Date:-

Assignment - C3

Title :- Bigmart Sales Analysis.

Problem Statement :-

For data comprising of transactions records of a sales store, The data has 8523 rows of 12 variables. Predict the sales of the store.

Objective:-

To predict the sales for each item (product) per store for a particular supermarket chain.

Outcome :-

Identify products which play a key role in the sales of the supermarket chain (best and worst performing) to enable proper stategies to be put in Place to ensure the business. success.

Software & Hardware Requirements:-

Python 3, Jupyter, sklearn, madplotlib, UNIX / LINUX based OS, 64 bit CPU, 8 GB RAM, 128 GB SSD ; pandas, numpy.

Theory:-

The Bigmart sales Analysis (Prediction) is a supervised machine learning, regression task, where an algorithm is expected to predict the sale price for a given product and store.

There are multiple influencing factors on the sales of a particular product, mainly the product itself, and the tape of store it is being sold at.

A more in-depth analysis of the two main factors is as below.

Store level Hypothesis.

1) City type: stores in urban areas should have higher sales due to the high income household
2) Population density: density populated areas will have more sales.
3) store capacity.
4) Competitors.
5) Establishment year.

Product Level Hypothesis:

1) Item advertisement (visibility).
2) Item utility (type)
3) Price.

Exploratory Data Analysis showed that:

1) Item visibility did not have a high correlation (positive) as expected. It also had a lot of 0 values.

2). No huge variations in sales due to Item. Type either.

3) Item weight and Outlet_size have 0 values or NaN values.

4) Item_Fat_Content contains varying values for 'lowfat'.

5) Item_Type can be converted to a more useful feature.

- These values (missing, and NaN values) were imputed with the mean values for their respective columns, since keeping the values may result in incorrect or flawed predictions
- Item_weight, Outlet_Size, were imputed accordingly along with Item_visibility.
- Item_fat content and Item_Type were modified as mentioned before into (Food, Drink, Non-comsumable) and (low fat, regular) respectivly
- The categorical variables were then converted to numerical values since the python library for machine learning, scikit-learn, only accepts numerical values.
- One-Hot Encoding was used for the purpose it creates dummy variables, one for each type of category in a particular categorical variable.
- This can be done easily through the pandas function get_dummies.