Documentation

1. **Approach**:

In the problem the test data was on monthly basis so I grouped by the train data to make it on monthly basis. Then I joined the expense column corresponding to each product and its country on month, corresponding NANs coming after joining were filled with 0.0, as there must be no expenses made on the product if no data were given. Also, I created the no. of holidays in a month to check the effect of holidays on sales.

Then I transformed country variable into numerical feature to use this feature in my model by using sklearn Label Encoder.

I created 5 models i.e. for each product a model was made and the prediction was done.

1. **Feature Generation/Data Pre-processing**:

I have formed a new features – Expense\_Price.

I have formed this new features from existing data ‘promotional\_expense’. This feature have been formed by joining the promotional\_expense to train data on the basis of product and country on monthly basis.

Expense\_Price – It represents the promotional expenses done on the product for that month.

1. **Key Observations/Trends**:

My manual interpretation I observed that for some weeks particular product’s Sales value was negative so I made them positive and trained my model.

1. **Model Selection**:

I created 5 models, modelling is done on the product level then, I used Grid Search CV for the best model given by XGBoost Regressor.

I tuned the parameters intuitively.

1. **Expected error for submission**:

Error may occur while reading the train and test file (‘yds\_train.csv’ and ‘yds\_test.csv’) if proper current directory is not specified.