Production Prediction Challenge

Motivation: It's always nice to have an accurate prediction of the revenue a business makes for the subsequent month(s). This would help the business plan accordingly and adjust their investments and time. For a dental practice, prediction of revenue is equivalent to prediction of production. An accurate forecast would help the practice in inventory planning, capacity planning, planning working hours, hiring and many more factors.

Problem Statement: Given the historical data of monthly production for various practices, predict the monthly production for the period Jan 2021 to April 2021 (inclusive) which would be 4 data points per practice.

Model Requirements: You can build a single model that can predict the monthly production of all the practices for 4 months **OR** you can build multiple models and predict the monthly production in the 4 month period for each practice. The choice of modelling is yours.

Data:
The dataset includes monthly production data for 284 practices (unique ids) with the features as mentioned in the following section.

Features:
id: The id which indicates a unique practice.
month: Month of the year for the production value
year: Year for the production value
production: It contains a decimal value which indicates the total production for that month, year combination
visits: Number of visits for that month and year combination for a particular practice identified by *id*.
no_of_appts: Number of appointments made for that month and year combination for a particular practice identified by *id.*

Metric:
MAPE (Mean Absolute Percentage Error)

Definition and Formula:
https://www.statisticshowto.com/mean-absolute-percentage-error-mape/

Code to calculate MAPE:

```
import numpy as np
def mean_absolute_percentage_error(y_true, y_pred):
    return np.mean(np.abs((y_true - y_pred) / y_true)) * 100
```

where y_true (true response) and y_pred (predicted response) are numpy arrays of the same size.

Let's look at an example -
Let's say you predict 4 production values for a practice as [100, 200, 300, 400] and the actual values of production are [120, 205, 315, 390] . The MAPE value using the code is 7.5% - which means the model on an average has an error of 7.5% for every prediction.

We calculate the MAPE for every practice to assess how your model would perform. Since there are 284 practices and you predict for 4 months, we would have 284 MAPE values. For instance,

MAPE for id = 1 is 4.5%
MAPE for id = 2 is 16% and so on.

The final score is calculated as follows:

Every MAPE value gets a reward (positive points) or a penalty (negative points) which can be either 2, 1, -1, or -2. Based on the MAPE values, you either get a reward or a penalty based on the criteria defined below:

1. If the MAPE value is <=5% then the reward is 2. If 40 out of 284 practices have MAPE <=5% then the score would be **224/284 * 2 = 0.78**
2. If the MAPE value is >5% and <=10% then the reward is 1. If 30 out of 284 practices have MAPE in the range mentioned then the score would be **30/284 * 1 = 0.105**
3. If the MAPE value is >10% and <=15% then the penalty is -1. If 20 out of 284 practices have MAPE in the range mentioned then the score would be **20/284 * (-1) = - 0.07**
4. If the MAPE value is >15% then the penalty is -2. If 10 out of 284 practices have MAPE in the range mentioned then the score would be **10/284 * (-2) = -0.07**

The final score is the summation of (1) + (2) + (3) + (4). Looking at the numbers for the example, it would be **0.78 + 0.105 + (-0.07) + (-0.07) = 0.745**

The maximum score you can get is 2: all the MAPE values are <=5%.
The minimum score you can get is -2: all the MAPE values are >15%.

Tie Breaker: If the final score is the same, the average MAPE across all practices will be used as a tie-breaker.

Submission: You will have to submit 3 things as described below:

result.csv
Please provide a CSV file with the following fields
Id: The id of the practice
month: The month for which the prediction is. Please keep it numeric (1/2/3/4) and do not have it as a string like “Jan”, “Feb” etc.
year: The year for which the prediction is (has to be 2021) Because you are predicting for 2021
production: The predicted production value for the month and year combination

writeup.pdf - A file explaining how you came up with the solution. There is no pre-defined format for this. You are free to use whichever format you prefer.

The github link - The repository which contains the code for the solution.