**Data Science Challenge**

**Approach Document**

*1. Before writing any code, how would you approach this challenge? Create a word document in which you describe in a few short sentences your plan.*

Step 1: Peek into the data manually.

Step 2: Note down key observations.

Step 3: Create an EDA notebook to materialize your thoughts from Step 1 and 2.

Step 4: Identify correlations.

Step 5: Handle null / missing values

Step 6: Formulate ideas about feature engineering

Step 7: Based on intial judgement, choose a baseline model.

Step 8: Write code for the baseline model.

Step 9: Evaluate it’s performance.

Step 10: Formulate ideas about tweaking model.

Step 11: Apply the ideas from Step 10 to the model in a new notebook.

Step 12: Evaluate new model’s performance.

Step 13: Compare the accuracy of two classification models on the same dataset.

**Hypothesis**

1. Null Values in Outcome Columns

Assumption (1): The null values in outcome columns are unreported or missing values.

Assumption (2): For the cells where outcome is null, the product was still on the shelf and selling.

2. Hypothesis testing for comparing models in Step 13:

Compare the accuracy of two classification models on the same dataset.

Hypothesis:

H₀: There is no difference in the performance of the two models.

H₁: There is a difference in the performance of the two models.

**Observations**

* Random Forest seemed like a good place to start as there were:

- A lot of features (comparing with the number of rows)

- And a lot of null values

* Random Forest takes care of both (viz. Feature Importance and Handling Null Values)
* Any study done on the dataset after Random Forest did not improve the performance.
* The other model tried was: Multi Layer Perceptron from scikit-learn.
* Prediction is done for Outcome\_M1.
* Outcome\_M2 is correlated with Outcome\_M1. Outcome\_M3 with Outcome\_M2, so on, so forth.

PTO

**Evaluation And Results**

|  |  |  |
| --- | --- | --- |
| # | Model | Mean Squared Error  Note: **These are not cross-validated.** |
| 1 | Random Forest | 1843739464.84375 |
| 2 | MLP (No Scaling, Fillna(0)) | 2723151302.5054903 |
| 3 | Scaling in MLP is increasing the error | 3095528419.9108686 |
| 4 | MLP (No Scaling, Fillna(median)) | 2951453700.45971 |
| 5 | MLP (With Scaling. Fillna(median)) | 3040307598.6931763 |
| 6 | MLP after dropping columns with too many NAs | 3206736619.740567 |

**Instructions to Run The Code**

1. Assumptions: Anaconda Distribution is installed on your machine.

2. Unzip the code.zip file in a folder.

3. Launch Jupyter Lab or Jupyter Notebook in this ‘code’ directory.

4. Now you use functionalities like ‘Run All Cells’ or ‘Run Cell’ to execute the Jupyter notebooks.