ASSIGNMENT DA-3

Roll No: 41205

Problem Statement:

Bigmart Sales Analysis: For data comprising of transaction records of a sales store. The data has 8523 rows of 12 variables. Predict the sales of a store. Sample Test data set available here https://datahack.analyticsvidhya.com/contest/practice-problem-big-mart-sales-iii/

Objective:

- 1. Predict sales for a dataset
- 2. Understand data cleaning and regression

Outcome: One will be able to implement data cleaning and regression models

Pre-requisites:

- 1. 64-bit Linux OS
- 2. Programming Languages: Python

Hardware Specification:

- 1. x86 64 bit
- 2. 2/4 GB DDR RAM
- 3. 80 500 GB SATA HD
- 4. 1GB NIDIA TITAN X Graphics Card

Software Specification:

1. Ubuntu 14.04

Theory:

- Predictive modeling is a method of predicting future outcomes by using data modeling.
- It's one of the premier ways a business can see its path forward and make plans accordingly.
- While not foolproof, this method tends to have high accuracy rates, which is why it is so commonly used.
- Analyzing representative portions of the available information -- sampling -can help speed development time on models and enable them to be deployed more quickly.
- Once data scientists gather this sample data, they must select the right model.
 Linear regressions are among the simplest types of predictive models. Linear
 models essentially take two variables that are correlated -- one independent
 and the other dependent -- and plot one on the x-axis and one on the y-axis.
 The model applies a best fit line to the resulting data points. Data scientists can

use this to predict future occurrences of the dependent variable.

- Decision tree:
 - Decision tree algorithms take data (mined, open source, internal) and graphs it out in branches to display the possible outcomes of various decisions. Decision trees classify response variables and predict response variables based on past decisions, can be used with incomplete data sets and is easily explainable and accessible for novice data scientists.
- Time series analysis:
 - This is a technique for the prediction of events through a sequence of time. You can predict future events by analyzing past trends and extrapolating from there.
- Logistic regression:
 - This method is a statistical analysis method that aids in data preparation.
 As more data is brought in, the algorithm's ability to sort and classify it improves and therefore predictions can be made.

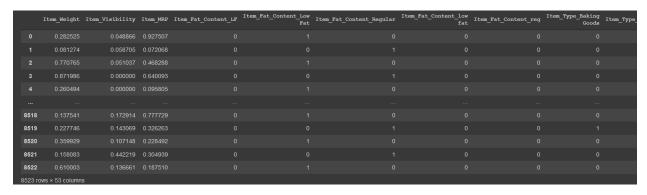
Input: Sales Dataset

Output:

Null Values in the dataset

```
Item_Identifier: 1559, NaN count: 0
Item_Weight: 416, NaN count: 1463
Item_Fat_Content: 5, NaN count: 0
Item_Visibility: 7880, NaN count: 0
Item_Type: 16, NaN count: 0
Item_MRP: 5938, NaN count: 0
Outlet_Identifier: 10, NaN count: 0
Outlet_Establishment_Year: 9, NaN count: 0
Outlet_Size: 4, NaN count: 2410
Outlet_Location_Type: 3, NaN count: 0
Outlet_Type: 4, NaN count: 0
Item_Outlet_Sales: 3493, NaN count: 0
```

Dataset after cleaning



Output metrics

MSE: 1160191.9712718462, RMSE: 1077.1220781656302

Conclusion: Thus, we have created a regression model to predict sales of big mart data.