**Delivery 3: EDA Implementing Machine Learning Algorithms**

- David Gill *(Data 606)*

**METHODOLOGY:**

1. Csv files were saved on my local machine
2. Python was used to perform data cleansing and exploratory data analysis

**PRELIMENARY EDA RESULTS:**

1. Stores file contains 45 Rows 🡪 which means there are 45 stores
2. It contains 3 Columns/Attributes - **Store, Type, Size**
3. There are 3 distinct Store Types - **A, B, C,** with **A** beingthe largest and **C** the smallest
4. There is no overlapped area in the sizes of the stores
5. There is no missing data in this file
6. Features file contains 8190 rows and 12 attributes, showing the features/factors affecting sales
7. Train data contains 421570 records and test data contains 115064 records
8. Sales on holidays was found to be a little bit more compared to non-holidays
9. The Department with highest sales lies b/w 60 – 80

**FURTHER EDA RESULTS AND DATA CLEANSING:**

1. Merged the Store, Feature and Train dataset
2. Plotted the correlation between the features and noted how they affect weekly sales.

* Higher departments have more sales (+ve correlation)
* Larger stores have more sales (+ve correlation)
* Markdowns/ discounts are weakly correlated and are not much significant
* Temperature, Fuel Price, CPI and Unemployment also do not affect weekly sales greatly

1. Markdowns with missing values are filled with 0
2. CPI and unemployment with missing values filled with their mean respectively
3. Holidays also do not have much of an affect on the weekly sales, though sales on holidays is a bit more than non-holidays.
4. Split Date into Day, month and Day, converting it in integer values
5. Testing data also merged with the features.
6. Testing and training datasets created choosing all the required attributes – Store, Size, Dept, Temperature, Fuel\_Price, CPI, Unemployment, MarkDown1, MarkDown2, MarkDown3, MarkDown4, MarkDown5, IsHoliday, Day, Month, Year, Weekly\_sales.
7. I have used 80% of my train dataset and 20% of my test dataset

**ML ALGORITHMS:**

1. Did some research and found sales prediction is most efficient using regression model, random forest, KNN (K-Nearest Neighbor), Extra Trees, etc.
2. I have used the following models –

* **K- Nearest Neighbor:**
* It is robust

- Its implementation is simple

- It is very useful in regression and classification problems

* **Random Forest:**
* It uses ensemble method
* It handles missing values
* It has higher accuracy
* **Extra Trees:**

**-** It is equally robust as Random Forest

- It is quicker than Random Forest

- The downside to this is the over-fitting trees

3) For testing the accuracy of my models, I have used validation metrics – Root Mean Square Error (RMSE), Mean Absolute Error (MAE)

4) The results obtained are as shown below-

|  |  |  |  |
| --- | --- | --- | --- |
| **MODEL** | **ROOT Mean square error** | **Mean absolute error** | **Accuracy** |
| **KNN** | **39.81** | **32.85** | **43.25** |
| **Random Forest** | **22.31** | **14.03** | **76.55** |
| **Extra Trees** | **10.1** | **1.5** | **98.81** |

5) I am still in the process of training the model further to get more accurate results and cater for the extra trees. Based on the inference made I will produce the predicted result set.