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:

Initial dataset without cleaning

	Item_Identifier	Item_Weight	Item_Fat_Content	Item_Visibility	Item_Type	Item_MRP	Outlet_Identifier	Outlet_Establishment_Year	Outlet_Size	Outlet_Location_Type	Outlet_Type
0	FDA15	9.300	Low Fat	0.016047	Dairy	249.8092	OUT049	1999	Medium	Tier 1	Supermarket Type1
1	DRC01	5.920	Regular	0.019278	Soft Drinks	48.2692	OUT018	2009	Medium	Tier 3	Supermarket Type2
2	FDN15	17.500	Low Fat	0.016760	Meat	141.6180	OUT049	1999	Medium	Tier 1	Supermarket Type1
3	FDX07	19.200	Regular	0.000000	Fruits and Vegetables	182.0950	OUT010	1998	NaN	Tier 3	Grocery Store
4	NCD19	8.930	Low Fat	0.000000	Household	53.8614	OUT013	1987	High	Tier 3	Supermarket Type1
...	...	...	...	...	...	...	...	...	...	...	...
8518	FDF22	6.865	Low Fat	0.056783	Snack Foods	214.5218	OUT013	1987	High	Tier 3	Supermarket Type1
8519	FDS36	8.380	Regular	0.046982	Baking Goods	108.1570	OUT045	2002	NaN	Tier 2	Supermarket Type1
8520	NCJ29	10.600	Low Fat	0.035186	Health and Hygiene	85.1224	OUT035	2004	Small	Tier 2	Supermarket Type1
8521	FDN46	7.210	Regular	0.145221	Snack Foods	103.1332	OUT018	2009	Medium	Tier 3	Supermarket Type2
8522	DRG01	14.800	Low Fat	0.044878	Soft Drinks	75.4670	OUT046	1997	Small	Tier 1	Supermarket Type1

8523 rows × 12 columns

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

	Item_Weight	Item_Visibility	Item_MRP	Item_Fat_Content_LF	Item_Fat_Content_Low Fat	Item_Fat_Content_Regular	Item_Fat_Content_low fat	Item_Fat_Content_reg	Item_Type_Baking Goods	Item_Type_
0	0.282525	0.048866	0.927507	0	1	0	0	0	0	
1	0.081274	0.058705	0.072068	0	0	1	0	0	0	
2	0.770765	0.051037	0.468288	0	1	0	0	0	0	
3	0.871986	0.000000	0.640093	0	0	1	0	0	0	
4	0.260494	0.000000	0.095805	0	1	0	0	0	0	
...	...	...	...	...	...	...	...	...	...	
8518	0.137541	0.172914	0.777729	0	1	0	0	0	0	
8519	0.227746	0.143069	0.326263	0	0	1	0	0	1	
8520	0.359929	0.107148	0.228492	0	1	0	0	0	0	
8521	0.158083	0.442219	0.304939	0	0	1	0	0	0	
8522	0.610003	0.136661	0.187510	0	1	0	0	0	0	

8523 rows × 53 columns

Output metrics

```
MSE: 1160191.9712718462, RMSE: 1077.1220781656302
```

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