Assignment A3 41403

- Title: Big mart 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 per store for a particular supermarket chain.
- Outcome: Identify products which play a key role in the sales of the supermarket chain (best & worst performing) to enable proper strategies to be put in place to ensure the business's success
- SW & HW: Python 3, jupyter environment.

- Theory: The Big mart sales analysis is a supervised machine learning regression task where an algorithm is expected to predict the sale price for a given product & store.

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

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

41403

Store level hypothesis:
- City type: Stores in urban areas should have higher sales due to the high income households
- population density.
- Store capacity
- competitions
- establishment year

Product level hypothesis
- Item advertisement
- Item utility
- price

- Exploratory data analysis showed that
- item visibility did not have a high correlation as expected. It also has a lot of 0 values
- No huge variations in sales due to Item-type either
- Item-weight & Outlet size has 0 values or Non values
- Item_Fat_content contains varying values for lowfat
- Item-type can be converted to a more useful feature.

- These values (missing & 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 & item-type were modified as mentioned before into food, drink non consume & low fat regular resp.
- 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

- 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 regression model was then built resulting in an improved RMSE

Root mean squared error represents the square root of the second sample moment of the differences b/w Predicted & observed values or the mean quadratic of these differences.

- conclusion: Successfully predicted bigmart sales using linear, ridge & decision tree regression models.