Report on CO2 Concentration Prediction Using ARIMA

1. Introduction 1.

The aim of this project is to analyze CO2 sensor data, determine the stationarity of the data, and develop an ARIMA-based time-series forecasting model. This report presents the results of the analysis, discusses the model's performance, and reflects on the challenges faced during the process.

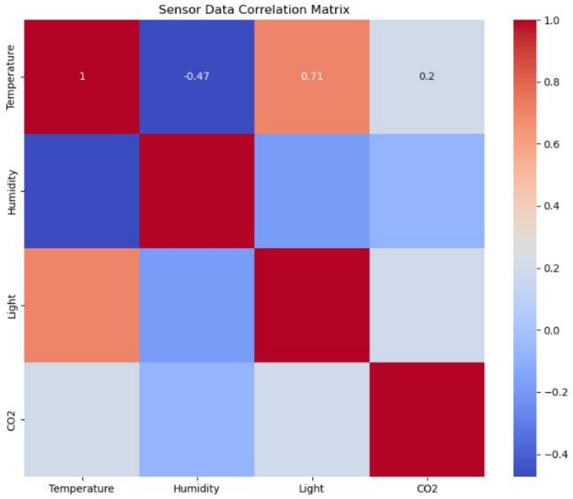
2. Dataset Analysis

The dataset consists of sensor readings for temperature, humidity, light intensity, and CO2 concentration. A summary of the dataset is as follows:

- **Observation Period:** Time-series data spanning several days.
- **Key Metrics (CO2):**
  - Mean: 753.22
  - Standard Deviation: 297.09
  - Range: 484.67 to 2076.50

Correlation Analysis

The correlation matrix reveals:



Correlation Analysis:

```
             Temperature  Humidity     Light       CO2
Temperature     1.000000 -0.472921  0.705538  0.199646
Humidity       -0.472921  1.000000 -0.187477 -0.079224
Light           0.705538 -0.187477  1.000000  0.190213
CO2             0.199646 -0.079224  0.190213  1.000000
```

The correlation analysis reveals several key relationships among the variables in the dataset. Temperature shows a weak positive correlation (0.20) with CO2, indicating that as temperature increases, CO2 levels tend to rise slightly, although the relationship is not particularly strong. Similarly, Light also demonstrates a weak positive correlation (0.19) with CO2, suggesting that higher light intensity might be associated with slightly elevated CO2 levels. On the other hand, Humidity exhibits a very weak negative correlation (-0.08) with CO2, implying that changes in humidity have minimal influence on CO2 concentrations. These findings suggest that while Temperature and Light might provide some predictive power for CO2 levels, their individual contributions are limited. Additionally, the strong positive correlation (0.71) between Temperature and Light highlights a significant interdependence, likely driven by environmental factors such as heat generated by light sources. The moderate negative correlation (-0.47) between Temperature and Humidity further reflects their inverse relationship, consistent with physical phenomena where rising temperatures often lead to lower relative humidity. Overall, while these variables offer some insights into CO2 behavior, their weak individual correlations with CO2 suggest that predictive models may benefit from incorporating additional features or exploring more complex relationships.

---

3. Data Stationarity

Using the Augmented Dickey-Fuller (ADF) test, we tested the stationarity of the CO2 data. The ADF statistic confirmed that the data is non-stationary in its raw form. Differencing (d=1) was applied to achieve stationarity.

---

4. ARIMA Model Selection

The auto_arima function was used to determine the best-fit ARIMA parameters:

- Best Model: **ARIMA(2,1,2)** with the following:
  - **p** (autoregressive order): 2

- **d** (degree of differencing): 1
- **q** (moving average order): 2

- AIC Score: **59231.77**
- Total model fit time: ~38 seconds.

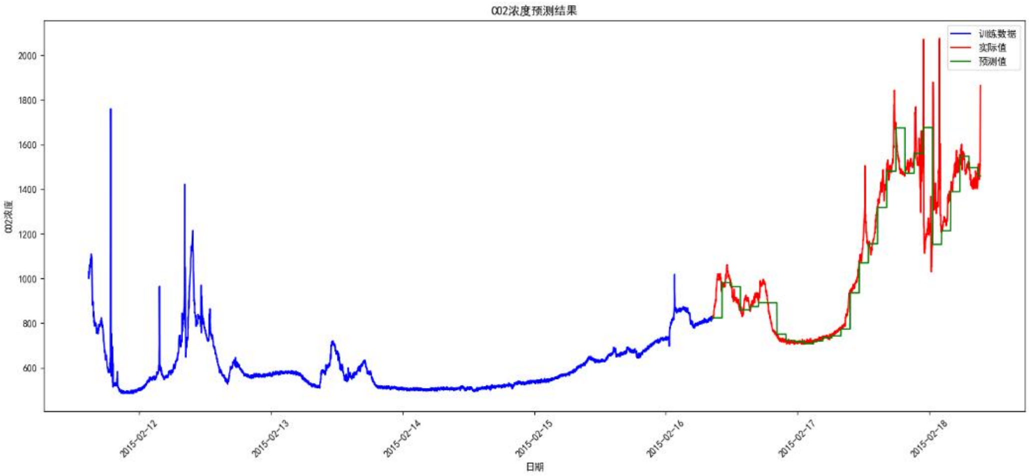
5. Forecasting and Results

The CO2 data was split into 70% training and 30% test sets. Predictions were made using a stepwise approach, updating the model every fixed number of steps.

Performance Metrics:

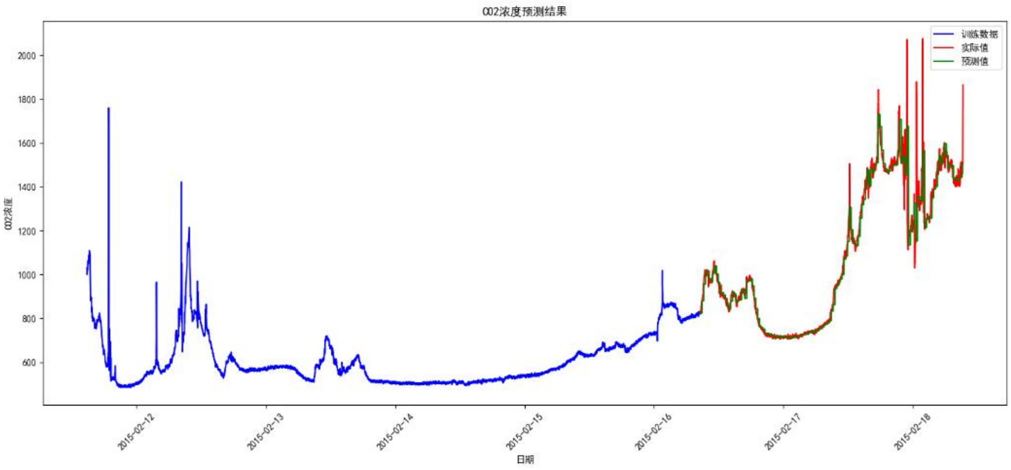
Refit Every 100 Steps (i % 100):

MAE:75.29



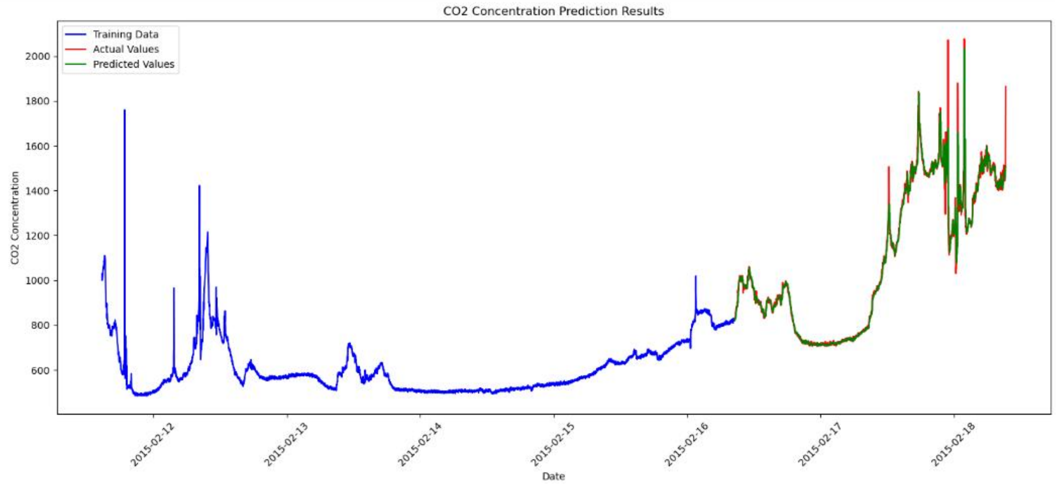
Refit Every 20 Steps (i % 20):

- MAE: 31.73
- RMSE: 73.28



Refit Every 5 Steps (i % 5):

- MAE: 16.44
- RMSE: 40.37



Observations:

- Frequent refitting (i.e., every 5 steps) resulted in the lowest prediction error (MAE: 16.44, RMSE: 40.37).
- While computationally expensive, frequent updates significantly improve accuracy by capturing short-term variations in the data.
- Predictions struggle with outliers, highlighting the need for better anomaly handling.

6. Challenges and Reflections

Handling Non-Stationarity:

- Achieving stationarity was straightforward using differencing (d=1). However, this transformation may have led to the loss of some long-term trends.

Outlier Sensitivity:

- Predictions were less accurate for outliers. In real-world applications like industrial or financial forecasting, such anomalies are common and need robust handling mechanisms.

Trade-Off Between Accuracy and Computation:

- Frequent refitting (e.g., every 5 steps) improved accuracy but increased computation time. Optimizing the refit frequency based on prediction errors could balance this trade-off.

Model Generalizability:

- The ARIMA model performed well for linear and stationary data but may not generalize to non-linear relationships. Hybrid models (e.g., ARIMA + LSTM) can be explored.

---

7. Recommendations

Incorporate Anomaly Detection

- Use statistical or machine-learning methods to identify and preprocess outliers.

Hybrid Modeling:

- Combine ARIMA with neural networks to handle both linear and non-linear patterns in the data.

Dynamic Refit Intervals:

- Implement dynamic refitting, where the model updates only when prediction errors exceed a defined threshold.

Feature Engineering:

- Integrate additional external variables, such as weather or industrial activity, to improve the model's explanatory power.

The ARIMA(2,1,2) model effectively forecasted CO2 concentration, achieving a mean absolute error of 16.44 with frequent refitting. While the model demonstrates the potential of ARIMA for sensor data analysis, challenges like outlier sensitivity and computational efficiency must be addressed for industrial-scale applications.

APPIDEX: CO2 Concentration Forecasting Experiment (5 Steps)

Dataset Basic Information

```
                     Temperature   Humidity        Light          CO2
date
2015-02-11 14:48:00      21.7600  31.133333  437.333333  1029.666667
2015-02-11 14:49:00      21.7900  31.000000  437.333333  1000.000000
2015-02-11 14:50:00      21.7675  31.122500  434.000000  1003.750000
2015-02-11 14:51:00      21.7675  31.122500  439.000000  1009.500000
```

```
2015-02-11 14:51:00        21.7900   31.133333   437.333333   1005.666667
```

Descriptive Statistics of the Dataset:

```
        Temperature     Humidity        Light          CO2
count   9752.000000  9752.000000  9752.000000  9752.000000
mean      21.001768    29.891910   123.067930   753.224832
std        1.020693     3.952844   208.221275   297.096114
min       19.500000    21.865000     0.000000   484.666667
25%       20.290000    26.642083     0.000000   542.312500
50%       20.790000    30.200000     0.000000   639.000000
75%       21.533333    32.700000   208.250000   831.125000
max       24.390000    39.500000  1581.000000  2076.500000
```

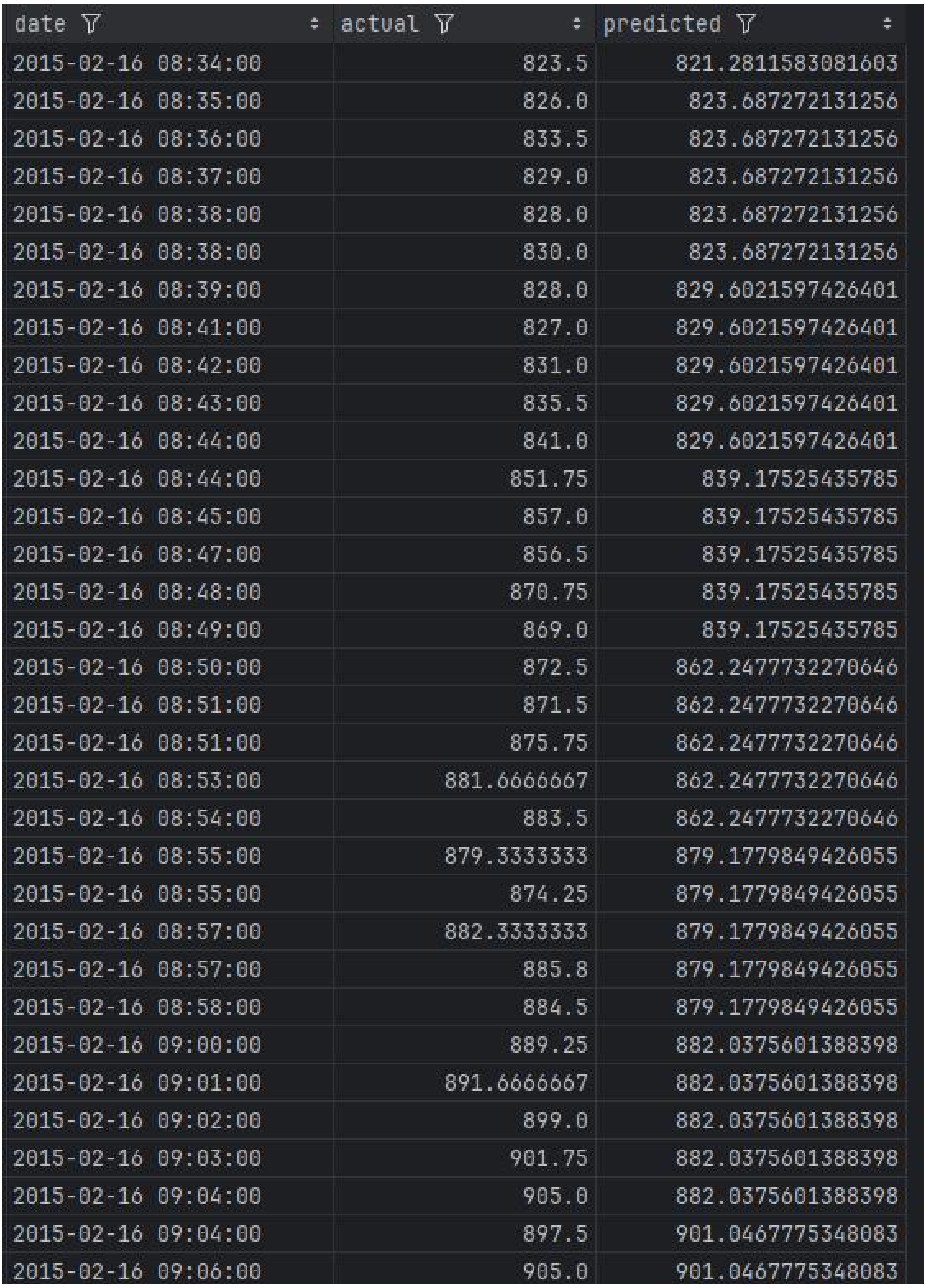
Searching for the best ARIMA parameters...

```
Performing stepwise search to minimize aic
 ARIMA(0,1,0)(0,0,0)[0] intercept   : AIC=60029.820, Time=0.09 sec
 ARIMA(1,1,0)(0,0,0)[0] intercept   : AIC=59997.355, Time=0.20 sec
 ARIMA(0,1,1)(0,0,0)[0] intercept   : AIC=59979.849, Time=0.62 sec
 ARIMA(0,1,0)(0,0,0)[0]             : AIC=60027.837, Time=0.04 sec
 ARIMA(1,1,1)(0,0,0)[0] intercept   : AIC=59656.767, Time=1.10 sec
 ARIMA(2,1,1)(0,0,0)[0] intercept   : AIC=59244.234, Time=1.56 sec
 ARIMA(2,1,0)(0,0,0)[0] intercept   : AIC=59748.541, Time=0.28 sec
 ARIMA(3,1,1)(0,0,0)[0] intercept   : AIC=59234.899, Time=2.29 sec
 ARIMA(3,1,0)(0,0,0)[0] intercept   : AIC=59531.164, Time=0.34 sec
 ARIMA(4,1,1)(0,0,0)[0] intercept   : AIC=59235.967, Time=3.16 sec
 ARIMA(3,1,2)(0,0,0)[0] intercept   : AIC=59234.998, Time=5.93 sec
 ARIMA(2,1,2)(0,0,0)[0] intercept   : AIC=59233.687, Time=2.47 sec
 ARIMA(1,1,2)(0,0,0)[0] intercept   : AIC=59312.969, Time=2.03 sec
 ARIMA(2,1,3)(0,0,0)[0] intercept   : AIC=59235.078, Time=2.29 sec
 ARIMA(1,1,3)(0,0,0)[0] intercept   : AIC=59262.631, Time=1.99 sec
 ARIMA(3,1,3)(0,0,0)[0] intercept   : AIC=59237.204, Time=2.38 sec
 ARIMA(2,1,2)(0,0,0)[0]             : AIC=59231.767, Time=0.94 sec
 ARIMA(1,1,2)(0,0,0)[0]             : AIC=59311.070, Time=0.83 sec
 ARIMA(2,1,1)(0,0,0)[0]             : AIC=59242.323, Time=0.72 sec
 ARIMA(3,1,2)(0,0,0)[0]             : AIC=59246.320, Time=0.99 sec
 ARIMA(2,1,3)(0,0,0)[0]             : AIC=59233.157, Time=1.21 sec
 ARIMA(1,1,1)(0,0,0)[0]             : AIC=59654.883, Time=0.45 sec
 ARIMA(1,1,3)(0,0,0)[0]             : AIC=59260.736, Time=0.56 sec
 ARIMA(3,1,1)(0,0,0)[0]             : AIC=59232.982, Time=0.91 sec
 ARIMA(3,1,3)(0,0,0)[0]             : AIC=59235.282, Time=1.09 sec

Best model:  ARIMA(2,1,2)(0,0,0)[0]
Total fit time: 34.481 seconds
```

Best model parameters: (2, 1, 2)

Predicted data:



date	actual	predicted
2015-02-16 08:34:00	823.5	821.2811583081603
2015-02-16 08:35:00	826.0	823.687272131256
2015-02-16 08:36:00	833.5	823.687272131256
2015-02-16 08:37:00	829.0	823.687272131256
2015-02-16 08:38:00	828.0	823.687272131256
2015-02-16 08:38:00	830.0	823.687272131256
2015-02-16 08:39:00	828.0	829.6021597426401
2015-02-16 08:41:00	827.0	829.6021597426401
2015-02-16 08:42:00	831.0	829.6021597426401
2015-02-16 08:43:00	835.5	829.6021597426401
2015-02-16 08:44:00	841.0	829.6021597426401
2015-02-16 08:44:00	851.75	839.17525435785
2015-02-16 08:45:00	857.0	839.17525435785
2015-02-16 08:47:00	856.5	839.17525435785
2015-02-16 08:48:00	870.75	839.17525435785
2015-02-16 08:49:00	869.0	839.17525435785
2015-02-16 08:50:00	872.5	862.2477732270646
2015-02-16 08:51:00	871.5	862.2477732270646
2015-02-16 08:51:00	875.75	862.2477732270646
2015-02-16 08:53:00	881.6666667	862.2477732270646
2015-02-16 08:54:00	883.5	862.2477732270646
2015-02-16 08:55:00	879.3333333	879.1779849426055
2015-02-16 08:55:00	874.25	879.1779849426055
2015-02-16 08:57:00	882.3333333	879.1779849426055
2015-02-16 08:57:00	885.8	879.1779849426055
2015-02-16 08:58:00	884.5	879.1779849426055
2015-02-16 09:00:00	889.25	882.0375601388398
2015-02-16 09:01:00	891.6666667	882.0375601388398
2015-02-16 09:02:00	899.0	882.0375601388398
2015-02-16 09:03:00	901.75	882.0375601388398
2015-02-16 09:04:00	905.0	882.0375601388398
2015-02-16 09:04:00	897.5	901.0467775348083
2015-02-16 09:06:00	905.0	901.0467775348083