Programing – Hands-On Assignment

Note:

1. This assignment is to be accompanied by the 'sales_pred_case.zip' dataset. The unzipped contents should be a single csv formatted file.
2. Total Time allocated for this assignment is 24 hours (starting hand-over to the candidate; you will need ~2-3 dedicated hours to complete the case).
3. The submission should be a single executable Jupyter notebook.
4. Following criteria will be used for assessment:
    a. Code Quality: Readability, Conciseness, CPU, Memory & Time Efficiency.
    b. Use of Standard Libraries: Libraries should be installable through pip.
    c. Case Adherence: Code should address the problems given i.e., no filler code.
    d. Algorithm Quality: How well suited is the algorithm to the problem?
    e. Result Quality: This is the model accuracy obtained on held out samples.
    f. Scope for improvement: Can the model be improved with further tuning.

Data Description:

The csv file 'sales_pred_case.csv' contains sales data for ~1000 Material & Customer Pairs gathered over ~3 years at a weekly granularity. The following is the data dictionary:

1. "Key": A concatenation of Material & Customer codes. There are about 970 Keys & predictions are to be made against these.
2. "Material", "Customer", "CustomerGroup"," Category": Label encoded strings with obvious meanings.
3. Columns 'H' – 'N': Common time & holiday features stored either as integers or one-hot values.
4. Columns "O"- "T": Sales promotion related features. "DiscountedPrice" is a float column while the rest are categorical columns.
5. "Sales": The target column; can be treated as float.
6. "YearWeek": 4-digit year + 2-digit week no. concatenation. This is the time index for the dataset.

Problem Statement:

For each of the keys predict the sales for the weeks starting from "2022-46" – "2023-02" (both weeks included). You may use data on or before "2022-45" in any manner to come up with train/test/validation split scheme for training & evaluating the model(s).

You can use any algorithm for this exercise – Time Series, Traditional ML or Deep Learning.

Accuracy Metric:

Accuracy of predictions will be estimated using the Weighted MAPE (WMAPE) metric. This is calculated (across all the keys & prediction periods) as:

Accuracy = (1- SUM(Absolute Error)/SUM(Sales)) where, "Absolute Error" = |Sales - Prediction|

In addition, Bias will be calculated as well:

Bias = (SUM(Sales)/SUM(Prediction) -1)

The objective is to get as high an accuracy with as low a bias as possible.

Additional notes:

1. **You must use Python for this assignment.**
2. **You must only use open-source, standard libraries.**
3. **You may use any ML or DL package (Scikit-learn, TensorFlow, Pytorch etc.)**
4. **Plots are optional.**
5. **Feature Engineering steps, if any, should have clear comments.**
6. **Finally, do justify your choice of Model, Loss function & include general observations about the data, further improvement methods etc. as part of conclusion in the bottom cell of your notebook.**