|  |  |  |  |
| --- | --- | --- | --- |
|  | 4/23/2017 |  | |
|  | |  | |
| BIG MART SALES PREDICTION  *BigMart 2013 sales data* | | | |
|  |  | |  |
|  |  | | Jaivignesh |

BIG MART SALES PREDICTION

BigMart 2013 sales data

**Problem Statement**

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.

Using this model, BigMart will try to understand the properties of products and stores which play a

key role in increasing sales.

Please note that the data may have missing values as some stores might not report all the data due

to technical glitches. Hence, it will be required to treat them accordingly.

**Data Exploration:**

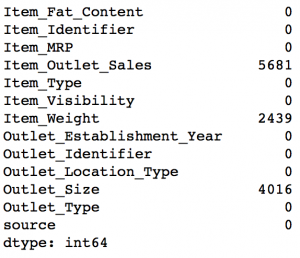
We have train (8523) and test (5681) data set, train data set has both input and output variable(s).

You need to predict the sales for test data set.

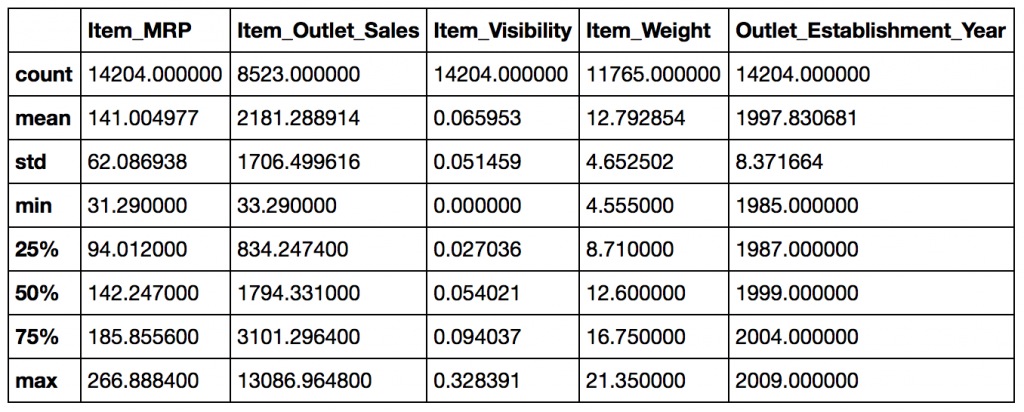
**Variable Description**

1. Item\_Identifier Unique product ID
2. Item\_Weight Weight of product
3. Item\_Fat\_Content Whether the product is low fat or not
4. Item\_Visibility The % of total display area of all products in a store
5. allocated to the particular product
6. Item\_Type The category to which the product belongs
7. Item\_MRP Maximum Retail Price (list price) of the product
8. Outlet\_Identifier Unique store ID
9. Outlet\_Establishment\_Year The year in which store was established
10. Outlet\_Size The size of the store in terms of ground area covered
11. Outlet\_Location\_Type The type of city in which the store is located
12. Outlet\_Type Whether the outlet is just a grocery store or some sort of
13. supermarket
14. Item\_Outlet\_Sales Sales of the product in the particulat store. This is the
15. outcome variable to be predicted.

**Missing Variables on Coloums**

[](https://www.analyticsvidhya.com/wp-content/uploads/2016/02/2.-missing.png)

Note that the Item\_Outlet\_Sales is the target variable and missing values are ones in the test set. So we need not worry about it. But we’ll impute the missing values in Item\_Weight and Outlet\_Size in the data cleaning section.



**Data Cleaning:**

This step typically involves imputing missing values and treating outliers. Though outlier removal is very important in regression techniques

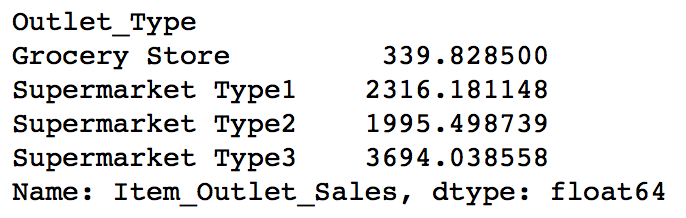
**Imputing Missing Values**

We found two variables with missing values – Item\_Weight and Outlet\_Size. Lets impute the former by the **average** weight of the particular item

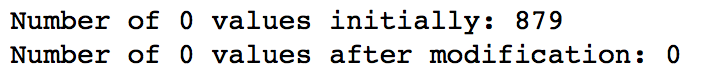
impute Outlet\_Size with the **mode** of the Outlet\_Size for the particular type of outlet.

Feature Engineering

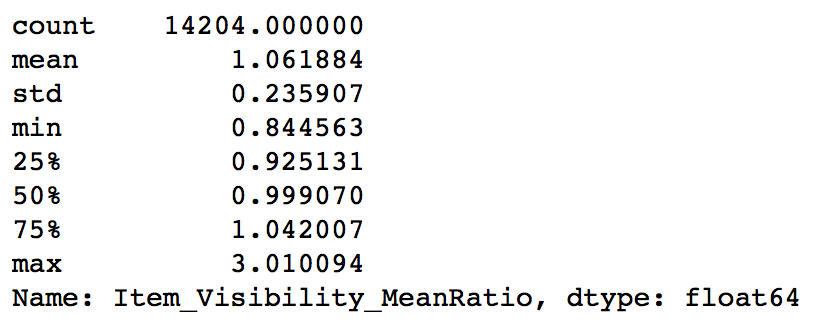
1: Consider combining Outlet\_Type

[](https://www.analyticsvidhya.com/wp-content/uploads/2016/02/9.-outlet-type.png)

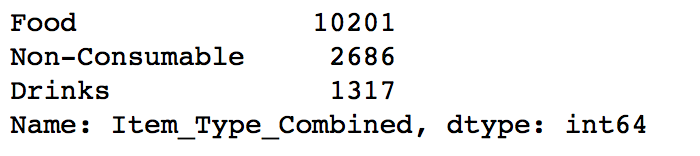
 2: Modify Item\_Visibility

[](https://www.analyticsvidhya.com/wp-content/uploads/2016/02/10.-visib-1.png)

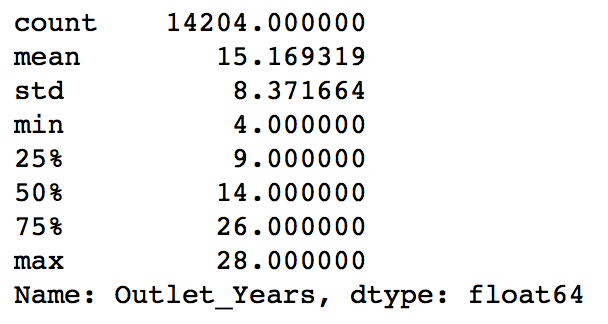
 3: Create a broad category of Type of Item

[](https://www.analyticsvidhya.com/wp-content/uploads/2016/02/11.-visb-2.png)

4: Determine the years of operation of a store

[](https://www.analyticsvidhya.com/wp-content/uploads/2016/02/12.-item_type.png)

5: Modify categories of Item\_Fat\_Content

[](https://www.analyticsvidhya.com/wp-content/uploads/2016/02/13.-years-of-oper.png)

**Exporting Data**

Final step is to convert data back into train and test data sets. Its generally a good idea to export both of these as modified data sets so that they can be re-used for multiple sessions.

1.Train.csv

2.Test.csv

**Model Building**

Trying Different Models for the Best Fit.

* Linear Regression Model:

RMSE:1129

* Decision Tree Model

RMSE:1058

* Random Forest Model

RMSE:1073

**CONCLUSION:**

We  data exploration where we found out some nuances in the data which required remediation. Next, we performed data cleaning and feature engineering, where we imputed missing values and solved other irregularities, made new features Finally we made regression, decision tree and random forest model for better results.