**Data Pre-processing**

**Table: holidays**

1. Year, Month columns obtained from Date column
2. Performed grouping based on year, Month and Columns and number of holidays are calculated for each sample in the group.
3. Further columns Holiday1, Holiday2…. Holiday 31 are generated to indicate whether a day is holiday for a particulate country, month and year. It’s preserves more details on aggregation.

**Table: yds\_train2018**

1. As the given data is in the weekly basis and we need monthly observation, grouped by columns Year, Month, Product\_ID and Country and weekly Sales are added to form monthly total.
2. Merchant\_ID columns dropped as it doesn’t give appropriate aggregation property.

**Table: promotional\_expense** already presents monthly observation

**Merging tables**

1. yds\_train2018 and holiday tables are joined on column Year, Month and Country. And the resultant table joined with promotional\_expense table on columns Year, Month, Product\_ID, Country

**Missing value treatment**

1. Missing values in Holiday column are filled with zero as it is making sense of no holidays.
2. Local currency columns Sales and Expense\_Price are converted into USD with the monthly average currency rate of observed month.

(Historical exchange rate is obtained from <https://www.x-rates.com/average>)

1. XGboost Regressor model with fully grown tree is applied to predict missing values of column Expense\_Price. Here overfitting is not a big issue as there won’t be a new sample to be predicted expect missing values.

The same process is also applied for table yds\_test2018 except predicting Expense\_Price. As Sales values need to be predicted for test data, we don’t have relation of columns Sales and Expense\_Price.

* Here missing Expense\_Price is filled with average of Expense\_Price filtered by Country and Month.

**Converting categorial columns to numeric**

1. Applied one-hot encoding on columns Country and Product\_ID.

Date columns is dropped as it’s already available as Year and Month Columns.

Now we have all numeric data with required transformation and no missing values.

**Prediction Approach**

As we have time series data, we can go with time series forecasting. But the size of dataset may not help with that. Further feature sets of data leverages supervised approach. And the correlation between sales and Expense\_Price columns adds additional point to supervised approach.

As per the experiment with dataset, it’s more sensible to tree-based algorithms.

The predictive model is a simple ensemble of Random Forest, Gradient Boosting algorithms.

Ensembled Model

Random Forest Regressor

Gradient Boosting Regressor

XGBoosting Regressor

The predicted values are converted back into local currency based on available country, month, year details.

Refer **ReadMe.txt** file to see how to run the program.