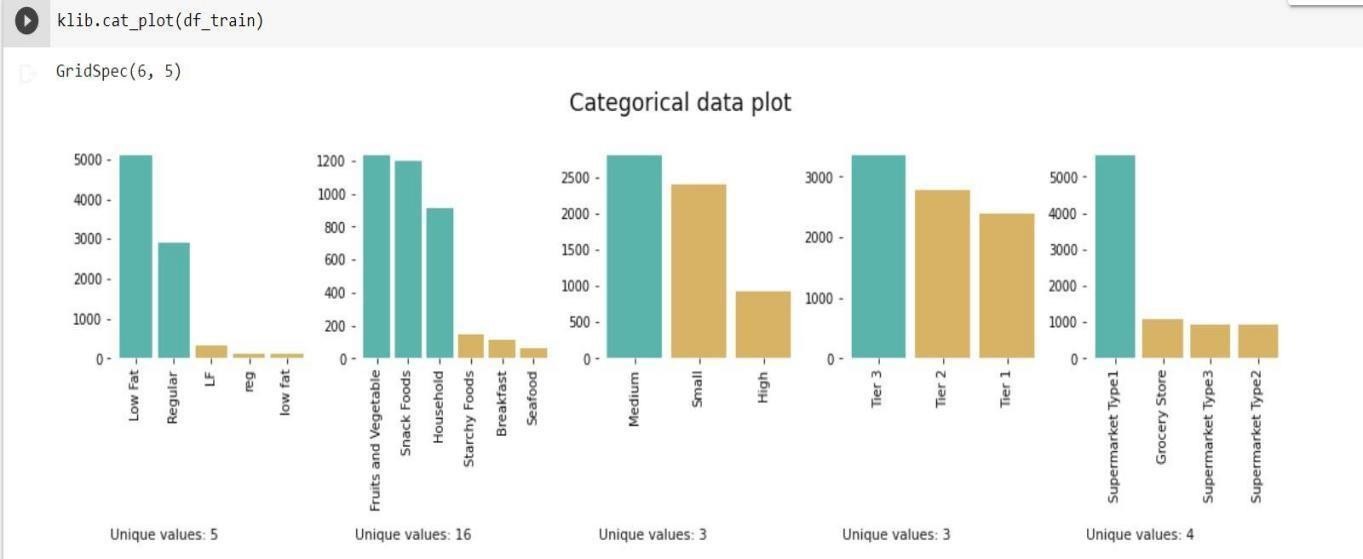


Fig 3.18 : This figure represents the Item\_Weight value range

### EDA USING KLIB LIBRARY

Klib is a python library which is used for importing, cleaning, analyzing and preprocessing the data.



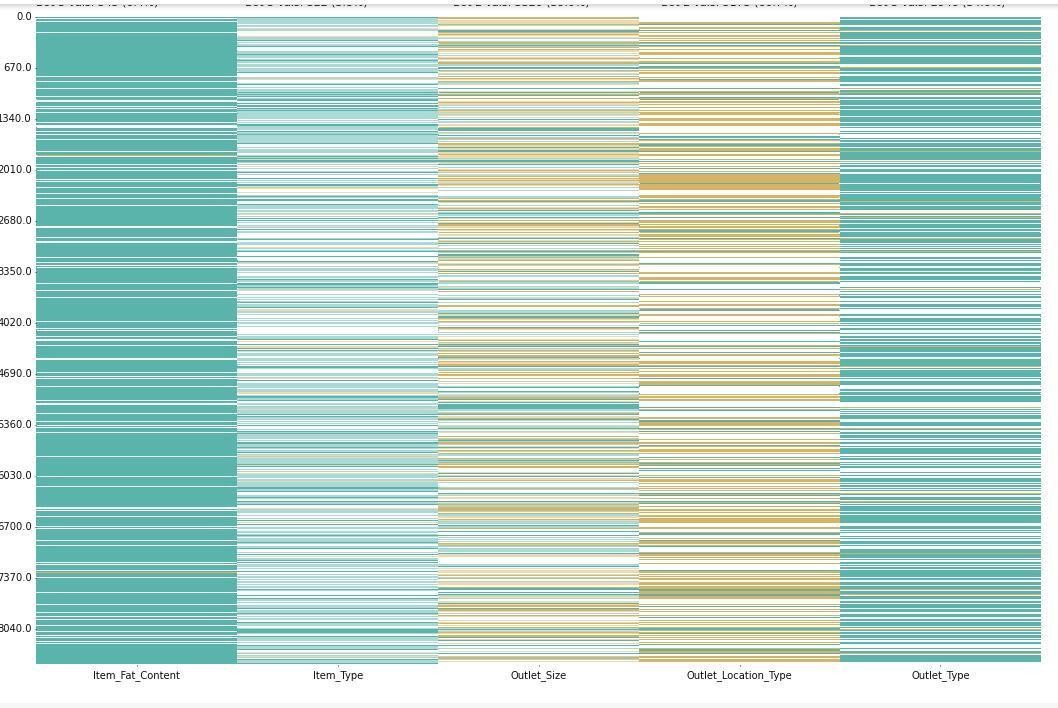


Fig 3.19 : Categorical data plot of all variables present in dataset using Klib Library

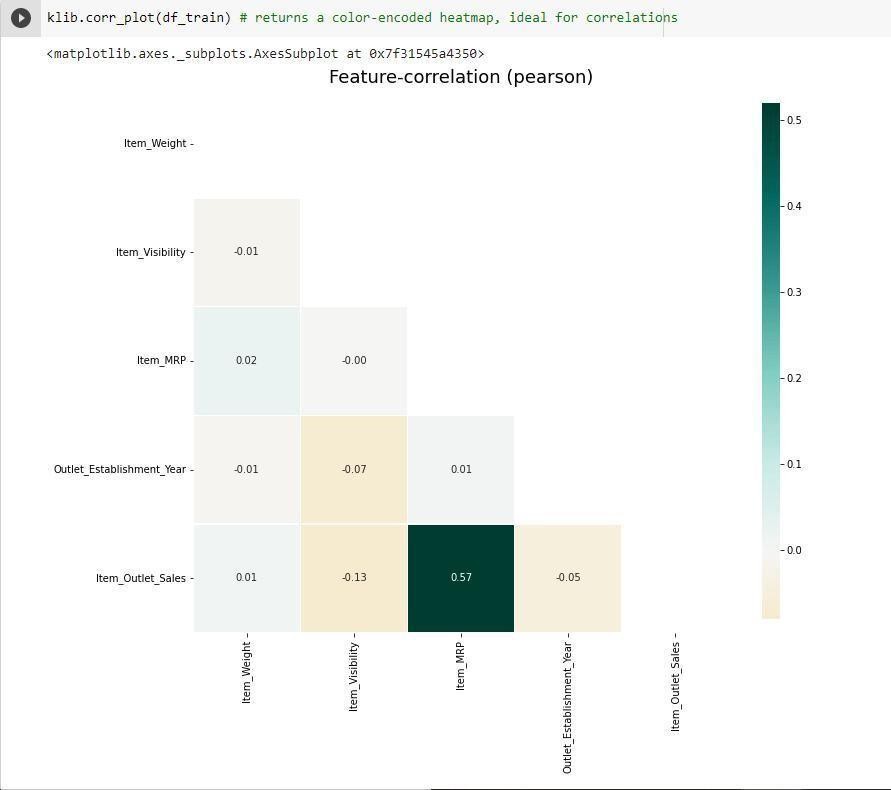


Fig 3.20 : Feature- correlation using klib Library



Fig 3.21 :Color- encoded correlation matrix.

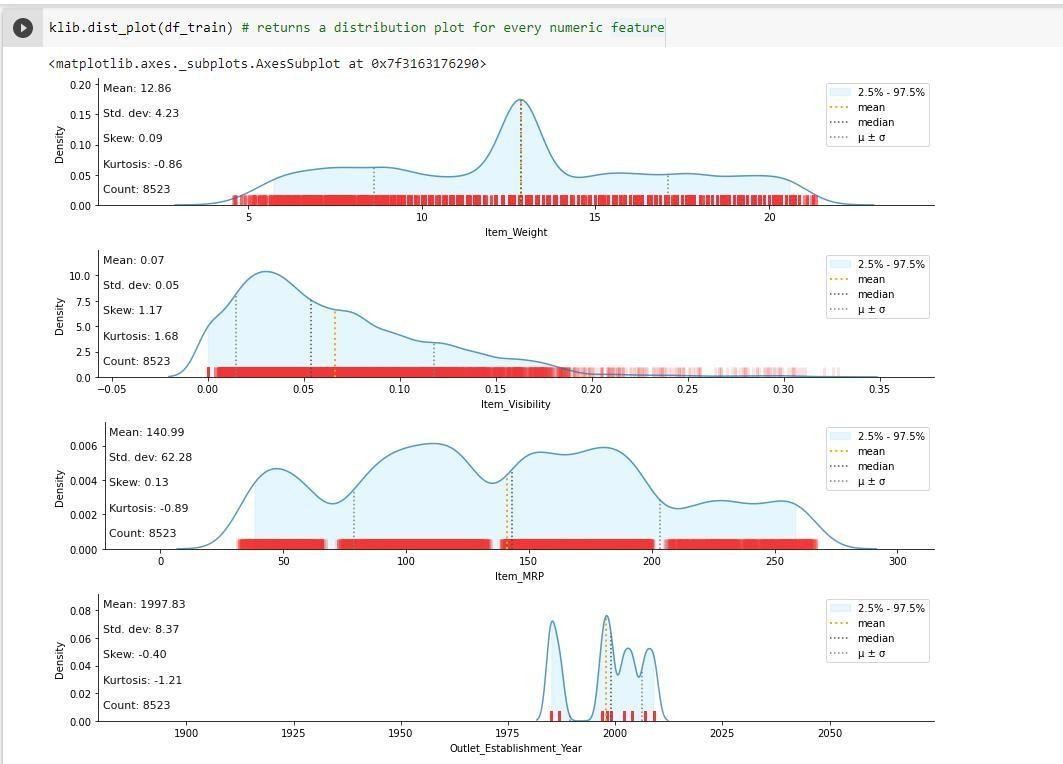
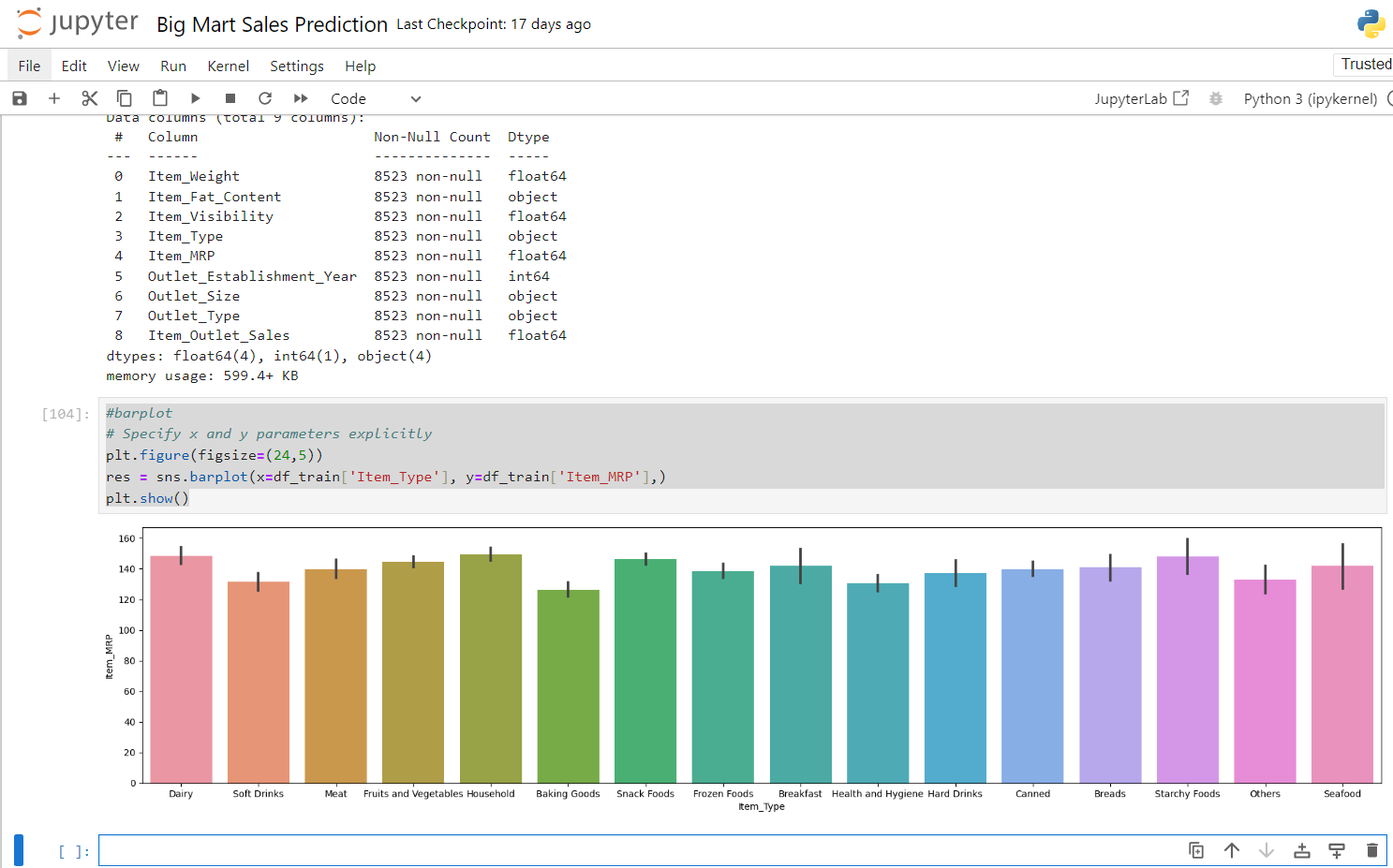


Fig 3.22: Distribution plot for every numeric feature.

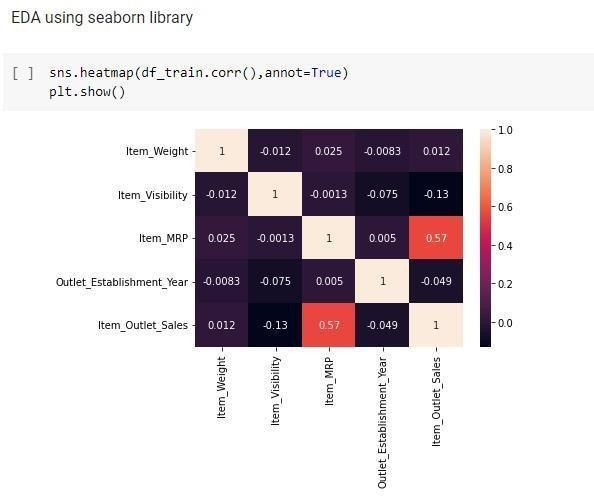
* + - 1. **EDA WITH SEABORN LIBRARY-** Seaborn is a data visualization library built on top of matplotlib

//*Majoring Item\_Type & Item\_MRP Data.*

A graph of different colored bars

Description automatically generated with medium confidence

-

  
Fig 3.23 : Correlation between different features

From the figure 3.23 we can clearly see that item\_visibility attribute has the lowest correlation with the other target variables and Item\_MRP has strong positive correlation with target variables i.e. 0.57.

* + 1. **DATA CLEANING USING KLIB LIBRARY**

Data cleaning is basically the process where the corrupt recordset, tables or databases are detected and then corrected by replacing, modifying, or deleting the dirty or coarse data.

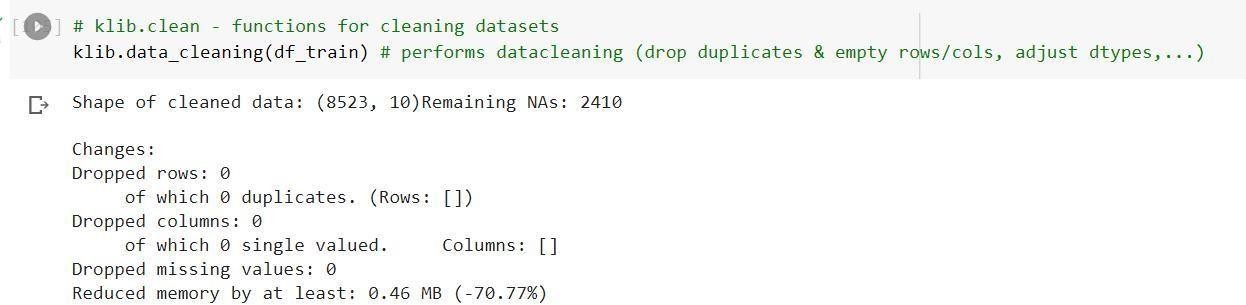
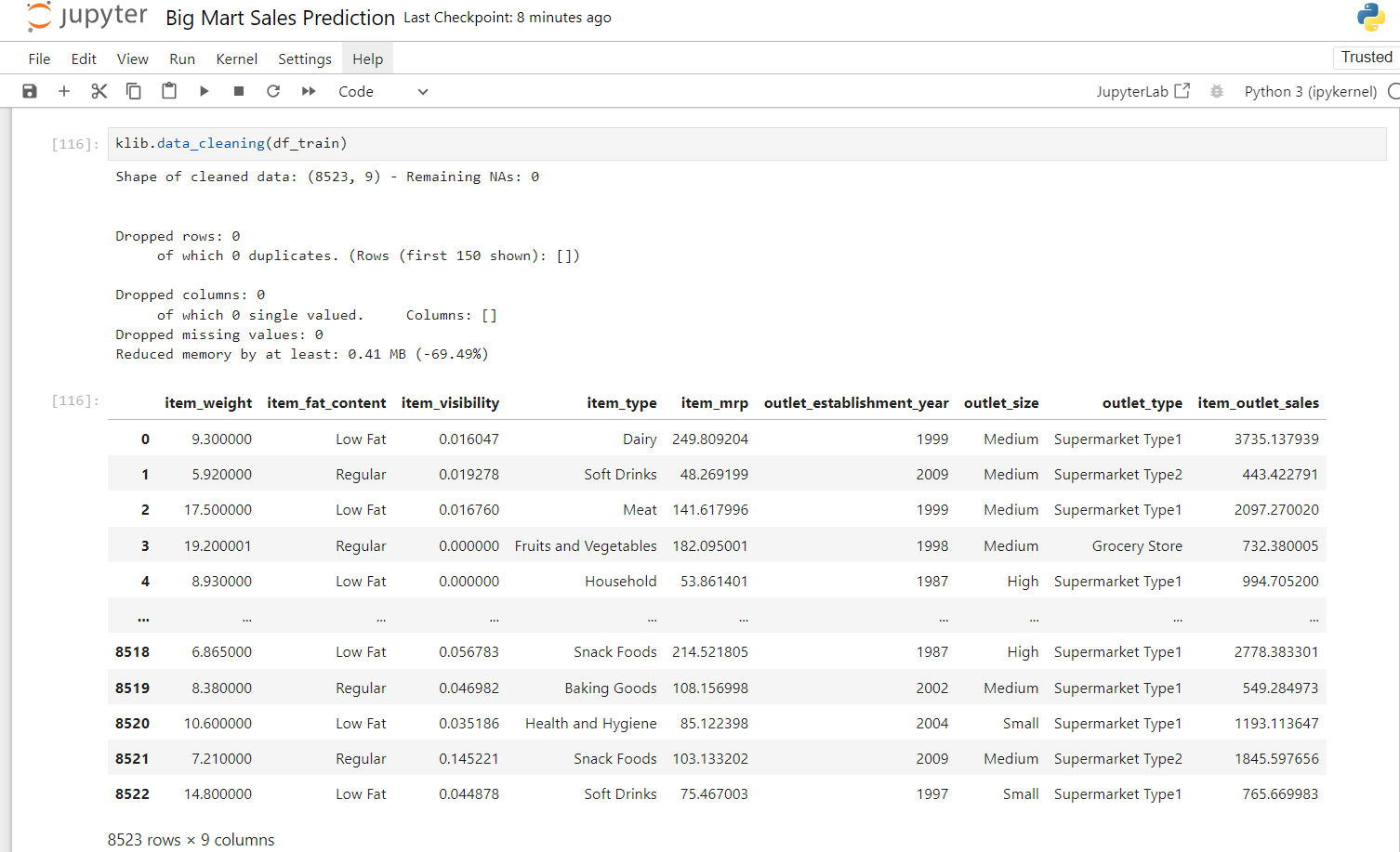


Fig 3.24 :Cleaning the data using klib library



A screenshot of a computer

Description automatically generated

Fig 3.26 : Represents the 12 features of the dataset ie numerical and categorical

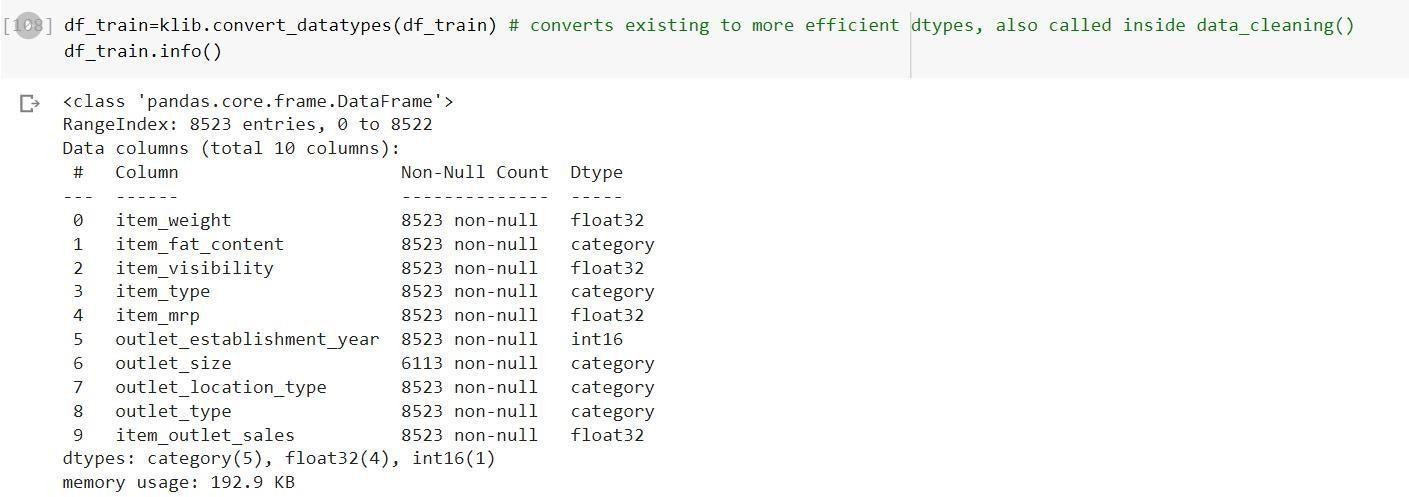
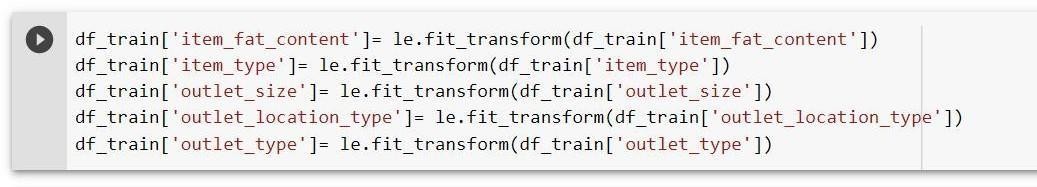


Fig 3.27 : Converting to more efficient data types using convert\_datatypes function

* + 1. **FEATURE ENGINEERING**

Feature Engineering is a way of using domain data to understand how to build mechanical operations learning algorithms. When feature engineering is done properly, the ability to predict ML algorithms are developed by creating useful raw data features that simplify the ML process. Feature engineering including correction of incorrect values. In the device database, object visibility has a small value of 0 which is unacceptable, because the object must be accessible to all, and so it is replaced by the mean of the column.



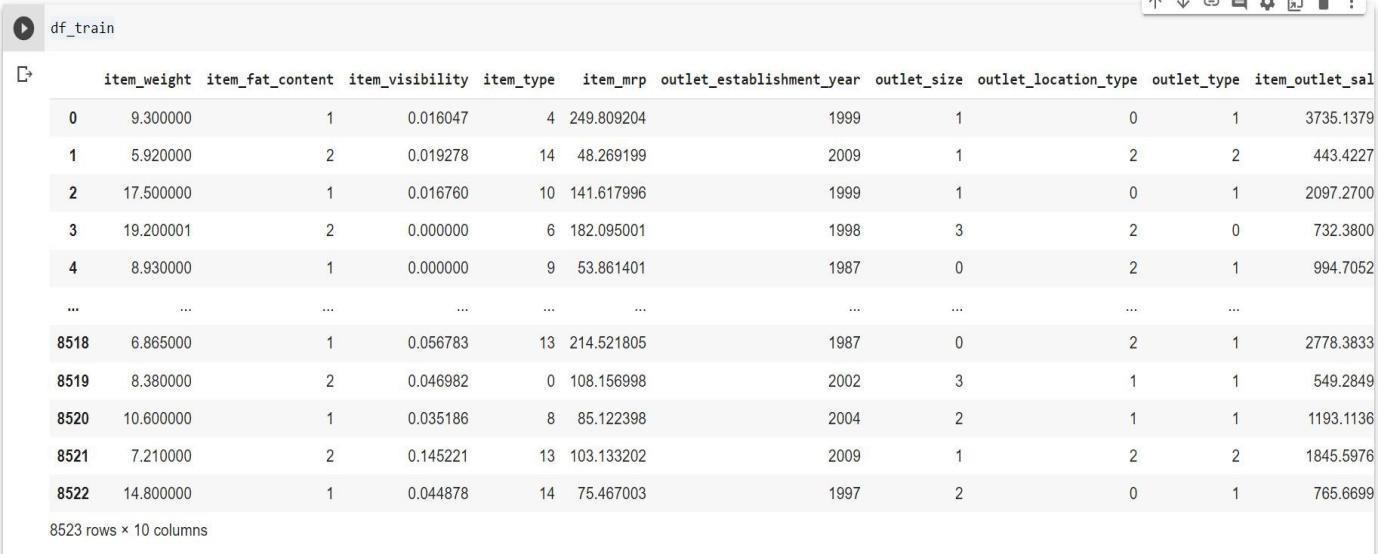


Fig 3.28 : Label Encoding Code



Fig 3.29 : Splitting of data into train and test data set.

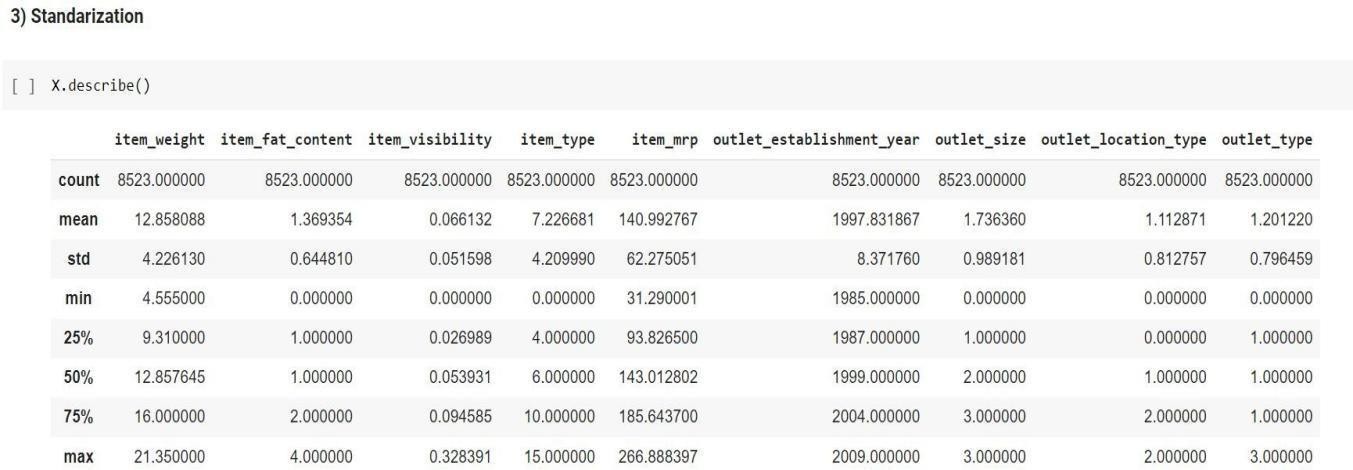


Fig 3.30 : Standardization of dataset

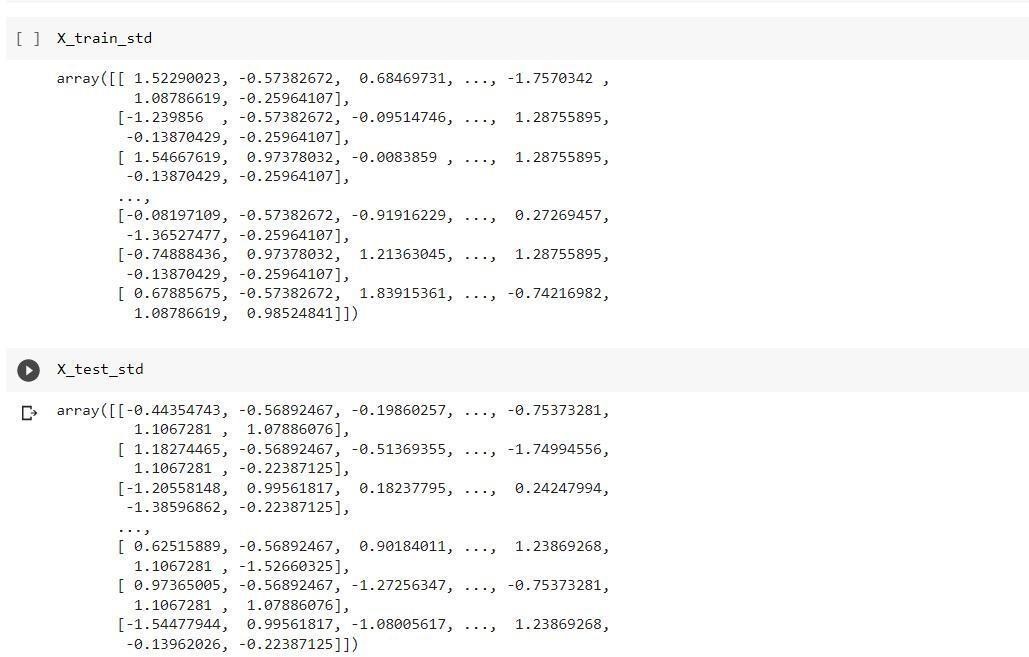


Fig 3.31 X\_train\_std array and X\_test\_std array

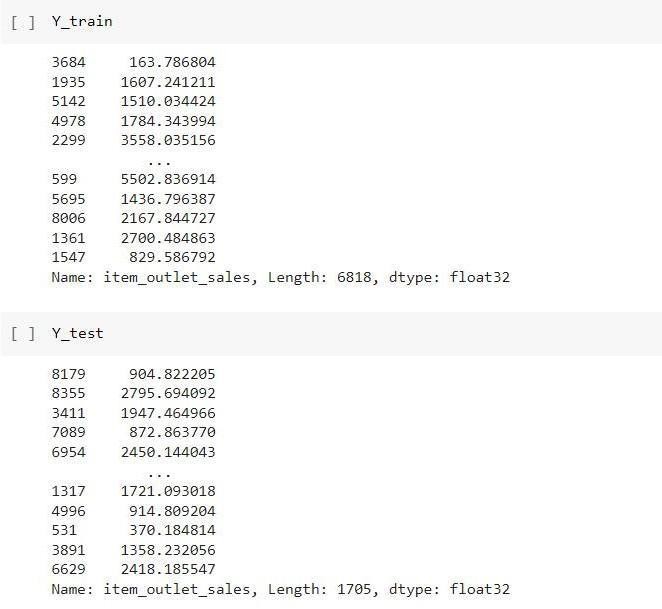


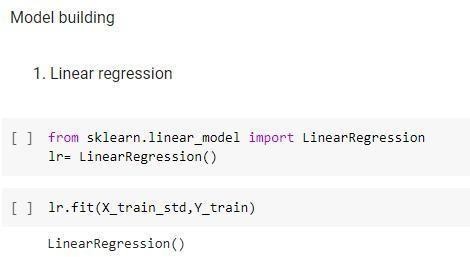
Fig 3.32 Y\_train array and Y\_test array

In figures 3.33 and 3.34 we just split the train and test data into X\_train\_std , Y\_train, X\_test\_std and Y\_test.

* + 1. **MODEL BUILDING**

Now the dataset is ready to fit a model after performing Data Preprocessing and Feature Transformation. The training set is fed into the algorithm in order to learn how to predict values. Testing data is given as input after Model Building a target variable to predict. The models are built using:

* + - 1. LR
      2. RF Regression
      3. Hyper Parameter Tuning
      4. XGBoost Regression
      5. Decision Tree
      6. Ridge Regression



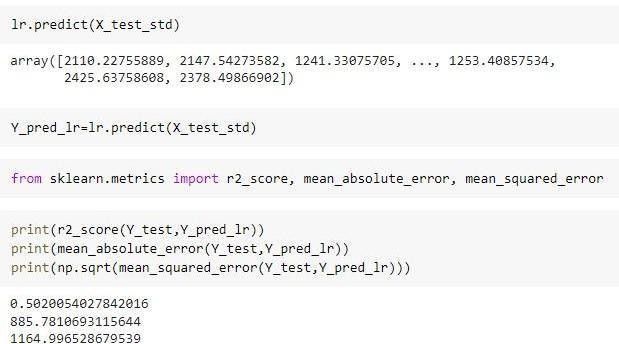


Fig 3.35: Value of R2 in Linear Regression = 0.50

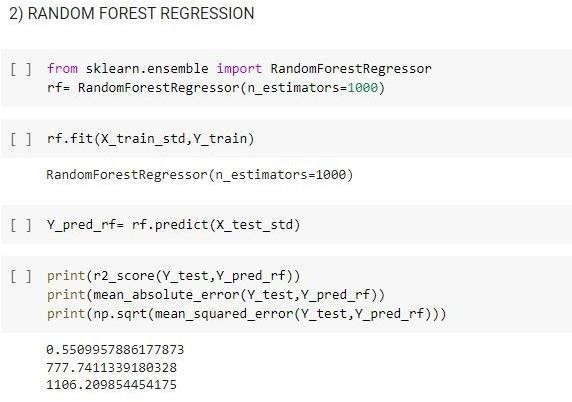
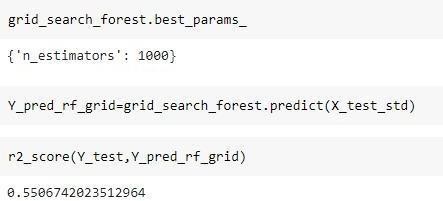


Fig 3.36: Value of R2 in Random Forest Regression = 0.55





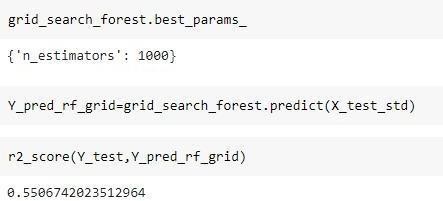


Fig 3.37: Value of R2 = 0.55



Fig 3.38: Value of R2 in XGBoost Regression = 0.63

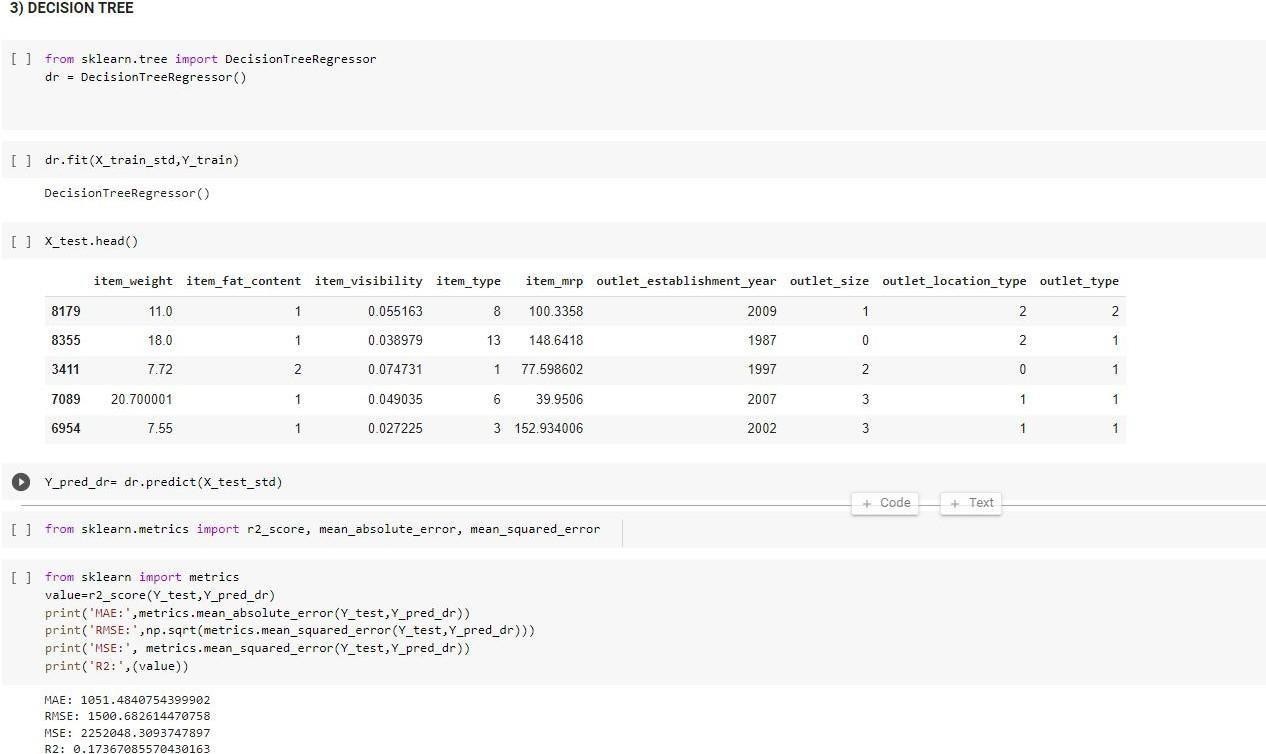


Fig 3.39: Value of R2 in Decision Tree = 0.17

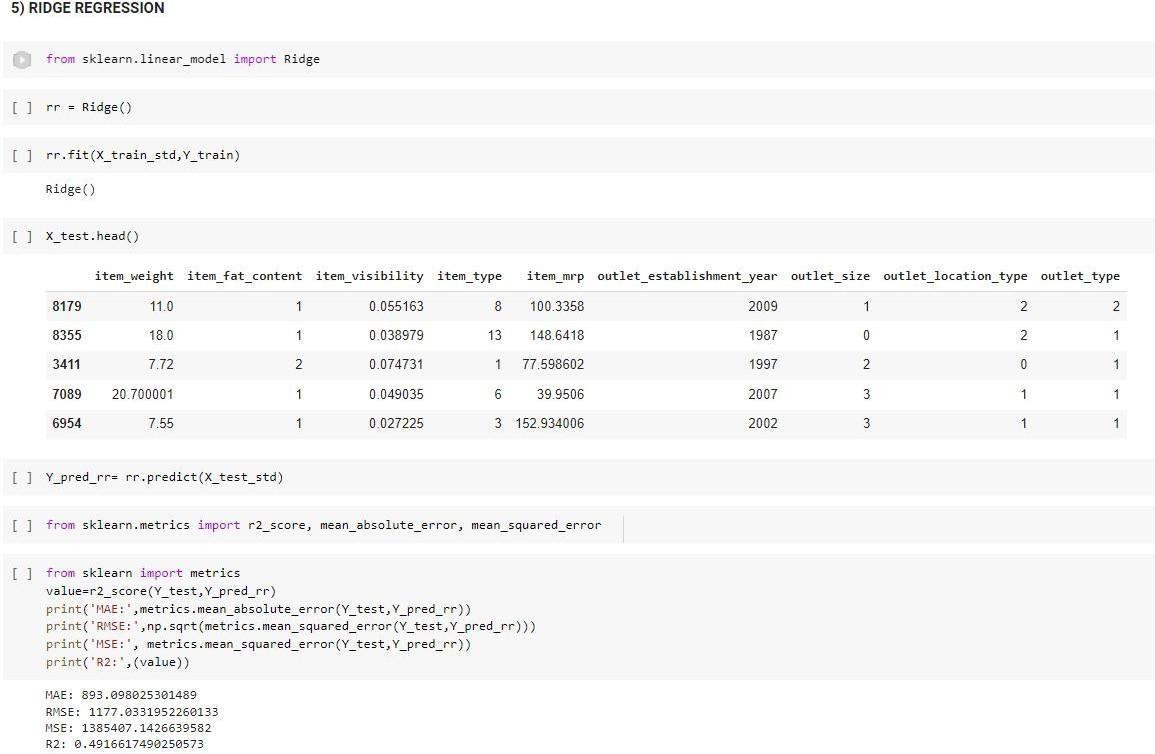


Fig 3.40: Value of R2 in Ridge Regression = 0.4916

**CHAPTER 4**

## . PERFORMANCE ANALYSIS

For the purpose of performance analysis we can go and look for the R2 value of the different algorithm performed and check for which algorithm gives us the best performance

LR

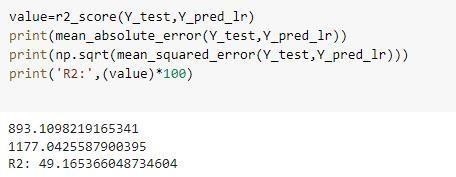


Fig 4.1 Performance of Linear Regression

RF regression

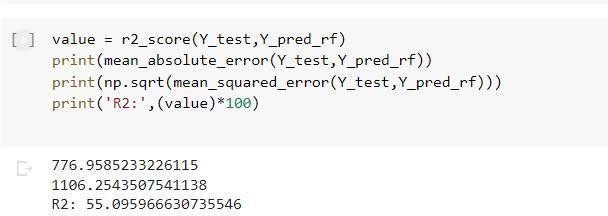


Fig 4.2 :Performance of Random Forest Regression

Hyper parameter tuning

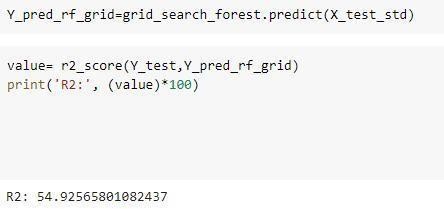


Fig 4.3: Performance of Hyper Tuning Parameter

Decision Tree

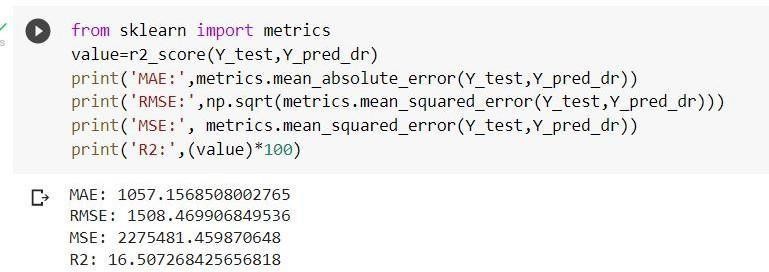


Fig 4.4: Performance of Decision Tree

XGBoost Regression

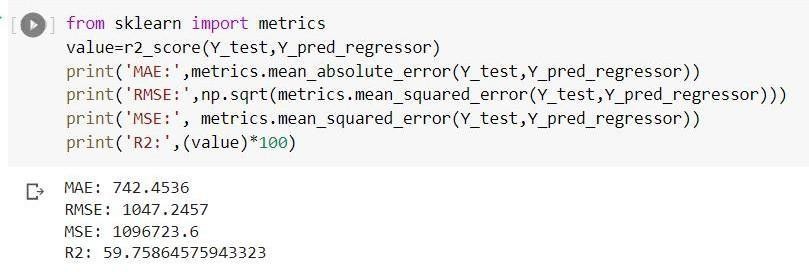


Fig 4.5: Performance of XgBoost Regression

Ridge Regression

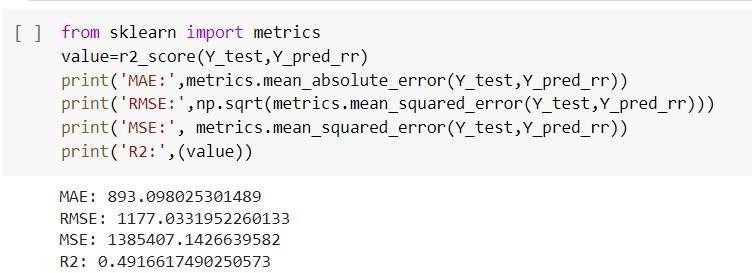


Fig 4.6: Performance of Ridge Regression

TABLE 4.1 : Algorithms Performance

|  |  |  |  |
| --- | --- | --- | --- |
| ALGORITHM | R2 | RMSE | MSE |
| Linear Regression | 49.165 | 1177.04 | 1385429.18 |
| Random Forest Regression | 55.09 | 1105 | 12222736.57 |
| Decision Tree | 16.50 | 1508.46 | 2275481.45 |
| XGBoost Regression | 59.75 | 1047 | 1096723.67 |
| Ridge Regression | 49.166 | 117.03 | 1385407.14 |

To forecast BigMart’s revenue, simple to advanced ML algorithms have been implemented, such as LR, Decision Tree, RF regression and XGBoost.

From the above table, we conclude that the XGBoost algorithm is more efficient and gives accurate and fast results.

### PERFORMANCE ANALYSIS USING GRAPHS RMSE AND MSE VALUES

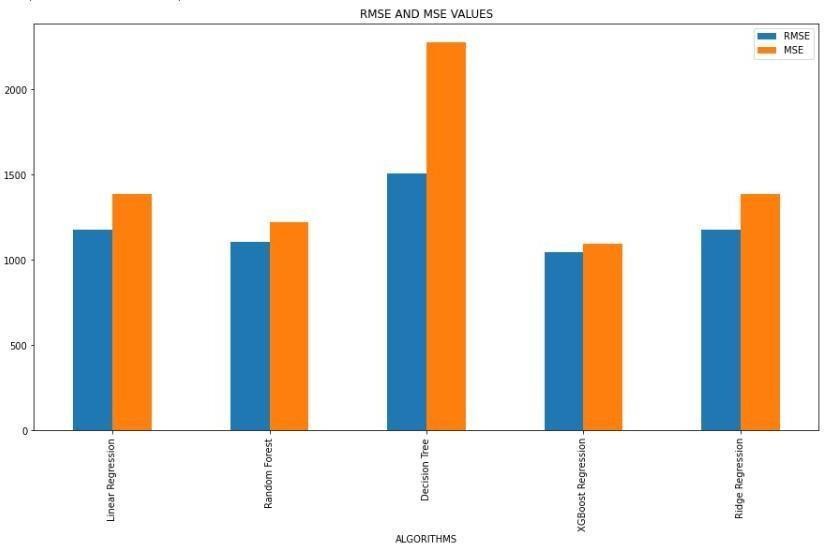


Fig 4.7:Comparison of RMSE and MSE values for ML Algorithms used

Figure 4.7 shows the comparative analysis of RMSE and MSE values. RMSE is the squared root of MSE and MSE is calculated by the squared difference between the original and predicted values in the data set. In this experiment Decision tree has the highest RMSE and MSE value and XgBoost Regression has the lowest RMSE and MSE value.

### R2 AND MAE VALUES

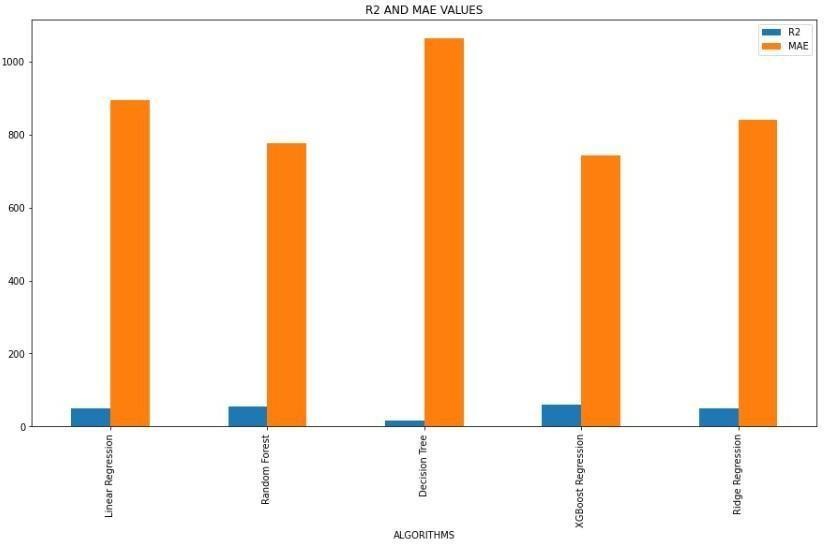


Fig 4.8:Comparison of R2 and MAE values for ML Algorithms used

Figure 4.8 shows the comparative analysis of R2 and MAE values. MAE is calculated by the average of the absolute difference between the actual and predicted values in the dataset and R2 is calculated by the sum of the residuals squared, and the total sum of squares is the sum of all the data's deviations from the mean. In this experiment Decision tree has the highest MAE value whereas XgBoost has the lowest and in case of R2 XgBoost has the highest value whereas Decision tree has the lowest value.

It has been observed that increased efficiency is observed with XGBoost algorithms with lower RMSE, MSE and MAE rating and higher R2 rating

## CHAPTER 5

1. **. CONCLUSIONS**
   1. CONCLUSION

So from this project we conclude that a smart sales forecasting program is required to manage vast volumes of knowledge for business organizations.

The Algorithms which are presented in this report , LR, RF regression, Decision tree and XGBoost regression provide an effective method for data sharing as well as decision-making and also provide new approaches that are used for better identifying consumer needs and formulate marketing plans that are going to be implemented.

The outcomes of ML algorithms which are done in this project will help us to pick the foremost suitable demand prediction algorithm and with the aid of which BigMart will prepare its marketing campaigns.

* 1. FUTURE SCOPE

The future scope of this project is that this project can further collaborate with any other devices which are supported with an in-built intelligence by virtue of the Internet of Things (I0T) which makes it more feasible to use.

Multiple instances parameters and various factors are also make this sales prediction project more

innovative and successful.

The most important term for any prediction-based system that is accuracy, is often significantly increased

because of the increase in the number of parameters.

## 6. REFERENCES

1. Beheshti-Kashi, S., Karimi, H.R., Thoben, K.D., Lutjen, M., Teucke, M.: A survey on retail sales forecasting and prediction in fashion markets. Systems Science &Control Engineering 3(1), (2015), pp.154–161
2. Bose, I., Mahapatra, R.K.: Business data mining ML perspective. Information & management 39(3),(2001), pp. 211–225
3. Mitchell, T. M. ML and data mining. Communications of the ACM, 42(11) , (1999), pp. 30- 36.
4. Das, P., Chaudhury, S.: Prediction of retail sales of footwear using feedforward and recurrent neural networks. Neural Computing and Applications 16(4-5),(2007), pp. 491–502
5. Punam, K., Pamula, R., Jain, P.K.: A two-level statistical model for big mart sales prediction. In: 2018 International Conference on Computing, Power and Communication Technologies (GUCON), IEEE (2018). pp. 617–620.