HACKATHON CHALLENGE - CMPE 274

Team Members:
Arpitha Srinivas - 015908880
Ravi Shanker Thadishetti - 016011983
Indhupriya Reddem - 015930148
Vineeth Reddy Govind - 015363556

Note: We have attached our entire python code in a separate file.

We have followed the KDD steps in solving this hackathon challenge.

KDD Steps:

1. Reviewed Problem statement from PotatoCo and Business objectives.
Reviewed various data sources that have been shared and reviewed all the reference material to understand the business domain.

2. Source Data:
We have downloaded all the required data sources for this challenge and also downloaded the time series data set from the FOASTAT and additional data set from worldbank.org.
Downloaded all the data from the assignment and also downloaded additional dataset from FOASTAT (climatology) and worldbank.org

3. Data Preparation:

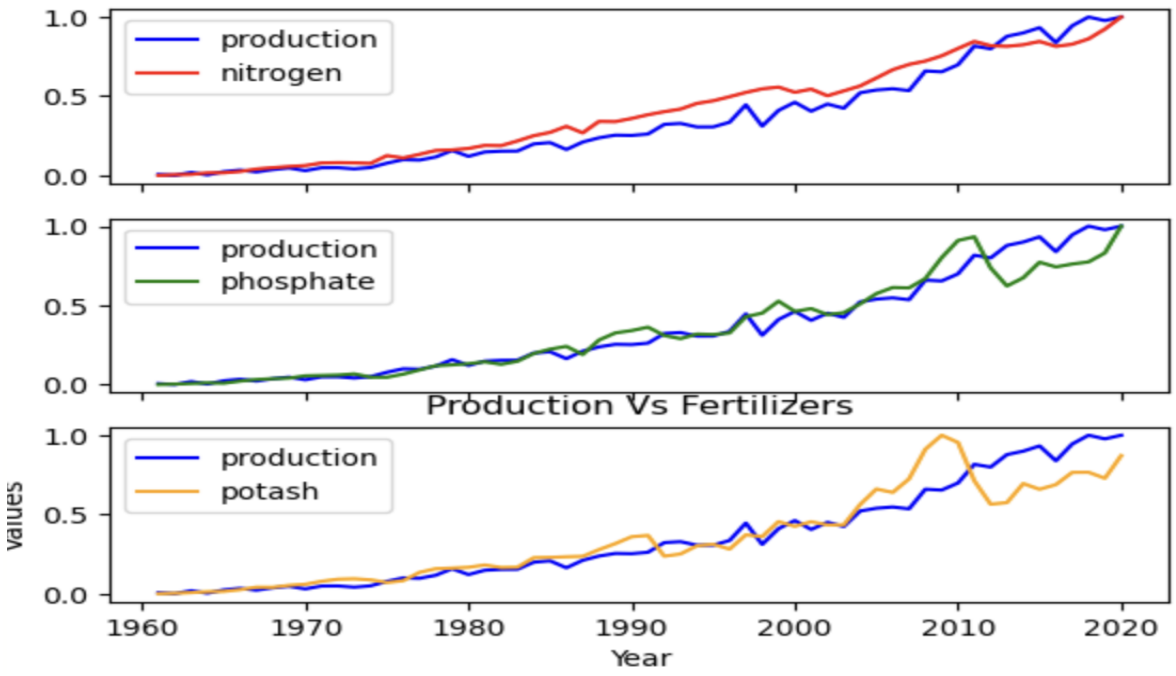
a. **Load the data set:** we have loaded the fertilizers, emissions and co2 data into a data frame.
b. **EDA :** we have completed the EDA on the three datasets.
c. **Data Cleaning:** Identified the null values for each feature in the given data sets and performed the data cleaning on this dataset and observed that the yield data set has no null values. We have preprocessed three data sets by removing null values. We observed that the yield data set doesn't contain null values.

4. Feature Engineering:

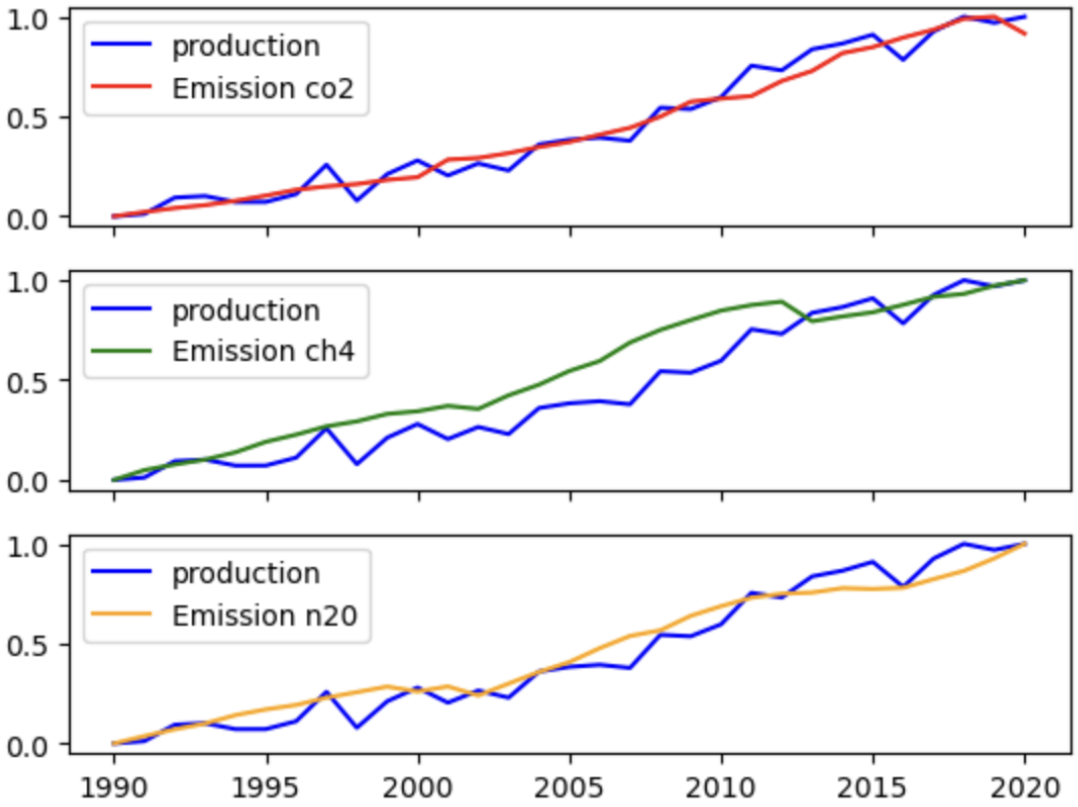
- Identified the unique features in the yield data set and removed the unwanted/not relevant features in the yield dataset.
- We made an observation that the main features for the yield data set are year, element and value.
- Similarly we found that the main features for the emissions data set are Domain, Year, Element, Item, Value.
- For the fertilizers data we have identified the main features domain, item, year, value.

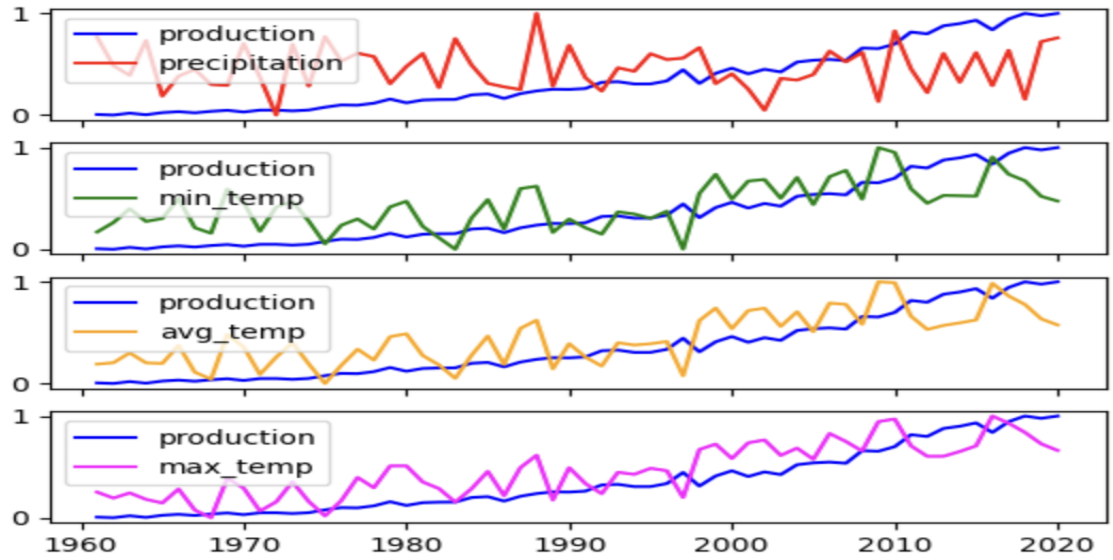
5. Scaling and Correlation:

- We have used the MinMaxScaler normalization technique and scaled the given data to fit in the range of -1 to 1.
- For the scaled data we observed the below patterns.



- We observed that with the increase the usage of fertilizers production increases over the years.
- We observed that the production has been increased over the years with increase in the emissions of C02, CH4, N2O with slight variations.



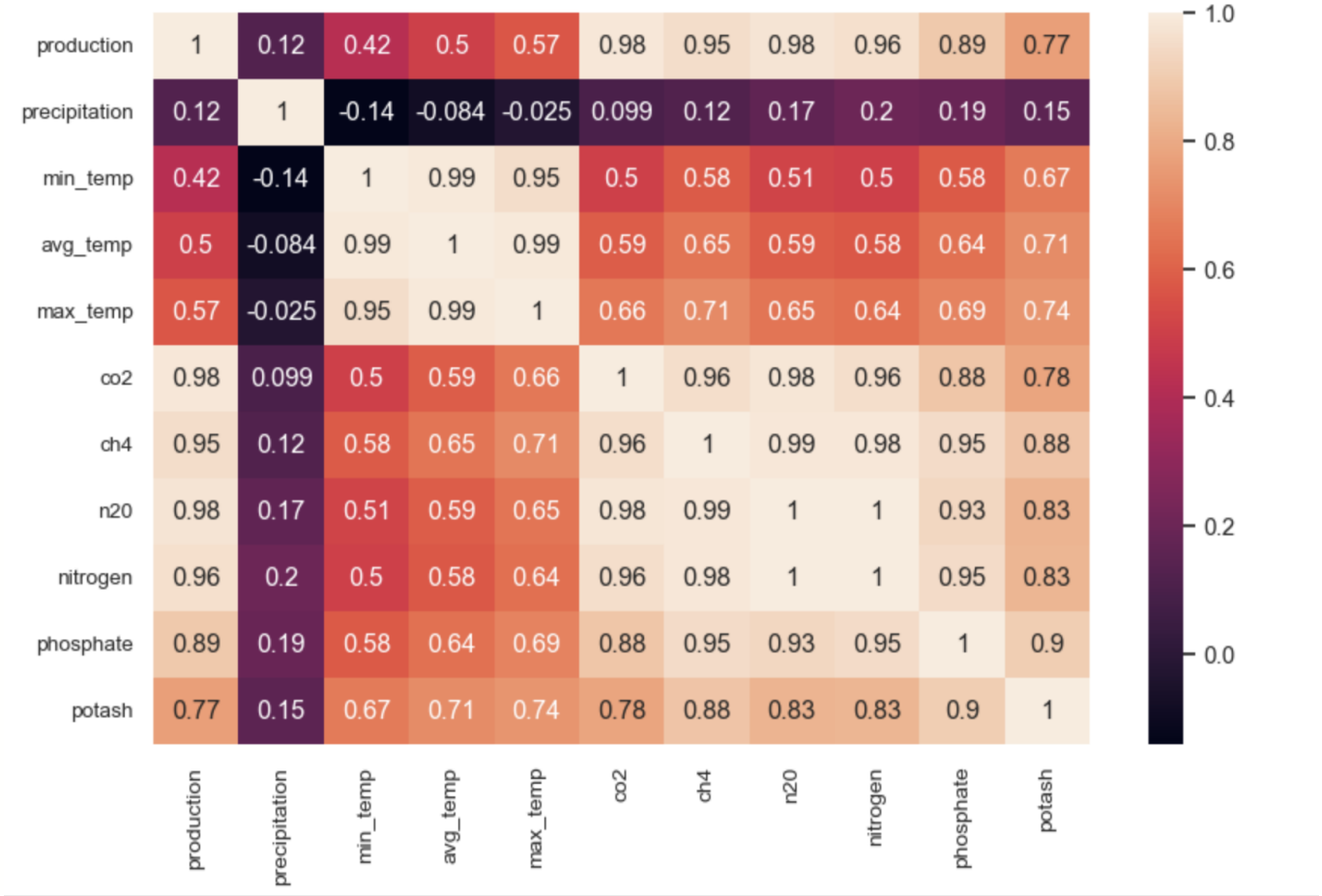


- We observed that the production is positively correlated with precipitation and temperature.

6. Analysis of final merged data:

- We have merged the three scaled datasets into a final data frame for our final analysis.

7. Pattern Evaluation:



- We have used the correlation values with heatmap for different features for identifying the relationship between production vs features.

8. Machine learning model:

- We have used the VAR model to forecast production of potatoes for the next six months.
- We have made our time series data stationary in order to apply VAR model.

- We have plotted graphs between year vs features on stationary data.
- VAR requires stationarity of the series which means the mean to the series do not change over time (we can find this out from the plot drawn next to Augmented Dickey-Fuller Test).

So, it will fit the VAR model on the training set and then use the fitted model to forecast the next year observation. These forecasts will be compared against the actual present in test data. I have taken the maximum lag (1) to identify the required lags for the VAR model.

- We have made transformed data to fit the VAR model and below is the summary of the regression results.

```
  Summary of Regression Results
==================================
Model:                         VAR
Method:                        OLS
Date:           Sun, 26, Mar, 2023
Time:                      00:34:36
--------------------------------------------------------------------
No. of Equations:         11.0000    BIC:                   -73.1314
Nobs:                     26.0000    HQIC:                  -77.6794
Log likelihood:           759.927    FPE:                6.69764e-35
AIC:                     -79.5187    Det(Omega_mle):     1.03041e-36
--------------------------------------------------------------------
```

9. Residual Plot:

- Residual plot looks normal with constant mean throughout apart from some large fluctuation during 2006 and 2010 etc.

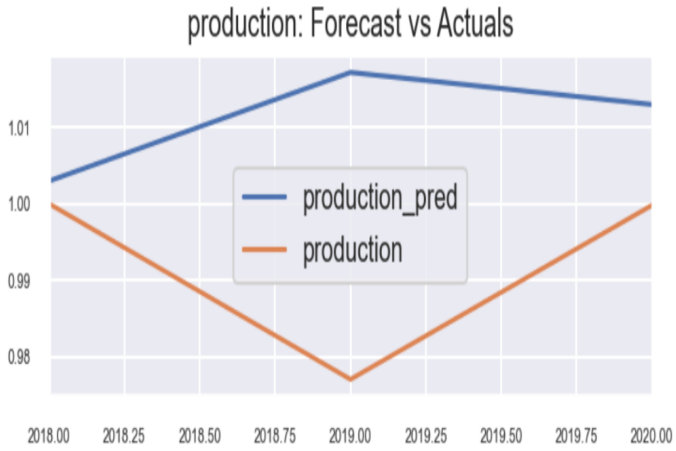
10. Prediction and Evaluation:

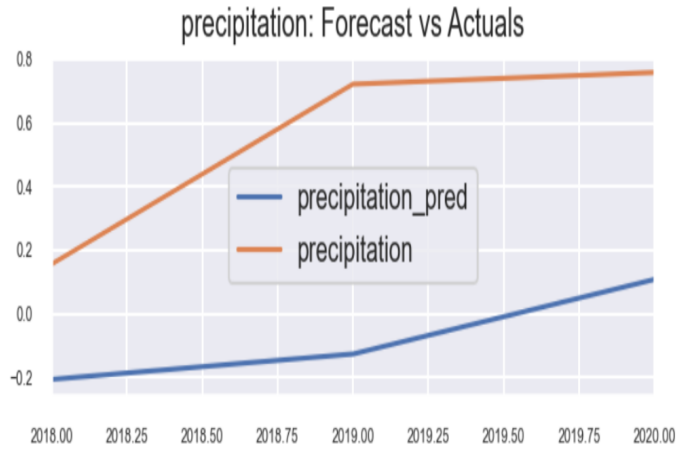
- We have combined the production data and real data for evaluating the final prediction or forecast.

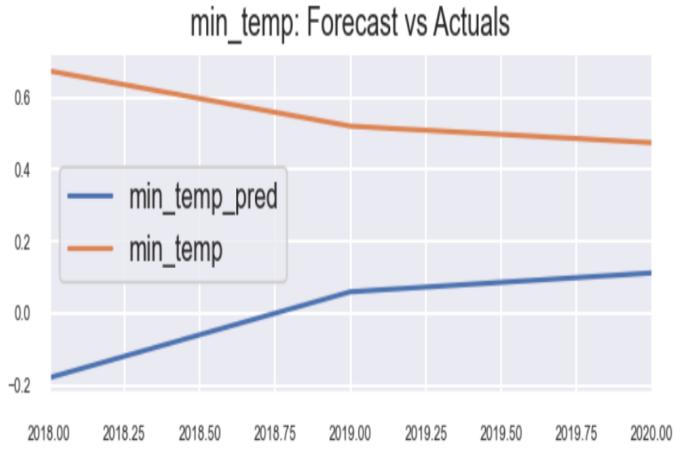
- For evaluating the final forecast we have used a set of metrics such as MAPE, ME, MAE, MPE and RMSE and calculated the results.

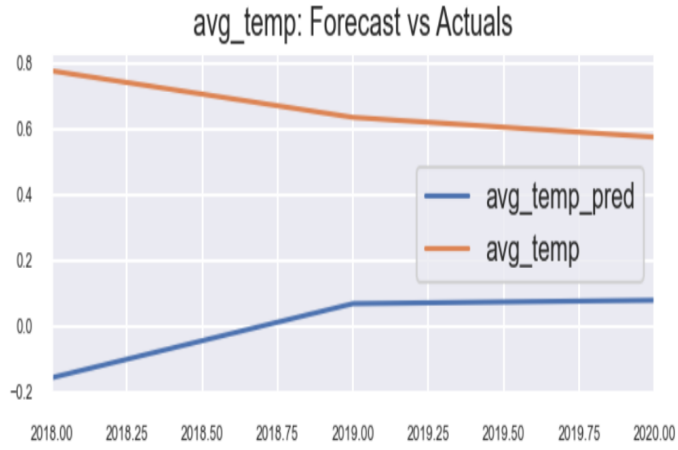
```
Bias: -0.016667
Mean absolute error: 0.01666666666666668
Mean squared error: 0.0005666666666666677
Root mean squared error: 0.023804761428476l9
```

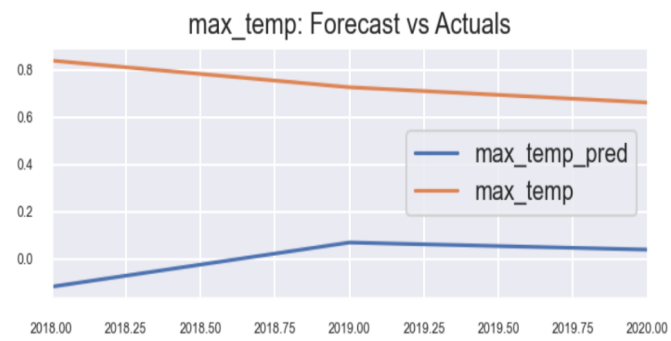
- We have plotted the graphs for different features shown below, where the relationship between the forecast data vs Actual data shown.

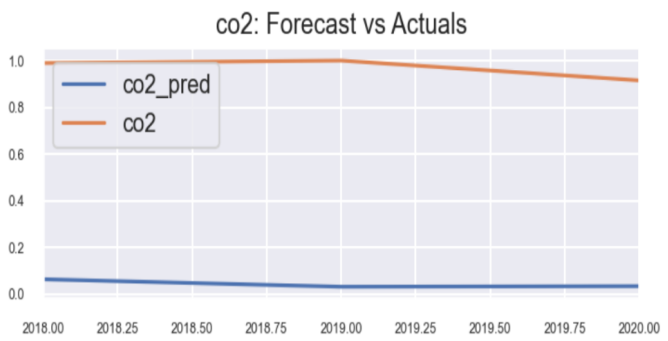


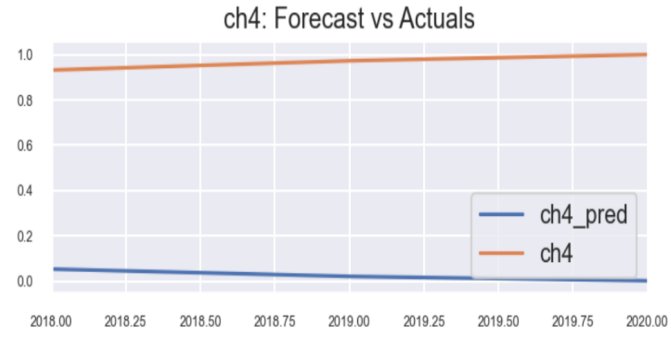


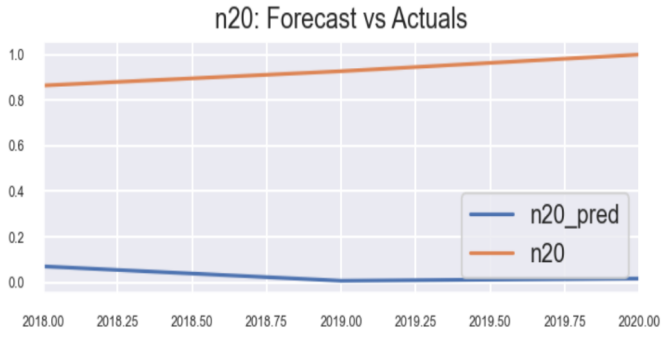


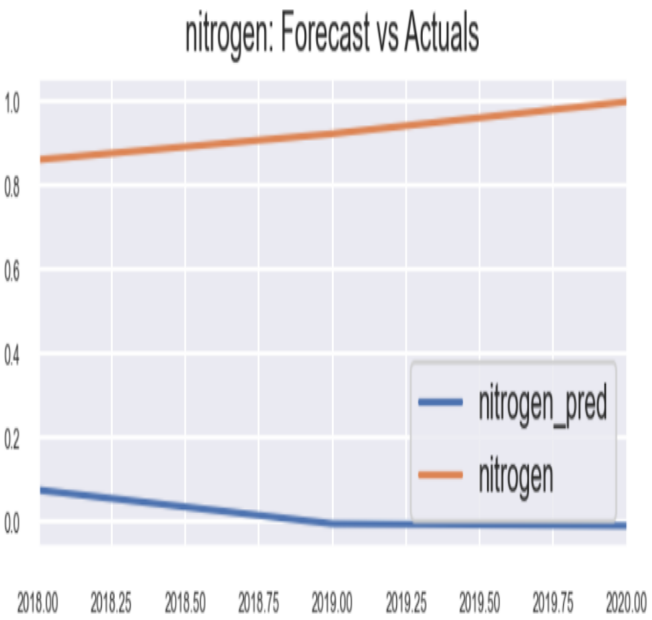


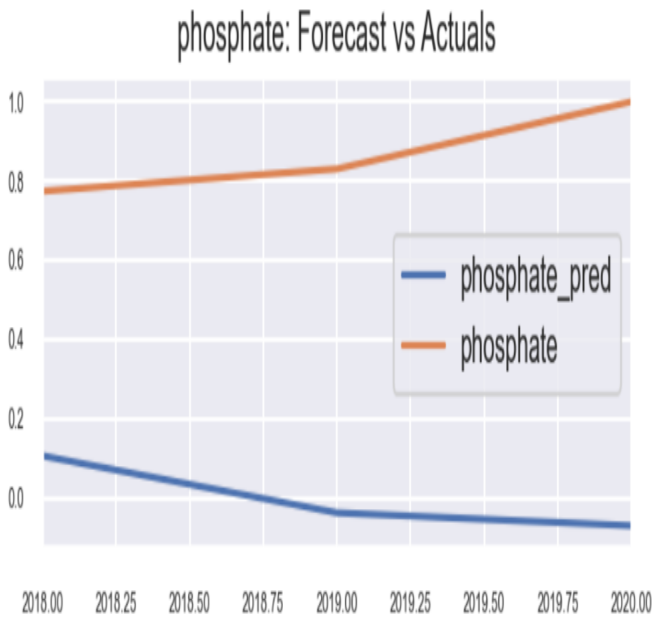

max_temp: Forecast vs Actuals
max_temp_pred
max_temp


co2: Forecast vs Actuals
co2_pred
co2


ch4: Forecast vs Actuals
ch4_pred
ch4


n20: Forecast vs Actuals
n20_pred
n20


nitrogen: Forecast vs Actuals
nitrogen_pred
nitrogen


phosphate: Forecast vs Actuals
phosphate_pred
phosphate