**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**

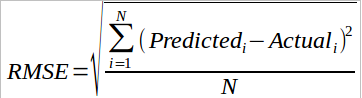
|  |  |
| --- | --- |
| **Variable** | **Description** |
| **Item\_Identifier** | Unique product ID |
| **Item\_Weight** | Weight of product |
| **Item\_Fat\_Content** | Whether the product is low fat or not |
| **Item\_Visibility** | The % of total display area of all products in a store allocated to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

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.

**Evaluation Metric:**

Your model performance will be evaluated on the basis of your prediction of the sales for the test data (test.csv), which contains similar data-points as train except for the sales to be predicted. Your submission needs to be in the format as shown in "SampleSubmission.csv".

We at our end, have the actual sales for the test dataset, against which your predictions will be evaluated. We will use the Root Mean Square Error value to judge your response.



Where,  
N: total number of observations  
Predicted: the response entered by user  
Actual: actual values of sales

Also, note that the test data is further divided into Public (25%) and Private (75%) data. Your initial responses will be checked and scored on the Public data. But, the final rankings will be based on score on Private data set. Since this is a practice problem, we will keep declare winners after specific time intervals and refresh the competition.

|  | | **Variables in Creation Order** | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **#** | **Variable** | | **Type** | **Var Type** | **Len** | **Format** | **Informat** |
| **1** | Item\_Identifier | | Char | NA | 5 | $5. | $5. |
| **2** | Item\_Weight | | Num | Cont | 8 | BEST12. | BEST32. |
| **3** | Item\_Fat\_Content | | Char | NA | 7 | $7. | $7. |
| **4** | Item\_Visibility | | Num | Cont | 8 | BEST12. | BEST32. |
| **5** | Item\_Type | | Char | NA | 21 | $21. | $21. |
| **6** | Item\_MRP | | Num | Cont | 8 | BEST12. | BEST32. |
| **7** | Outlet\_Identifier | | Char | NA | 6 | $6. | $6. |
| **8** | Outlet\_Establishment\_Year | | Num | Discrete | 8 | BEST12. | BEST32. |
| **9** | Outlet\_Size | | Char | NA | 6 | $6. | $6. |
| **10** | Outlet\_Location\_Type | | Char | NA | 6 | $6. | $6. |
| **11** | Outlet\_Type | | Char | NA | 17 | $17. | $17. |
| **12** | Item\_Outlet\_Sales | | Num |  | 8 | BEST12. | BEST32. |

Problems identified:

1. Item Weight has 17% missing values
2. MRP of the items are distributed as left skewed distribution – since mean value < median value (no problem)