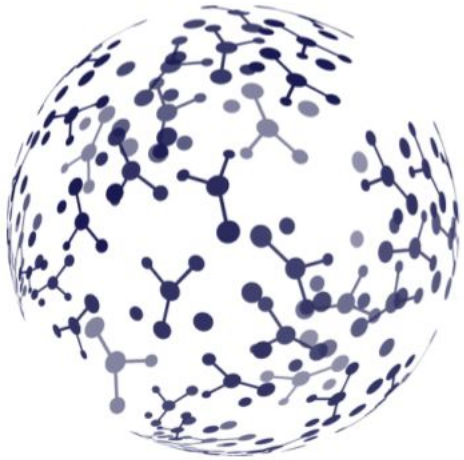


DATASCIENCEGO HACKATHON

TEAM 2

ORACLE Cloud Infrastructure | DATASCIENCEGO HACKATHON

1. Challenge Interpretation

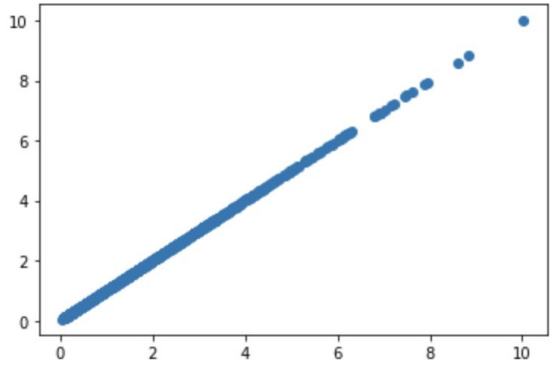
- Exploratory Analysis
- Initial Visualization
- Experimented with Linear Regression and XGBoost Regression Models

2. Solution

Examined data for consistency and missing values in key fields (Ca, Mg, etc)
Replaced missing missing values with average per site
Correlation analysis
Noted that predicted column was formula of two measurements

- TOTAL_NO3 = [tno3]+0.9841*[nhno3]
- Linear regression had best sum of squared errors

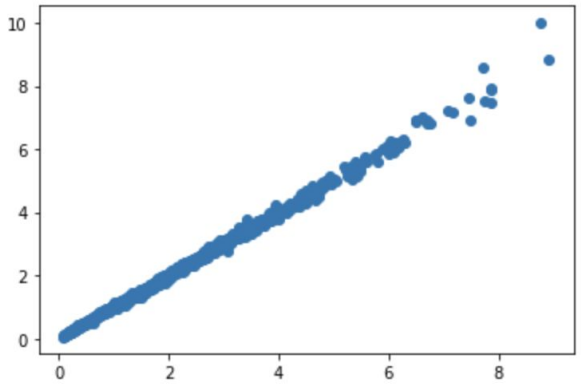
Linear Regression Model




```
print("R2=", lr.score(x_train, y_train))
print("sum of squared error=", mean_squared_error(y_predict, y_test) * len(y_test))
```

```
R2= 0.9999999990066074
sum of squared error= 2.8655499274745247e-06
```

XGBoost




```
print("xgb R2=", xgb.score(x_train, y_train))
print("xgb sum of squared error=", mean_squared_erro
```

```
xgb R2= 0.9987012809881468
xgb sum of squared error= 8.29997789716705
```

3. Recommendations

The United States Environmental Protection Agency (EPA) hired your team to help them understand air quality in protected ecosystems sensitive to pollution.

Use linear model for prediction
Troubleshoot missing values from stations
More exploration on time series and anomaly detection