**Intel College Excellence Program   
Project Synopsis**

**“Big Mart Sale prediction using Regression”**

|  |  |  |  |
| --- | --- | --- | --- |
| **Team member’s detail** | | | |
| **S.No.** | **Participant Name** | **Mobile No.** | **Email ID** |
| 1 | Manish Kumar | 7827043684 | ***sahmanish10987@gmail.com*** |
| 2 | Sakshi Pandey | 9821955176 | ***sakshipandey21022003@gmail.com*** |
|  |  |  |  |
| **Faculty(college) mentor detail** | | | |
| **S.No.** | **Mentor Name** | **Mobile No.** | **Email ID** |
| 1 | Anjum Mohd. Aslam | 7347058401 | ***anjum.aslam@galgotiasuniversity.edu.in*** |
| **College/University Name** | | | |
| ***Galgotias University*** | | | |

School of Computing Science and Engineering,

Galgotias University

**BACKGROUND**

Nowadays, shopping malls and Big Marts organizations are expanding their businesses globally, so Sales Prediction is a big matter these days for predicting future customer demand. Sales forecasting can assist a company in working and growing in the appropriate path. We propose a regression-based predictive model for Big Mart sales analysis. The sales volume of Big Mart is forecasted by analyzing the obtained data set using the Regression model.

**PROBLEM IDENTIFICATION**

*As business of big mart, shopping malls are increasing day by day so demand and need of predicting the customer demand and future sale is also increasing. We study the dataset of items sells in supermarket, grocery store, items fat content, type, item size, item visibility, mrp etc and tries to find which factor effect the sale of items.*

**PROPOSED SOLUTION**

We used some of the regression technique like linear regression, decision tree regression and random forest to predict the result with higher accuracy. To perform regression, firstly it is important to clean the data set. Then apply algorithms on it. Linear regression model to predict the value of a dependent variable (y) based on the value of an independent variable (x). As a result of this regression technique, a linear relationship between x i.e., input and y i.e., output is found.

y=a+b\*x

where a = intercept

b = slope of line

In Decision tree is used for regression problems where you are trying to predict something with infinite possible answers such as sale of big mart. Decision trees can be used for either classification or regression problems and are useful for complex datasets. Random forest is a supervised machine learning algorithm that is commonly used to solve classification and regression problems. It creates decision trees from various samples, using the majority vote for classification and the average for regression.

**Algorithm**

Step 1: Import the dataset

Step 2: Read the dataset

Step 3: Calculate the total missing in each column of dataset

Step 4: Perform Data cleaning

Step 5: Imputing Missing Values

Step 6: Data understanding through visualization (Compare every column with sales to observe which aspect is affecting sale of item)

Step 7: Apply different regression technique and observe the result.

**APPROACH TAKEN**

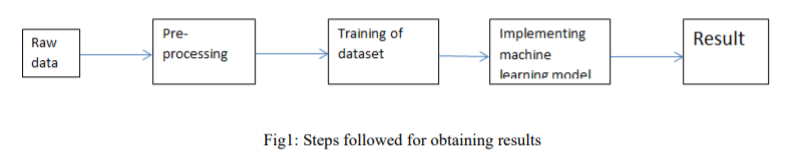
“To find out what role certain properties of an item play and how they affect their sales by

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.

**Methodology**



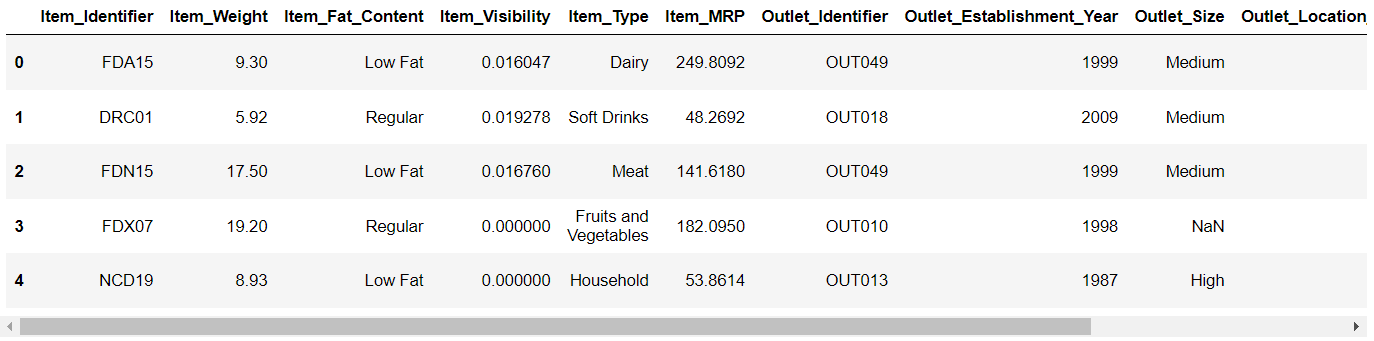
**Dataset and its processing**

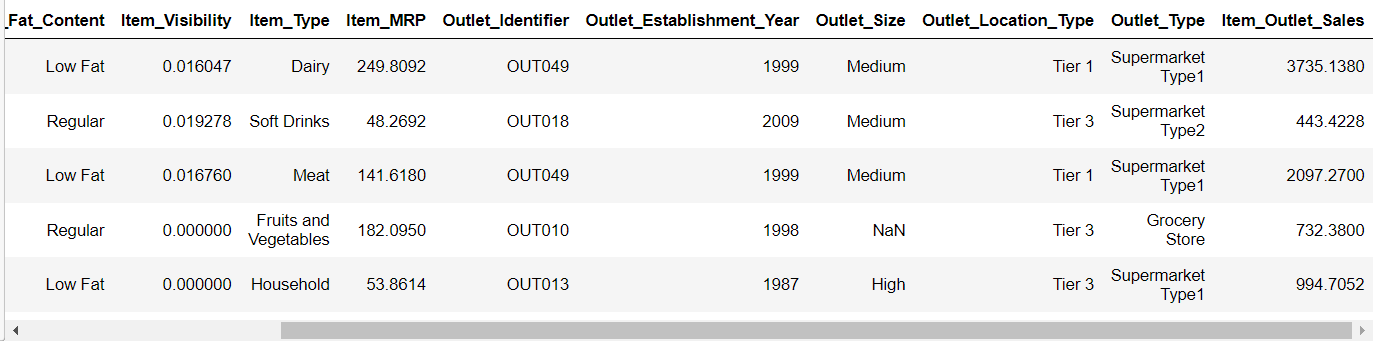
Big Mart’s data scientists collected sales data of their 10 stores situated at different locations

with each store having 1559 different products as per 2013 data collection. Using all the

observations it is inferred what role certain properties of an item play and how they affect its

sales.





**After cleaning of dataset**

Preprocessing of this dataset includes doing analysis on the independent variables like checking

for null values in each column and then replacing or filling them with supported appropriate data

types, so that analysis and model fitting is not hindered from its way to accuracy. Shown above

are some of the representations obtained by using Pandas tools which tells about variable count

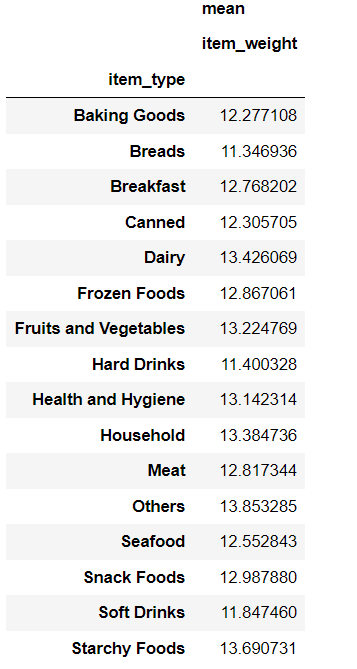
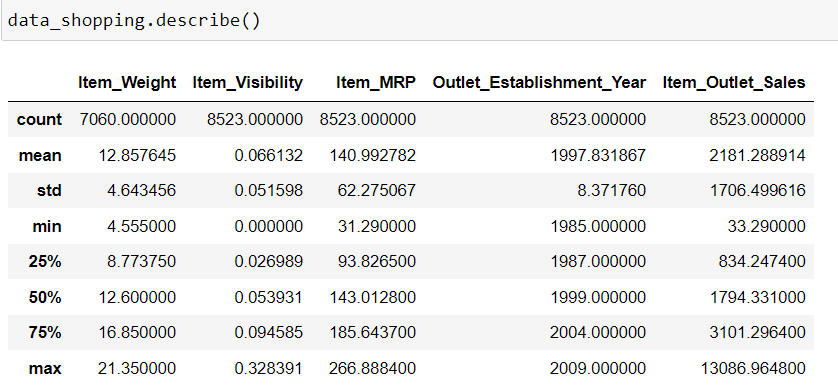
for numerical columns and modal values for categorical columns. Maximum and minimum

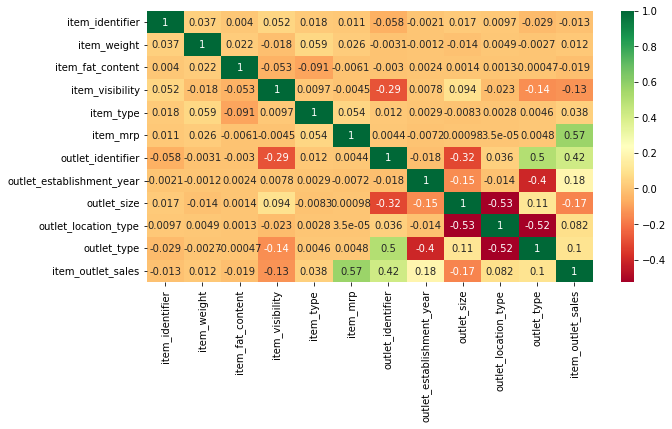
values in numerical columns, along with their percentile values for median, plays an important

factor in deciding which value to be chosen at priority for further exploration tasks and analysis.

Data types of different columns are used further in label processing and one-hot encoding

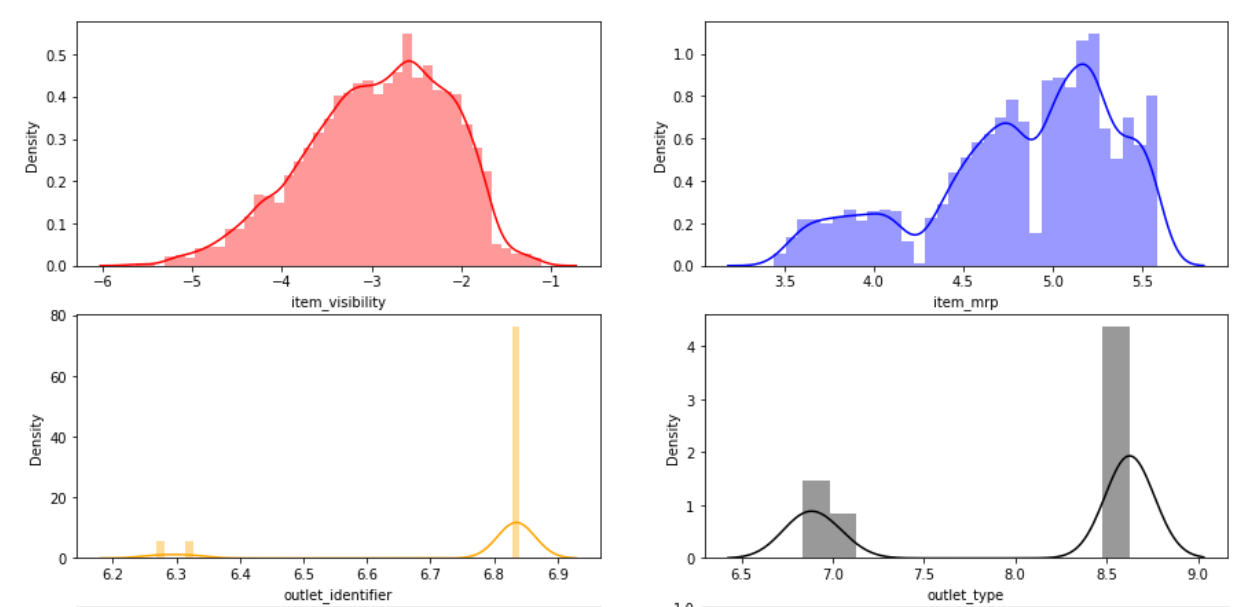
scheme during model building.

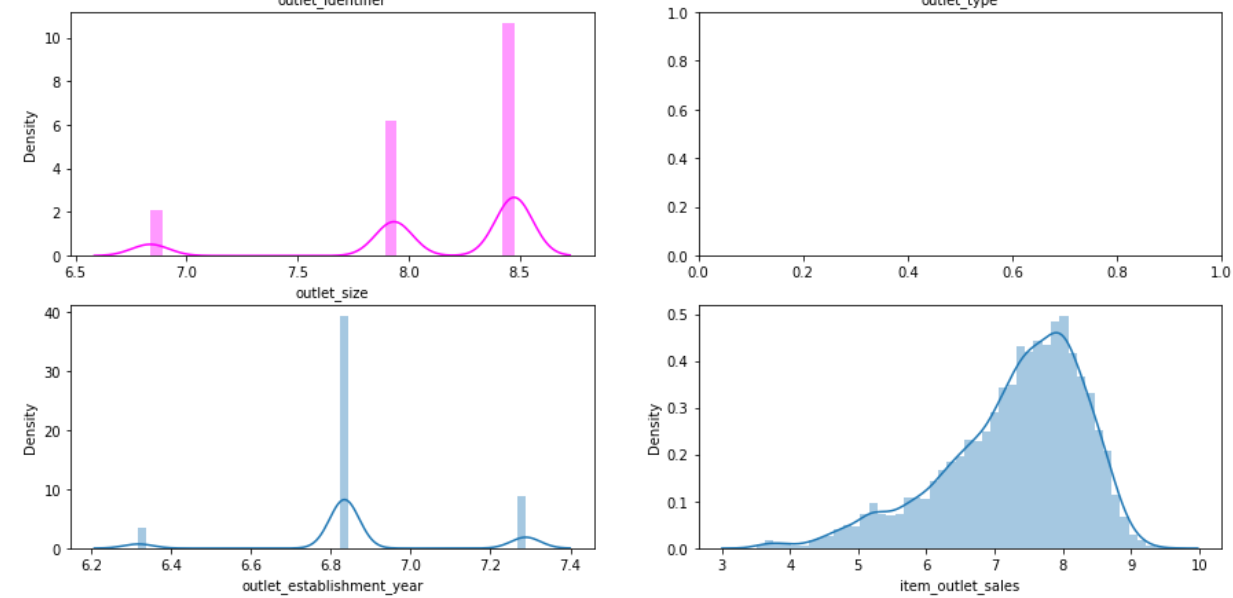
**** ****

****

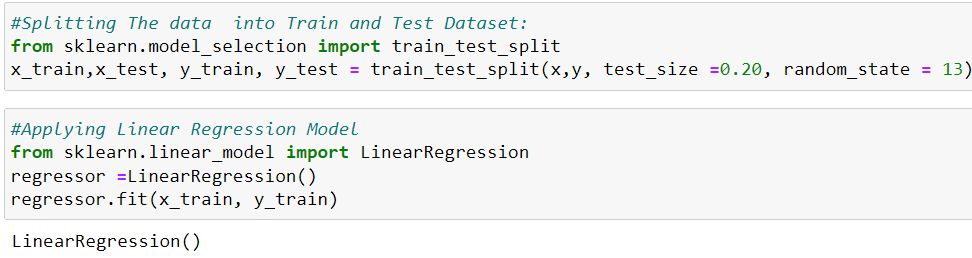
This Figure shows the correlation among each column of the dataset.

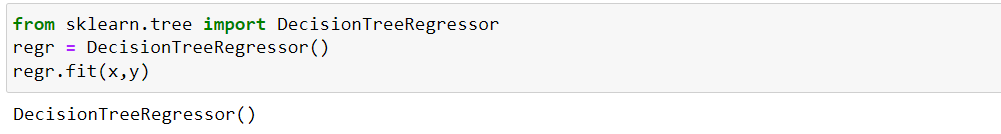
Visualizing the skewness of data

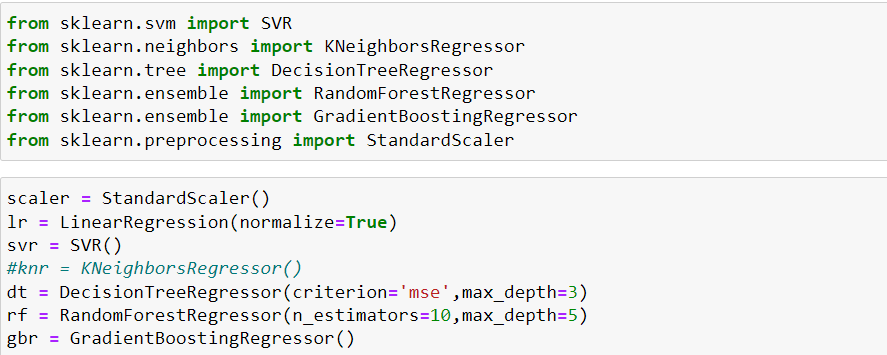




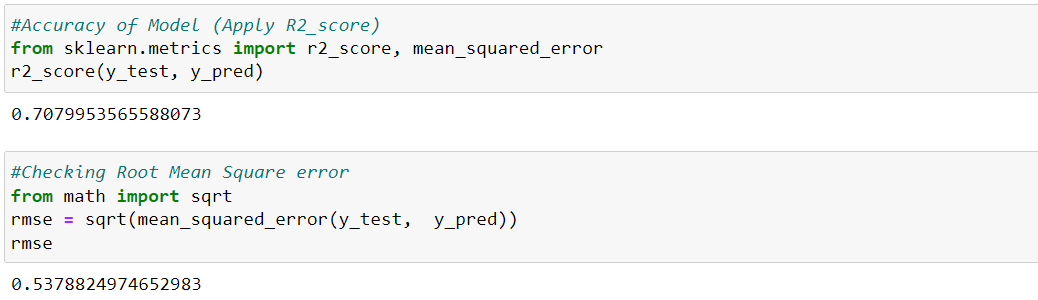
**Implementation of Machine Learning Models**

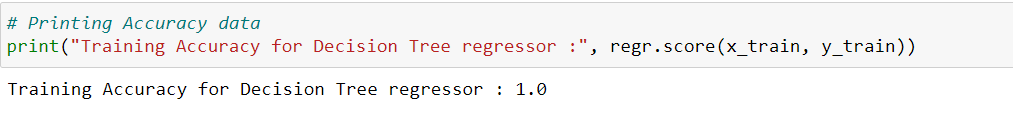
****

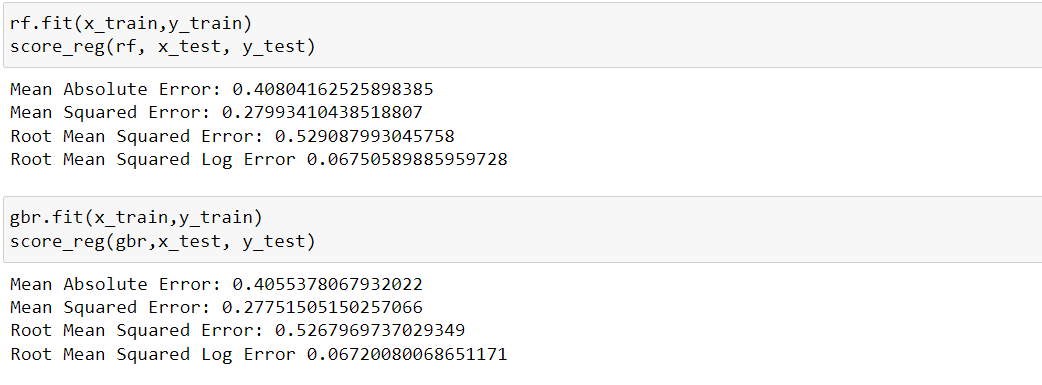
****

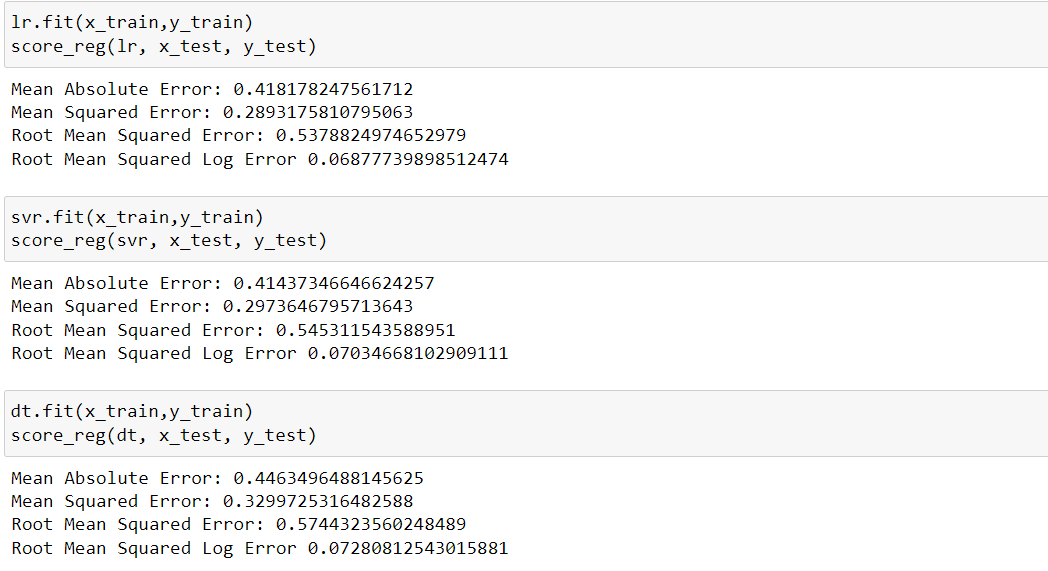
****

**Result of Machine Learning**

****

****

****

****

**HARDWARE & SOFWARE REQUIREMENTS**

*Hardware requirements:*

*1. PC/Laptop*

*Software requirements:*

*1. Anaconda*

2. Jupyter Notebook

3. Python Libraries installed

1. Scikit-Learn

2. Numpy

3. Pandas

4. Scipy

5. Matplotlib (Pyplot)

6. Seaborn

**BLOCK DIAGRAM & DESCRIPTION**

We have drawn block diagram for the project. This ER diagram is representing the process of project. Firstly, we have taken the dataset from Kaggle and processed the data. We used anaconda python platform for exploratory data analysis and visualization. Data preprocessing is mandatory for any machine learning or data mining approach, since the performance of a machine learning methodology depends on how well the dataset is prepared and structured



Import Dataset

Read Data

Calculating missing values in dataset

Data Processing

Replacing the missing values with mean value

Data Visualization (comparing each column with sales)

Applying Regression Algorithms

Random Forest

Linear Regression

Decision Tree

Performance Analysis

Result

**RESULT**

Machine Learning algorithms did a great job in the testing phase in this supervised learning environment, and the algorithms used in this will definitely perform well in real Big Mart company to predict the sales. But our best choice would be to work with decision tree regressor algorithm as it was the only one regression algorithm which was able to accomplish 100% accuracy. And the rest of the results does not exceed 70% (approx.) accuracy. So, in conclusion this project would definitely bring accurate results, if worked with decision tree regressor.

It must also be noted that all the accuracy mentioned in Approach Taken is done considering only those columns which have a influence in SALES. And those columns are 'item\_visibility', 'item\_mrp', 'outlet\_identifier', 'outlet\_establishment\_year', 'outlet\_size', and 'outlet\_type'.

**FUTURE SCOPE**

We can work on more dataset and try to apply more algorithm to increase accuracy as predicting the sale is currently on high demand as it helps business to grow positively and its scope will increase in future as well

**CONCLUSION**

So, in this project we observe that item fat contains, item price and selling sites is influencing the sale of item. By using algorithms and analyzing we tried to get more accurate result. Linear regression gives accuracy of 78% and decision tree algorithm is giving accuracy of 100%. This model can help to predict the future sale of Big Mart.

**REFERENCES**

https://github.com/Sah-Manish/Intel-COE-ML-Project.git