**Process document for CCI python code:**

|  |
| --- |
| Main Objective of this code is to generate ensemble predicted and forecasted results, which uses three main approaches named Ordinary Least Squares regression (OLS), Random Forest Regressor (RF), Artificial Neural network (ANN).  Ensemble takes a weight of individual predictions to form a final prediction. |

Code can run 15 states at a time for a particular category.

There are some formats that need to be followed before running the code.

There are 3 input files where order of excel sheets across all 3 files should be same.

i.e. ['WB','DE', 'MH', 'AP', 'UP', 'PU', 'KA', 'GU', 'RA', 'MP', 'OR', 'BI', 'CH','HA','TN']

Input files are CCI raw data, CCI ensemble data, CCI Coefficient data.

Input variables should be in the below format across all states. [without any space]

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **States** | **Date** | **Tmax** | **Avg\_precip** | **Economy or**  **(log\_economy)** | **Unemployment** | **Category Sales** | **Dummies if any** | **VolSales (KO sales)** |

Coefficient file should also follow the below format: Eg.

|  |  |
| --- | --- |
| FactorName | Estimate |
| Intercept | -1425992.59 |
| Tmax | 6846.6575 |
| Avg\_precip | -564.76 |
| log\_economy | 6678892.6 |
| Unemployment | -1093.56 |
| Category | 1.9656 |
| Dummies if any |  |

|  |
| --- |
| * First, import the required libraries such as pandas, numpy. |
| * Reading the Input files. [Raw data, transformed data (i.e. variables used in OLS technique and we are feeding the same variables to the RF and ANN approaches) and coefficient file]. |
| * Splitting the data set into training and testing (Code automatically picks last 3 months for the validation purpose). |

|  |
| --- |
| * First approach we used is OLS Technique which calculates Y=mx+c. Also, MAPE formula is used in order to check the accuracy of the OLS model.   MAPE Formula= [ Average(abs(actual-predicted)/actual) \*100] |
| * Defined a function to calculate MAPE for both training and testing data set and then this function is called in RF and ANN approaches |
| * Used r2\_score library in order to know proportion of the variance in the dependent variable that is predictable from the independent variable. |

|  |
| --- |
| * Second approach is Random Forest where we are using sklearn package and RandomForestRegressor library. * The only parameter used is: Random\_state   In order to forecast the values, we assuming entire dataset as training and then forecasting the results using predict function.  Final MAPE formula = [((mape of training data set\* count of training data set) + (mape of testing data set \* count of testing data set)) / Total count of data set] |

|  |
| --- |
| * Third approach is Artificial neural network where we are using sklearn package and MLPRegressor libr2ary.   The parameter used are: hidden\_layer\_sizes, max\_iter, batch\_size, verbose and  random\_state.   * In order to forecast the values, we assuming entire dataset as training and then forecasting the results using predict function.   Final MAPE formula =[(mape of training data set\* count of training data set) + (mape of testing data set \* count of testing data set) / Total count of data set] |

|  |
| --- |
| * Once we have predicted results for all 3 approaches, we are calculating weights for each algorithm based on their corresponding calculated mape’s.   Formula used is: Weight(i)= σ^−2 ( i ) / Σ σ^−2( j )  These weights then multiplied with predicted results of each approaches to get the ensemble predicted and forecasted results. |

############################## Due to Calculations #############################

In due to calculation we are only considering 4 variables: Tmax, Avg\_precip, Economy and Unemployment.

X: Tmax, Avg\_precip, Economy and Unemployment.

* Written a for loop to perform certain calculations across all variables:
* To calculate Diff YOY/MOM raw data is considered

|  |  |
| --- | --- |
| (X) Diff\_YOY : | Year over Year difference  Eg: Tmax(current year)- Tmax(previous year) |
| (X) Diff MOM : | Month over month difference |

* To perform below operations OLS transformed data are used:

|  |  |
| --- | --- |
| DT (X) | Beta \* X(current year) -Beta\*X(previous year) |
| Sales Index | Average sales of the particular month **/** Average of sales for entire time period |
| DT(X) using sales Index | Sales Index\* DT(X) |

* 3MMT (VolSales, Ensemble, Category etc) : 3 months Rolling sum of particular variables.
* 6MMT (VolSales, Ensemble, Category etc) : 6 months Rolling sum of variables.
* Few variables which are required for the Power BI tool are fed into power BI data frame.
* Once after these calculations, melt function is used to convert values into long format.

And finally writing these values into excel sheet.