Assignment - DAB

* TITLE: Bigmart Sales Analysis.

* PROBLEM STATEMENT: For data comprising of transaction 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 particular supermarket chain.

* OUTCOME: Identify products which play a key role in the sales of the supermarket chain to enable proper strategies to be put in place to ensure the business's success.

* SOFTWARE AND HARDWARE REQUIREMENTS: Python 3, Jupyter, Sklearn, matplotlib, UNIX/UNIX 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 algo is expected to predict the sale price for a given product & store.

There are multiple influencing factors on the sales of an particular product, mainly the product itself & type of store it is being sold at.

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

Store level Hypothesis:

i) City type: Stores in urban areas should have higher sales due to high income households.

ii) Population density: densly populated areas will have more sales.

iii) Store capacity
iv) Competitors
v) establishment year.

Product level Hypothesis
i) Item advertisement (visibility)
ii) Item utility (type)
iii) Price.

Exploratory Data Analysis showed that:
1) Item visibility did not have a high correlation as expected
2) No huge variations in sales due to Item_type either
3) Item_weight & outlet_size have 0 values or NaN.
4) Item_Fat_Content contains varying values for 'lowfat'
5) Item_type can be converted to a more useful feature.

- These values (missing & NaN) were imputed with mean values for their respective columns, since keeping the values may result in incorrect or flawed predictions.
- Item_weight, outlet_size, were imputed acc, along with Item_visibility.
- Item_fat_Content & Item_type were modified as (Food, Drink, Non-consumable) and (low fat, regular) resp.
- The categorical values were then converted to numerical values.
- one-hot Encoding was used for purpose; it creates dummy variables.

Linear Regression & Ridge Regression models were built to perform the actual prediction. Both models performed within the same range; giving a Root Mean Squared error of 1128 & 1129 resp.

Decision Tree model was then built, resulting in an improved RMSE of 1058.

Root mean Squared Error represents square root of second sample moment of differences b/w predicted and observed values, or the quadratic mean of these differences.

* CONCLUSION:

Successfully predicted Bigmart Sales using Linear, Ridge & Decision Tree regression models.