**BigMart Sales Prediction**

**Business Context**

Bigmart is a large supermarket chain, with stores all around the country. The data scientists at BigMart 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 is to build a predictive model and find out the sales of each product at a particular store. This will help the management take decisions like which type of products to concentrate on at a particular store for increasing sales and so on.

**Objective**

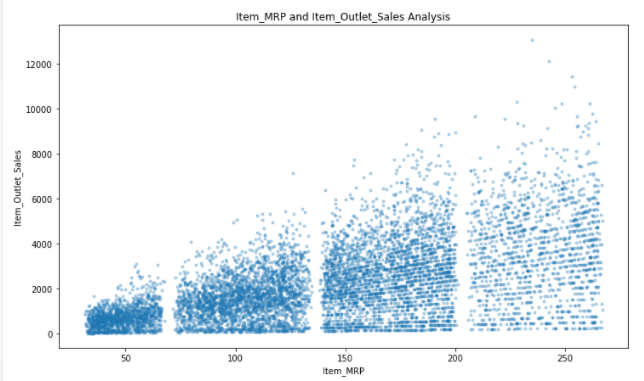
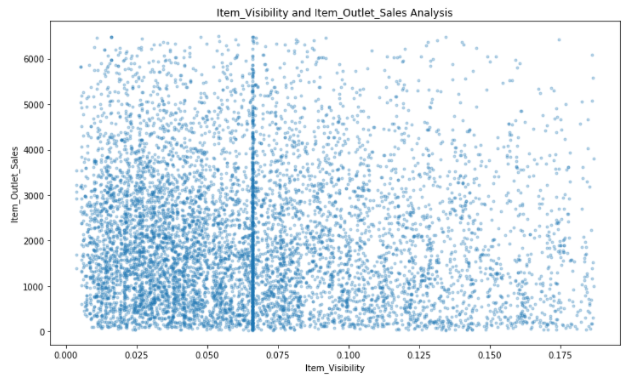
The aim is to build a predictive model and find out the sales of each product at a particular store. Using this model, BigMart will try to understand the properties of products and stores which play a key role in increasing sales.

This is a supervised machine learning problem with a target label as “Item\_Outlet\_Sales” Also since we are expected to predict a predict the sale price for a given product, it becomes a regression task.

**Dataset Information**

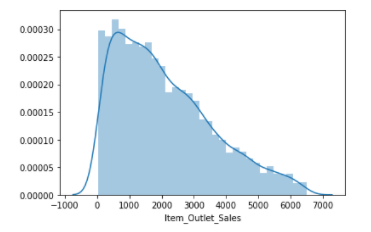
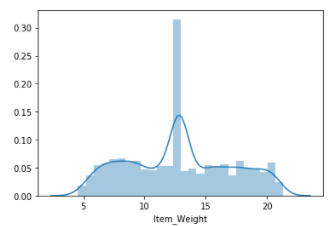
|  |  |
| --- | --- |
| **Variable** | **Description** |
| Item\_Id   |  |  | | --- | --- | | Item\_Identifier | Unique product ID | | Item\_Weight | Weight of product | | Item\_Fat\_Content | Whether the product is low fat or not | | Item\_Visibility | The % of total display area of all products in 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 |   entifier   |  |  | | --- | --- | | 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 | | 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 particulat store. This is the outcome variable to be predicted. |

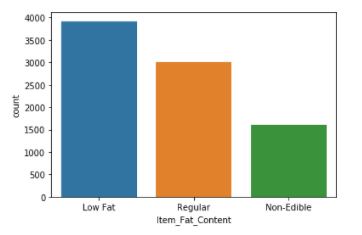
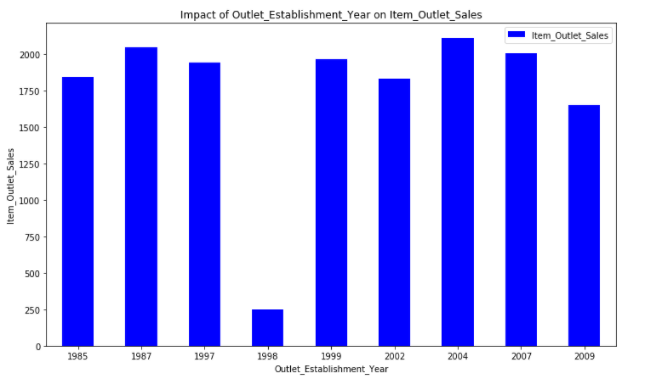
**Exploratory Data Analysis**



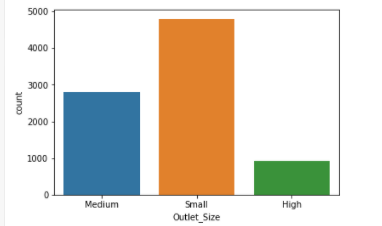
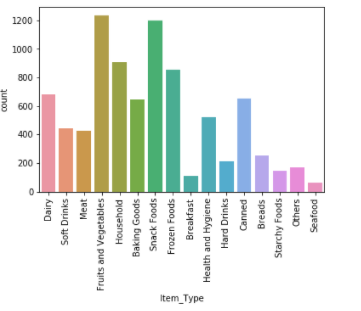
Item Visibility and Item Outlet Sales Item MRP and Item Outlet Sales

Density plots:





Outlet Establishment Year Item Fat Content



Item Type Outlet Size

**Outlier Detection and Treatment**

Number of Outliers in Item\_Weight = 0

Number of Outliers in Item\_Visibility = 99

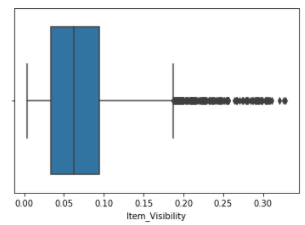
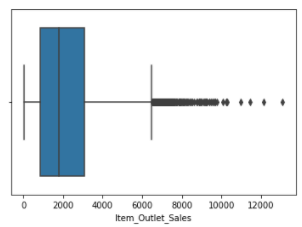
Number of Outliers in Item\_MRP = 0

Number of Outliers in Item\_Outlet\_Sales = 90

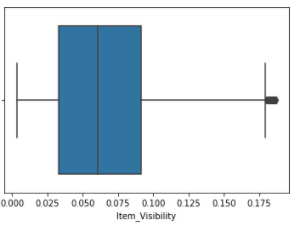
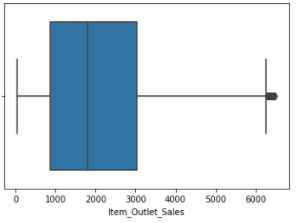
Number of Outliers in Outlet\_Years = 0

So, we have 2 features with outliers present. We will be removing these outliers from our dataset.

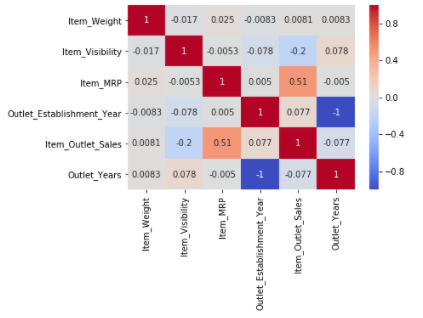
Before Outlier Treatment:

After Outlier Treatment:

Correlation Matrix



**Feature Engineering**

1. Item Weight: The missing values were imputed by mean of Item weights for a particular Item Identifier

2.Outlet Size: The missing values were imputed with mode (most frequent category) of Outlet Size for a particular Outlet Type (Grocery Store, Supermarket Type1, Supermarket Type2, Supermarket Type3)

3. Items for which Item\_Visibility was 0, was replaced with mean of Item\_Visibility column

4. The categories in Item\_Fat\_Content were renamed.

5. Categorical columns were converted to numeric using One-Hot Encoding.

|  |
| --- |

Creation of New Attributes:

1. A new feature called “New\_Item\_Type” was created by extracting the type of product from the Item\_Identifier. The categories were Food, Non-Consumable and Drinks.

2. A new feature called “Outlet\_Years” was created from “Outlet Establishment Year”. As the data was collected in 2013, the feature was created as,

Outlet\_Years= 2013 – Outlet\_Establishment\_Year

This will give meaningful data when fed to the model for better accuracy.

**Model Training and Evaluation:**

The ML models trained were:

1. Linear Regression

2. Ridge Regression

3. Lasso Regression

4. Decision Tree Regressor

5. Random Forest Regressor

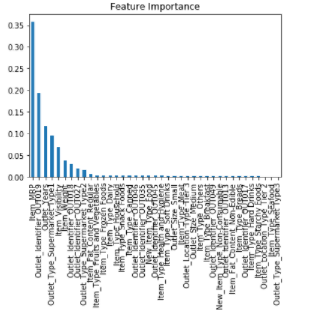
Random Forest Regressor gave the best results:

Model Report

MSE: 0.18512155196314878

CV Score: 0.28207932047494977

Feature Importances:



|  |  |
| --- | --- |