Technocolabs
BUILD WITH AI

If we cannot do great things, we can do small things in a great way.

Delivering smart solutions to smart people | follow us @technocollabs | mail us technocollabs@gmail.com

Final Project Report

Big Mart Sales Prediction using Python

Submitted by:

Dev Kumar Kashyap

Data Science Intern | Technocolabs

LinkedIn: DevKashyap

GitHub: devkashyap10

E-mail: dev10kashyap@gmail.com

Nowadays, shopping malls and big retail stores keep track of their sales of items or products for predicting future demand of the customers and update their stock records and inventory management system. Each organization is trying to attract more customers using personalized and short-time offers which makes the prediction of future volume of sales of every item an important asset in the planning and inventory management of every organization.

These data stores contain a large number of customer data and individual item attributes in a data warehouse. Due to the cheap availability of computing and storage, it has become possible to use sophisticated machine learning algorithms for predicting future sales volume for the retailers. This data can then further be used for forecasting future sales.

Project Description:

The data scientists at Big Mart have collected 2013 sales data for 1559 products across 10 stores in different cities. Also, certain attributes of each product and store have been defined. The aim of this data science project is to build a predictive model and find out the sales of each product at a particular store.

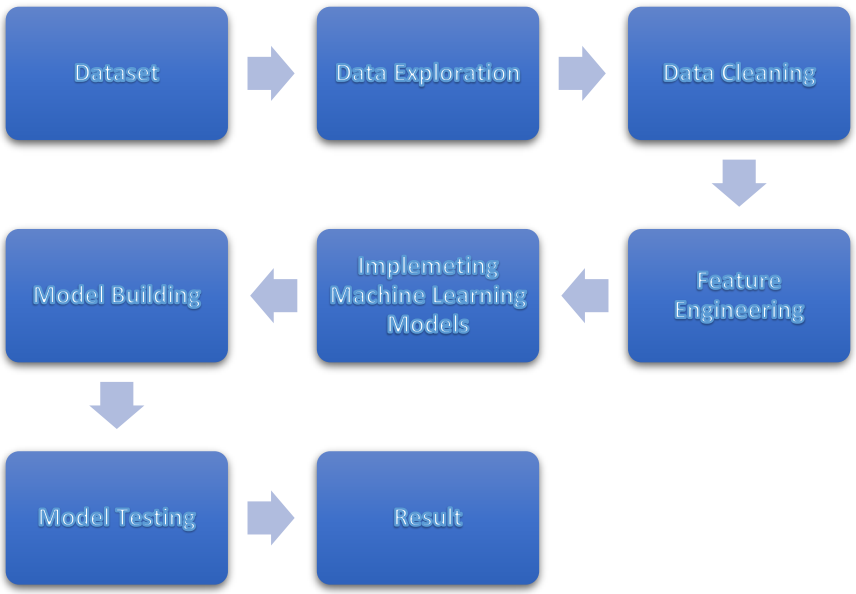
Using this model, Big Mart will try to understand the properties of products and stores which play a key role in increasing sales.

About Big Mart:

Big Mart is an International Retail Corporation. Big Mart is also One Stop Shopping center and Free Marketplace.

Methodology:

The sequence of steps that the dataset of Big Mart sales goes through to build up the proposed model to produce accurate results.



Dataset:

The dataset is taken from analyticsvidhya.com. There are two different .csv files namely, "*Train.csv*" having 8523 records and "*Test.csv*" having 5681 rows. Train data set has both input and output variable(s) We need to predict the sales for test data set.

Variable	Description
Item_Identifier	Unique product ID.
Item_Weight	Weight of product.
Item_Fat_Content	Whether the product is low fat or not.
Item_Visibility	The % of total display area of all products in a store allocated to the particular product.
Item_Type	The category to which the product belongs.
Item_MRP	Maximum Retail Price (list price) of the product.
Outlet_Identifier	Unique store ID.
Outlet_Establishment_Year	The year in which store was established.
Outlet_Size	The size of the store in terms of ground area covered.
Outlet_Location_Type	The type of city in which the store is located.
Outlet_Type	Whether the outlet is just a grocery store or some sort of supermarket.
Item_Outlet_Sales	Sales of the product in the particular store. This is the outcome variable to be predicted.

Hypotheses:

1. Store Level Hypotheses :-

a) City type	Stores located in urban or Tier 1 cities should have higher sales because of the higher income levels of people there.
b) Population Density	Stores located in densely populated areas should have higher sales because of more demand.
c) Store Capacity	Stores which are very big in size should have higher sales as they act like one-stop-shops and people would prefer getting everything from one place.
d) Competitors	Stores having similar establishments nearby should have less sales because of more competition.
e) Marketing	Stores which have a good marketing division should have higher sales as it will be able to attract customers through the right offers and advertising.
f) Location	Stores located within popular marketplaces should have higher sales because of better access to customers.
g) Customer Behavior	Stores keeping the right set of products to meet the local needs of customers will have higher sales.
h) Ambiance	Stores which are well-maintained and managed by polite and humble people are expected to have higher footfall and thus higher sales.

2. Product Level Hypotheses :-

a) Brand	Branded products should have higher sales because of higher trust in the customer.
b) Packaging	Products with good packaging can attract customers and sell more.
c) Utility	Daily use products should have a higher tendency to sell as compared to the specific use products.
d) Display Area	Products which are given bigger shelves in the store are likely to catch attention first and sell more.
e) Visibility in Store	The location of product in a store will impact sales. Ones which are right at entrance will catch the eye of customer first rather than the ones in back.
f) Advertising	Better advertising of products in the store will should higher sales in most cases.
g) Promotional Offers	Products accompanied with attractive offers and discounts will sell more.

Data Structure and Content:

Train dataset has 8523 rows and 12 features whereas, test dataset has 5681 rows and 11 features. Train has 1 extra column which is the target variable.

It is generally a good idea to combine both train and test data sets into one to perform feature engineering and then divide them later again. This saves the trouble of performing the same steps twice on test and train.

```
Data columns (total 13 columns):
 #   Column                     Non-Null Count  Dtype
---  ------                     --------------  -----
 0   Item_Identifier            14204 non-null  object
 1   Item_Weight                11765 non-null  float64
 2   Item_Fat_Content           14204 non-null  object
 3   Item_Visibility            14204 non-null  float64
 4   Item_Type                  14204 non-null  object
 5   Item_MRP                   14204 non-null  float64
 6   Outlet_Identifier          14204 non-null  object
 7   Outlet_Establishment_Year  14204 non-null  int64
 8   Outlet_Size                10188 non-null  object
 9   Outlet_Location_Type       14204 non-null  object
 10  Outlet_Type                14204 non-null  object
 11  Item_Outlet_Sales          8523 non-null   float64
 12  source                     14204 non-null  object
dtypes: float64(4), int64(1), object(8)
memory usage: 7.4 MB
```

- The resultant shape of concatenated data is 14204 rows and 13 columns.
- Item_Weight has 2439 (17.2%) missing values.
- Outlet_Size has 4016 (28.3%) missing values.
- The Item_Outlet_Sales is the target variable and missing values are ones in the test set. So, we need not worry about it.

```
Original Categories:
Low Fat    8485
Regular    4824
LF          522
reg         195
low fat     178
Name: Item_Fat_Content, dtype: int64

Modified Categories:
Low Fat    9185
Regular    5019
Name: Item_Fat_Content, dtype: int64
```

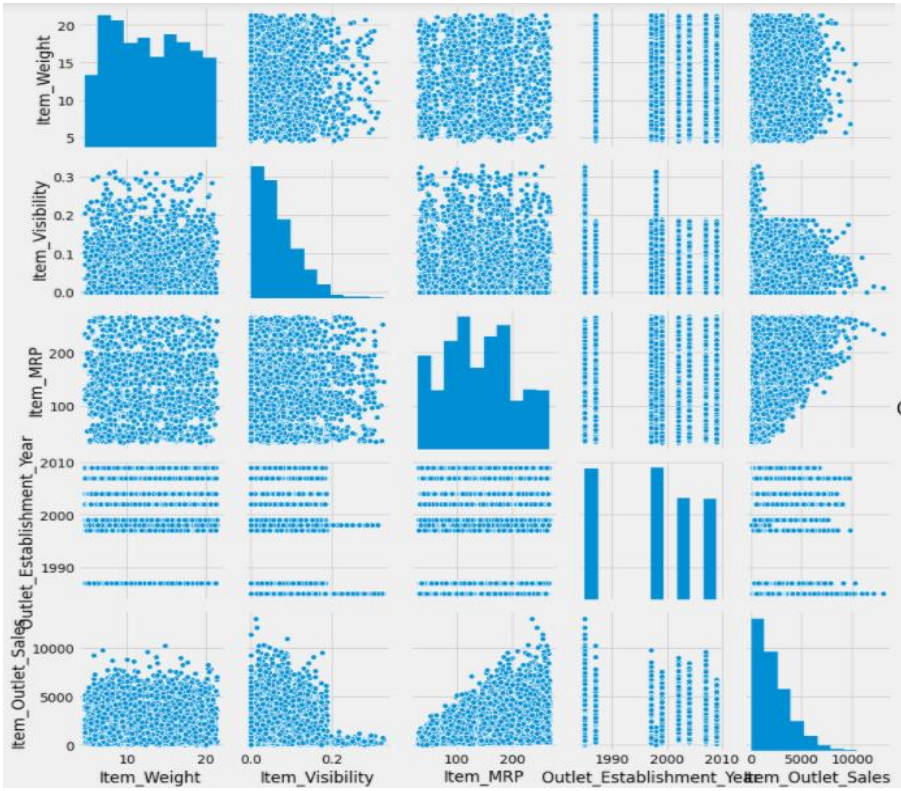
There are typos and difference in representation in categories of Item_Fat_Content variable. Some of 'Low Fat' values are mis-coded as 'low fat' and 'LF'. Also, some of 'Regular' are mentioned as 'regular'.

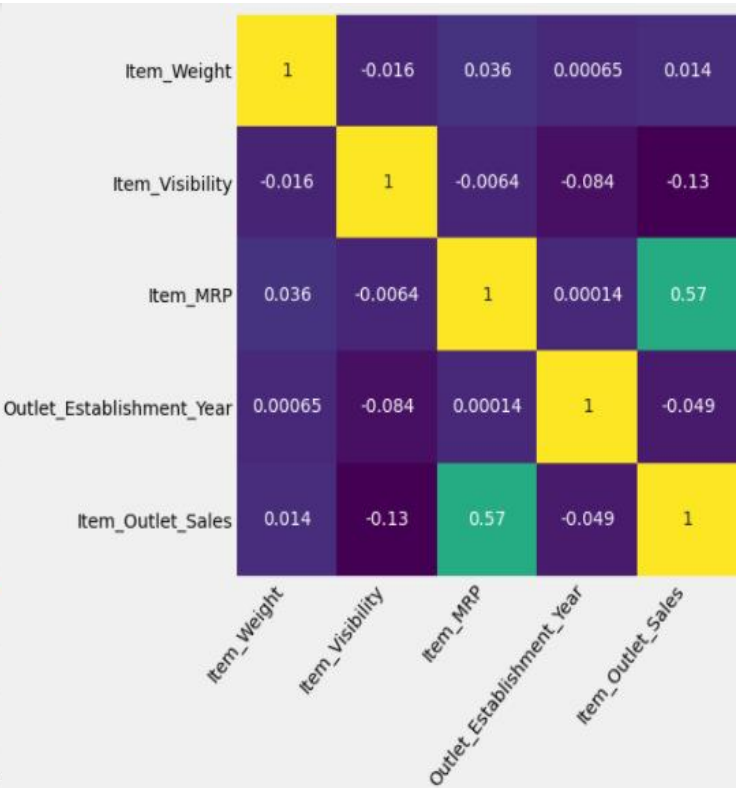
Description of The Data:

	Item_Weight	Item_Visibility	Item_MRP	Outlet_Establishment_Year	Item_Outlet_Sales
count	11765.000000	14204.000000	14204.000000	14204.000000	8523.000000
mean	12.792854	0.065953	141.004977	1997.830681	2181.288914
std	4.652502	0.051459	62.086938	8.371664	1706.499616
min	4.555000	0.000000	31.290000	1985.000000	33.290000
25%	8.710000	0.027036	94.012000	1987.000000	834.247400
50%	12.600000	0.054021	142.247000	1999.000000	1794.331000
75%	16.750000	0.094037	185.855600	2004.000000	3101.296400
max	21.350000	0.328391	266.888400	2009.000000	13086.964800

- The lower count of Item_Weight shows the presence of missing values.
- Item_Visibility has a min value of 0. It makes no practical sense as because when a product is being sold in a store, the visibility cannot be 0.
- Item_MRP has Q3 (third quartile) as 185.86 and maxiumum value as 266.89 which may indicate that there is a presence of outliers.
- Outlet_Establishment_Years vary from 1985 to 2009. If we can convert the years to how old the particular store is, it should have a better impact on sales.
- The lower 'count' of Item_Outlet_Sales shows the presence of missing values.
- There are total 1559 products but total IDs are 14204 which shows the presence of duplicate IDs. Hence, ID variable has duplicate values.
- Item_Type has 16 unique values.
- BigMart has 10 outlets/stores.
- There are 3 types of outlets/stores on the basis of area covered which is represented by Outlet_Size.
- There are 3 types of locations for an outlet/store.
- There are 4 types of outlets/stores on the basis of store capacity which is represented by Outlet_Type.
- There are 12645 duplicate IDs for 14204 total entries.

Pair plot and Correlations:



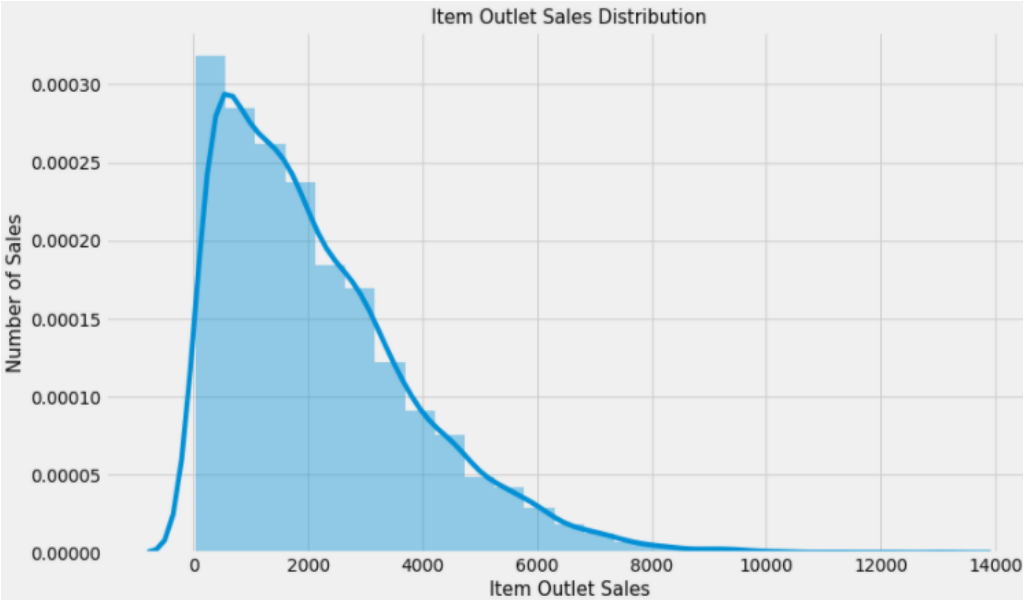


	Item_Weight	Item_Visibility	Item_MRP	Outlet_Establishment_Year	Item_Outlet_Sales
Item_Weight	1.000000	-0.015901	0.036236	0.000645	0.014123
Item_Visibility	-0.015901	1.000000	-0.006351	-0.083678	-0.128625
Item_MRP	0.036236	-0.006351	1.000000	0.000141	0.567574
Outlet_Establishment_Year	0.000645	-0.083678	0.000141	1.000000	-0.049135
Item_Outlet_Sales	0.014123	-0.128625	0.567574	-0.049135	1.000000

- Item_Weight has almost negligible correlation (1.4%) with the target variable.
- Item_Visibility is having nearly zero correlation (-13%) with the target variable Item_Outlet_Sales. This means that the sales are not affected by visibility of item which is a contradiction to the general assumption of “more visibility thus, more sales”.
- Item_MRP is positively correlated with sales at an outlet, which indicates that the price quoted by an outlet plays an important factor in sales. Variation in MRP quoted by various outlets depends on their individual sales.
- Outlets situated in location with type tier 2 and medium size are also having high sales, which means that a one-stop-shopping-center situated in a town or city with populated area can have high sales.

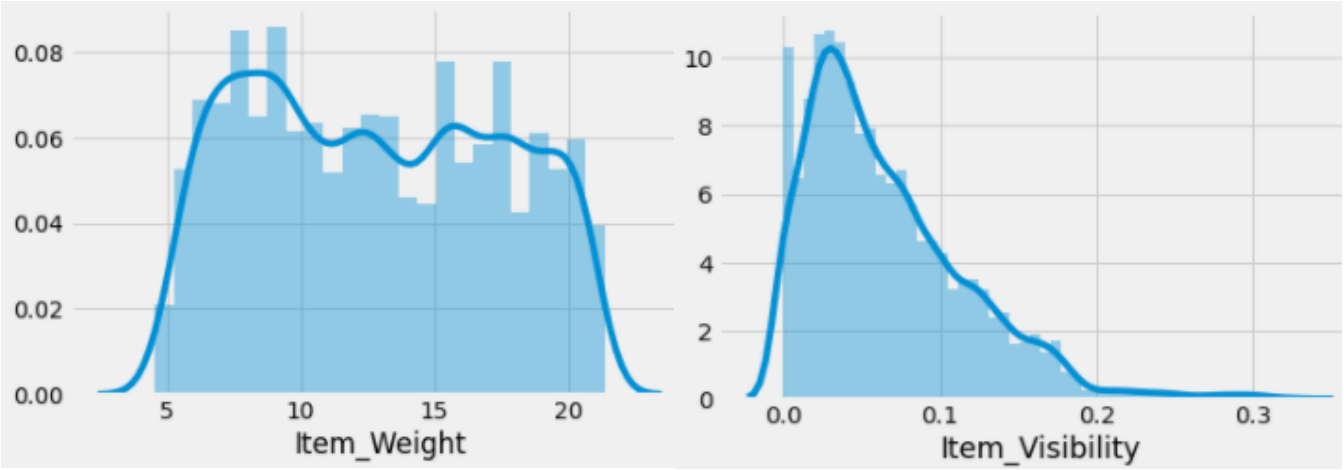
Univariate Analysis:

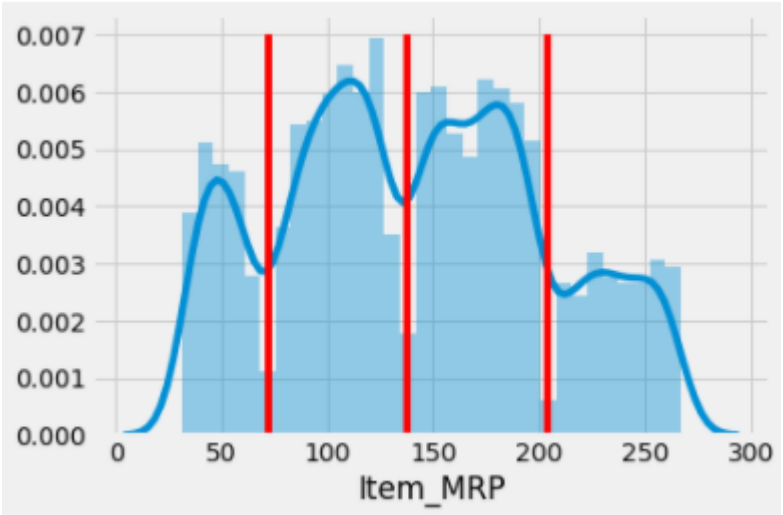
1. Distribution of target variable – Item_Outlet_Sales



- It is a right skewed variable and would need some data transformation to treat its skewness.
- Skewness > 1 which indicates that the distribution is highly positively skewed.
- Kurtosis > 1 shows that the distribution is leptokurtic.

2. Distribution of other independent variables

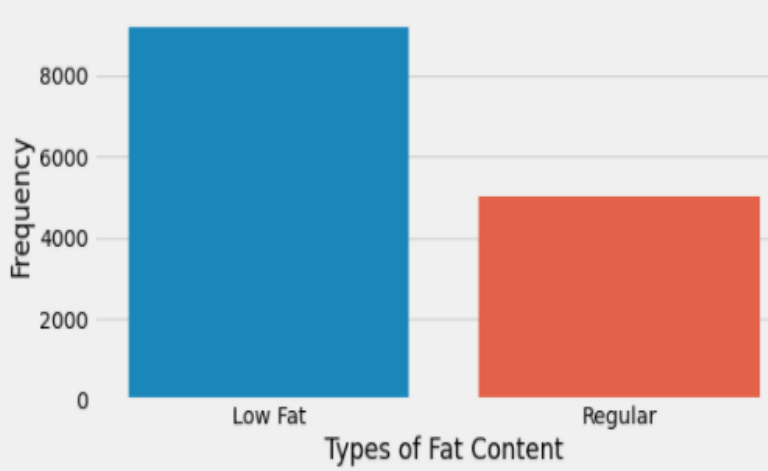




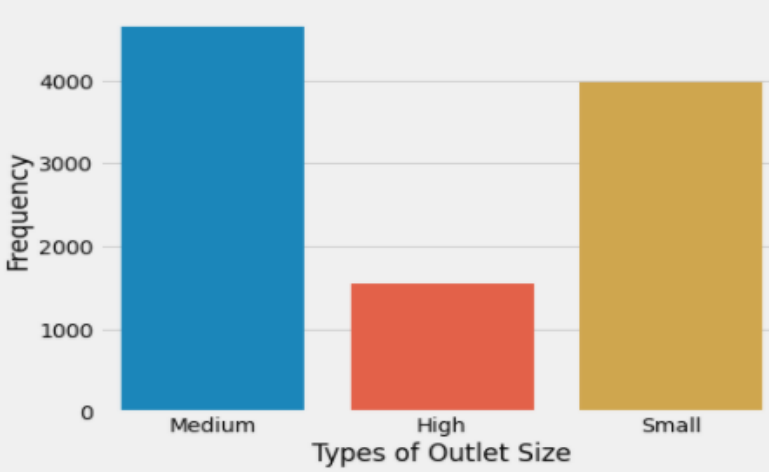
The Item_MRP clearly shows that there are 4 different price categories. So, I have defined them to be 'Low', 'Medium', 'High' and 'Very High'.

3. Categorical Variables

3.1. Distribution of the Item_Fat_Content

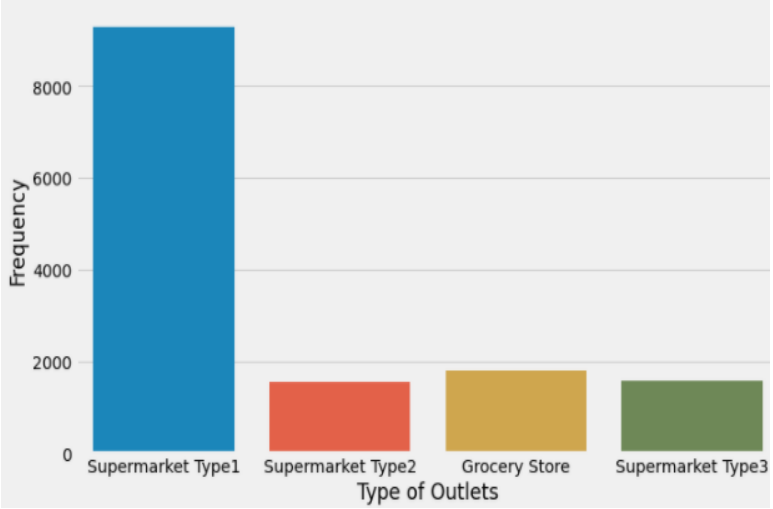


3.2. Distribution of the Outlet_Size

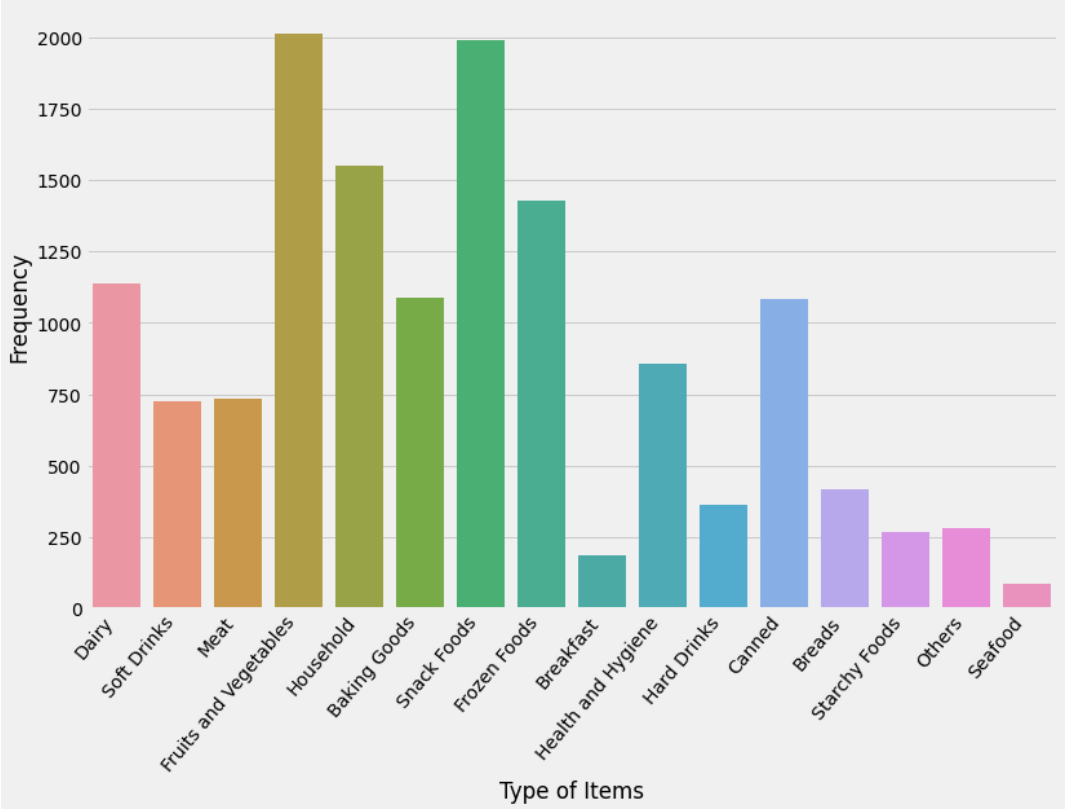


3.3. Distribution of the Outlet_Location

3.4. Distribution of the Outlet_Type



3.5. Distribution of the Item_Type



Bivariate Analysis:

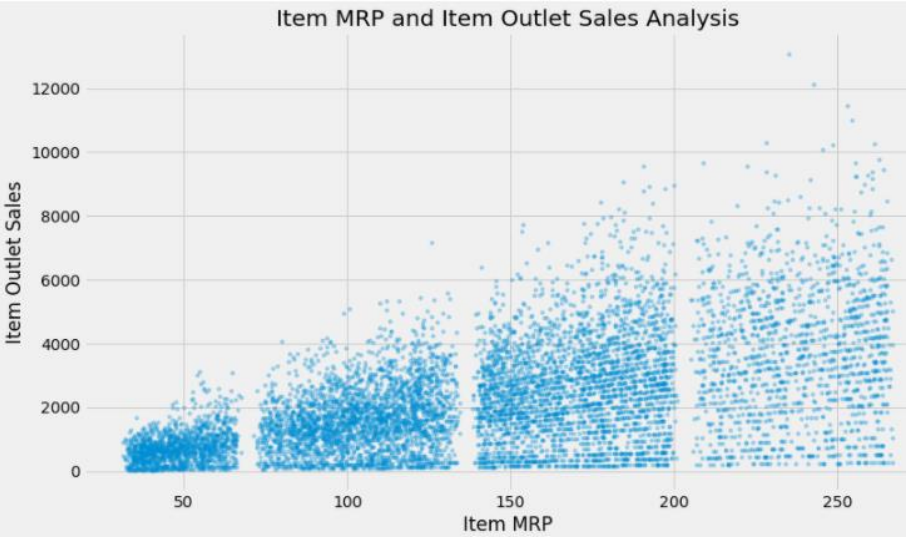
1. Item_Weight and Item_Outlet_Sales Analysis

- Item_Outlet_Sales is spread well across the entire range of the Item_Weight without any obvious pattern.
- Item_Weight is shown to have a low correlation with the target variable.

2. Item_Visibility and Item_Outlet_Sales Analysis

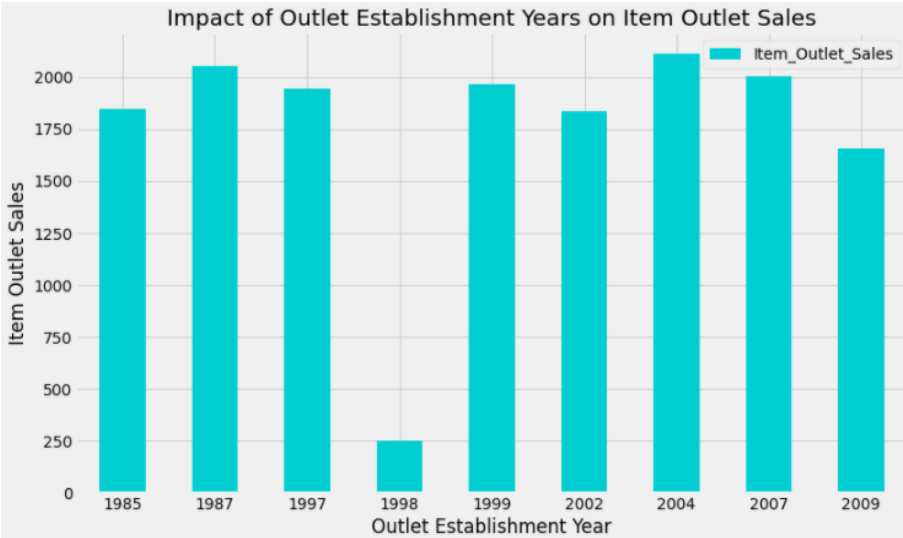
- Less visible items are sold more compared to more visibility items as outlet contains daily used items which contradicts the null hypothesis.
- There is a string of points at Item_Visibility = 0.0 which seems strange as item visibility cannot be completely zero.

3. Item_MRP and Item_Outlet_Sales Analysis



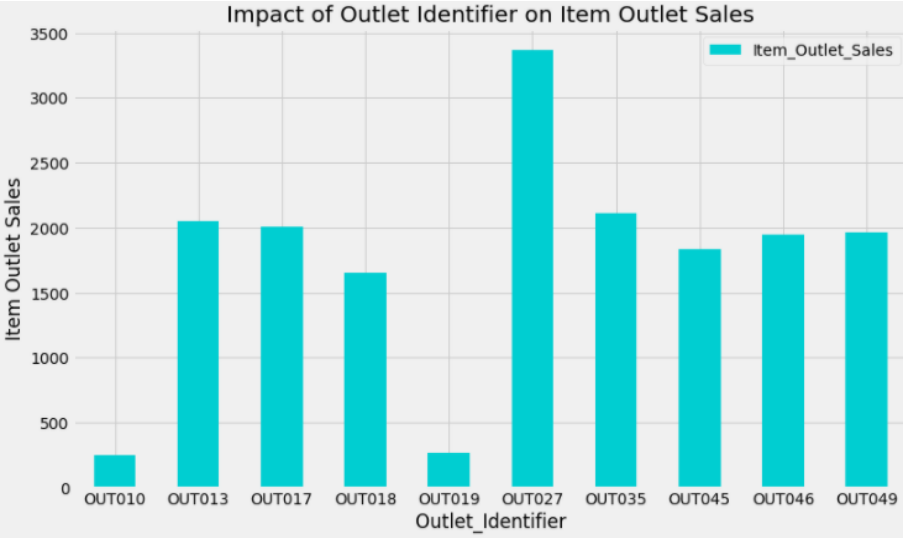
- We can clearly see that there are four segments of prices.
- The price range of MRP 150 to 250 has the highest range of products available.

4. Outlet_Establishment_Year and Item_Outlet_Sales Analysis



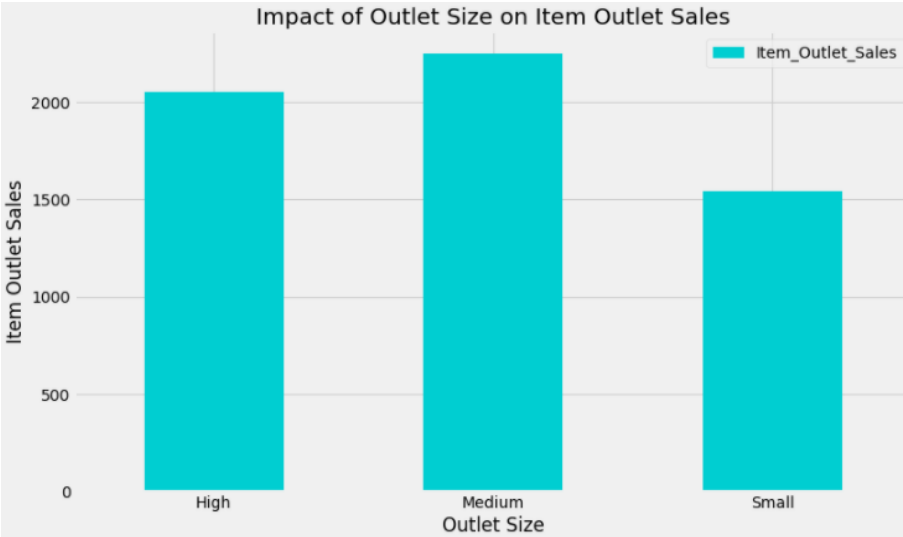
Every new outlet established at that particular year, surprisingly, has great sales except for the year 1998.

5. Impact of Outlet_Identifier on Item_Outlet_Sales



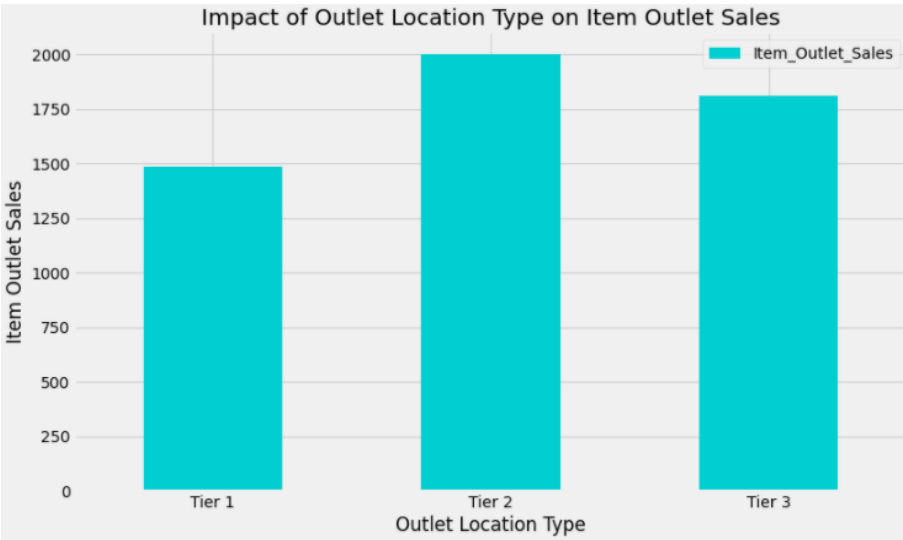
- The average sales are around 2000.
- 'OUT027' has the highest sales.
- 'OUT010' and 'OUT019' has a quite similar distribution depicting very smaller number of sales.

6. Impact of Outlet_Size on Item_Outlet_Sales



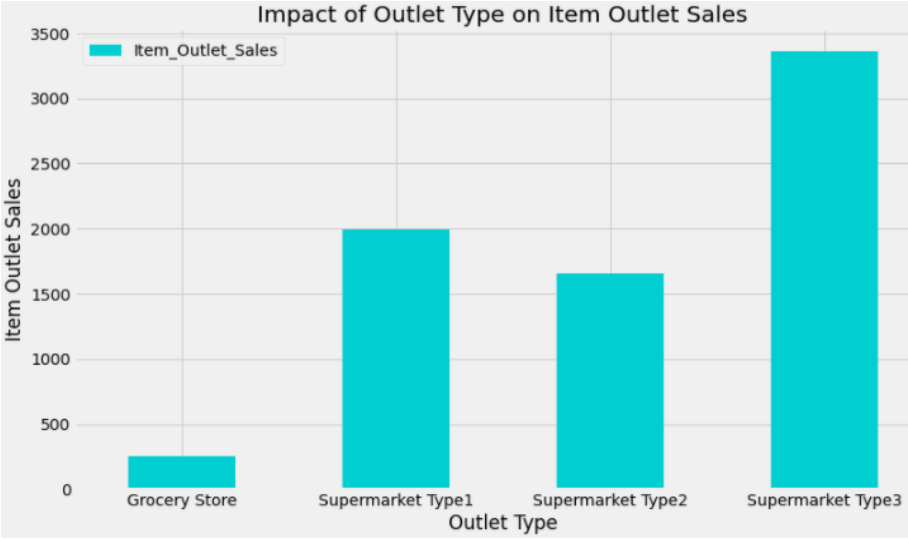
There is a very little difference between the sales of different outlets on the basis of the size of outlet i.e. the distribution is almost identical.

7. Impact of Outlet_Location_Type on Item_Outlet_Sale



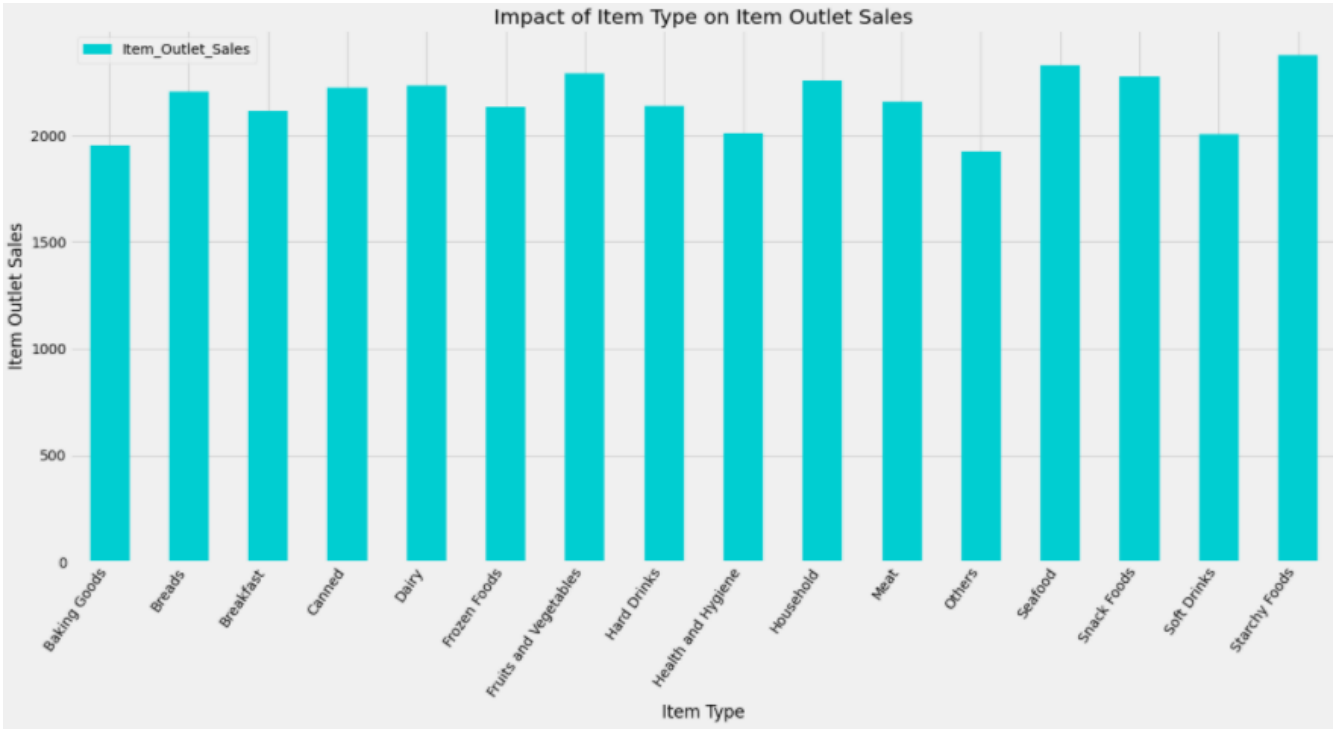
There is a very little difference between the sales of different outlets on the basis of the location of the outlet.

8. Impact of Outlet_Type on Item_Outlet_Sales



- Grocery Store has most of its data points around the lower sales values as compared to the other categories. Hence, we can say that it has the least sales.
- There is a very little difference between the sales of both Supermarkets Type 1 and Type 2, respectively.
- Supermarket Type 3 has the highest sales contribution in the organization.

9. Impact of Item_Type on Item_Outlet_Sales



Distribution of Item_Outlet_Sales across the categories of Item_Type is not very distinct.

Feature Engineering:

i. Modify Item_Visibility

The minimum value of variable is 0 which makes no practical sense.
Total number of 0 values initially: 879
Total number of 0 values after modification: 0

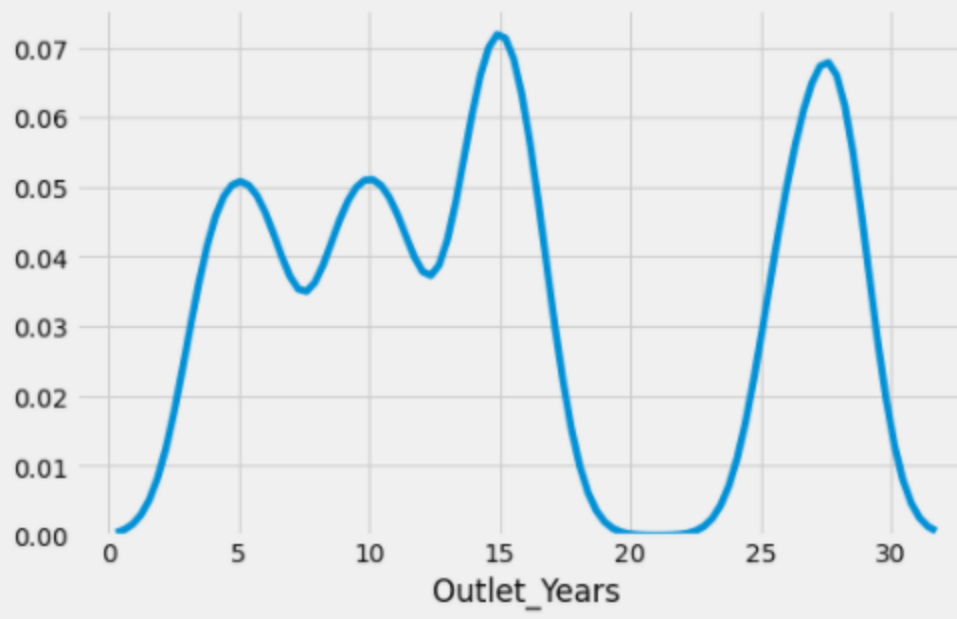
ii. Modify Item_Type

The Item_Type variable has 16 unique categories which might prove to be very useful in analysis. So, it is a good idea to combine them.
If we look at the Item_Identifier, i.e. the unique ID of each item, it starts with either FD, DR or NC. If we see the categories, these look like being Food, Drinks and Non-Consumables.

```
Food              10201
Non-Consumable     2686
Drinks             1317
Name: Item_Type_Combined, dtype: int64
```

```
Low Fat        6499
Regular        5019
Non-Edible     2686
Name: Item_Fat_Content, dtype: int64
```

iii. Year of Operations of a Store



Hence, as we can see, stores are 4 - 28 years old as compared from the year 2013, specifically.

Label Encoding and One Hot Encoding:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 14204 entries, 0 to 14203
Data columns (total 37 columns):
 #   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
 0   Item_Identifier             14204 non-null  object
 1   Item_Weight                 14204 non-null  float64
 2   Item_Visibility             14204 non-null  float64
 3   Item_Type                   14204 non-null  object
 4   Item_MRP                    14204 non-null  float64
 5   Outlet_Identifier           14204 non-null  object
 6   Outlet_Establishment_Year   14204 non-null  int64
 7   Item_Outlet_Sales           8523 non-null   float64
 8   source                      14204 non-null  object
 9   Item_Visibility_MeanRatio   14204 non-null  float64
 10  Outlet_Years                14204 non-null  int64
 11  Item_Fat_Content_0          14204 non-null  uint8
 12  Item_Fat_Content_1          14204 non-null  uint8
 13  Item_Fat_Content_2          14204 non-null  uint8
 14  Outlet_Location_Type_0      14204 non-null  uint8
 15  Outlet_Location_Type_1      14204 non-null  uint8
 16  Outlet_Location_Type_2      14204 non-null  uint8
 17  Outlet_Size_0               14204 non-null  uint8
 18  Outlet_Size_1               14204 non-null  uint8
 19  Outlet_Size_2               14204 non-null  uint8
 20  Outlet_Type_0               14204 non-null  uint8
 21  Outlet_Type_1               14204 non-null  uint8
 22  Outlet_Type_2               14204 non-null  uint8
 23  Outlet_Type_3               14204 non-null  uint8
 24  Item_Type_Combined_0        14204 non-null  uint8
 25  Item_Type_Combined_1        14204 non-null  uint8
 26  Item_Type_Combined_2        14204 non-null  uint8
 27  Outlet_0                    14204 non-null  uint8
 28  Outlet_1                    14204 non-null  uint8
 29  Outlet_2                    14204 non-null  uint8
 30  Outlet_3                    14204 non-null  uint8
 31  Outlet_4                    14204 non-null  uint8
 32  Outlet_5                    14204 non-null  uint8
 33  Outlet_6                    14204 non-null  uint8
 34  Outlet_7                    14204 non-null  uint8
 35  Outlet_8                    14204 non-null  uint8
 36  Outlet_9                    14204 non-null  uint8
dtypes: float64(5), int64(2), object(4), uint8(26)
memory usage: 4.6 MB
```

- Since, scikit-learn accepts only numerical variables, we need to convert all categories of nominal variables into numeric type variables. I have created a new variable Outlet same as Outlet_Identifier but encoded that.
- One-Hot-Coding refers to creating dummy variables, one for each category.
- For example, Item_Fat_Content has 3 categories - ‘Low Fat’, ‘Regular’ and ‘Non-Edible’. One hot coding will remove this variable and will generate 3 new variables instead of 1. Each will have binary numbers – 0 (if the category is not present) and 1(if category is present).
- Similarly, One-Hot Encoding has been performed on the variables, namely, 'Item_Fat_Content', 'Outlet_Location_Type', 'Outlet_Size', 'Outlet_Type', 'Item_Type_Combined' and 'Outlet'.

Modeling:

The following models have been used:

1. Linear Regression
2. Regularized Linear Regression
    2.1. Ridge Regression
    2.2. Lasso Regression
3. Decision Tree
4. Random Forest
5. XGBoost

- Mean Absolute Error (MAE) represents the difference between the original and predicted values extracted by averaging the absolute difference over the data.
- Root Mean Squared Error (RMSE) is the square root of mean squared error (MSE) which is the average of the square of the difference between the original and predicted values of the data. RMSE is basically the square root of the variance of the residuals.
- The RMSE indicates the absolute fit of the model i.e. how close the observed data points are to the model's predicted values.
- RMSE is a good measure of how accurately the model predicts the response and is the most important criterion for fit if the main purpose of the model is prediction. Lower values of RMSE indicate better fit.

Model	CV Score (Mean)	CV Score (Std)
Linear Regression	1129	43.72
Ridge Regression	1130	44.60
Lasso Regression	1129	43.64
Decision Tree	1091	45.42
Random Forest	1083	43.78
XGBoost	1163	52.12

Model	MAE	RMSE
Linear Regression	836.11	1127
Ridge Regression	836.03	1129
Lasso Regression	835.45	1128
Decision Tree	741.63	1058
Random Forest	748.31	1068
XGBoost	421.89	586.5

Conclusion:

As the profit made by the Big Mart is directly proportional to the accurate predictions of sales, they are desiring more accurate prediction algorithm so that the company will not suffer any losses. Xgboost has produced more accurate predictions as compared to the other available techniques like linear regression, regularized linear regression, random forest.
It is also concluded that XGBoost with lowest MAE and RMSE performs better as compared to the other existing models.