Air Quality Sensor Analysis Using Unsupervised Learning Anomaly Detection

Shan Hakani
shakani7@gatech.edu

1 BACKGROUND

1.1 Initial Case Study and Data Source

In a study conducted by the Italian National Agency for New Technologies, Energy and Sustainable Economic Development (ENEA), a multi-sensor device developed by Pirelli Labs was used to gather air quality data in a highly polluted and urbanized Italian city over a 13-month period (De Vito et al., 2008). The initial study collected this data with the sensor device to help predict the concentration of benzene in the environment. The multi-sensor device was, at the time, a novel tool designed specifically for highly polluted environments.

According to Clarity, a leading provider of air sensing technology, long-term air quality sensor degradation can be attributed to weather conditions, such as temperature and humidity, as well as long-term exposure to corrosive substances, dust, and debris.

We will perform this analysis using the same data set from the case study (sourced from UC Irvine Machine Learning Repository) to detect anomalies with unsupervised learning and determine periods of high frequency anomalies that attribute to long-term sensor degradation. The data consists of time series data of hourly response averages from March 2004 to February 2005.

The data set consists of 9358 records and 15 features. Of those features, there are five different types of gases that were measure and recorded over the duration of the study: CO(GT), NMHC(GT), C6H6(GT), NOx(GT), and NO2(GT). Additionally, the case study collected data on temperature (T) and relative and absolute humidity (RH, AH).

2 METHODOLOGY

In this project, there are four main stages: data cleaning and preprocessing, model creation, feature importance, and plot interpretation. Given that the dataset does not have clear labels for when there are spikes in harmful gas concentration, we will be leveraging the Isolation Forest algorithm to determine if there are any anomalies recorded. Then, we will use SHAP (SHapley Additive exPlanations) to find which features contribute most to the detected anomaly. Lastly, we will generate various plots to verify and interpret the model results (i.e. anomaly score distribution, time series plots, and comparative scatter plots of most influential features).

2.1 Data Cleaning and Preprocessing

Upon initial investigation of the data set, there were two main steps needed to be done to clean the data set: modify the date and time fields to a more useable format and handle missing data.

The date and time fields were separated into two separate fields of type string, so a new column "Datetime" was created to use in time series plots by concatenating the two fields and converting it to a formatted datetime object.

The data set uses the value -200 to indicate missing data. The missing value counts in each field are as follows:

Field Name	**Missing Value Count**	**Field Name**	**Missing Value Count**
CO(GT)	*1683*	*NOx(GT)*	*1639*
PT08.S1(CO)	*366*	*PT08.S3(NOx)*	*366*
NMHC(GT)	8443	NO2(GT)	1642
C6H6(GT)	366	PT08.S4(NO2)	366
PT08.S2(NMHC)	366	PT08.S5(O3)	366
T	366	AH	366
RH	366	Datetime	0

Given that there are 9358 total records, the field “NMHC(GT)” was dropped due to having majority missing data. All other values were replaced using a simple median imputer.

Lastly, the data was standardized and scaled prior to running the model to ensure all features are treated equally.

2.2 Isolation Forest Modeling

In this analysis, there is no target label to predict, and we are attempting to draw insights from the features to determine spikes in harmful gases that can attribute to both poor air quality and long-term sensor degradation. Given that this requires an unsupervised learning model to identify anomalous data points, we can leverage the Isolation Forest algorithm.

The algorithm will build out decision trees and randomly select a feature to split on based on an arbitrary threshold value. It will continue to split until an anomaly is detected by finding outliers in unique combinations of features.

Since we are analyzing data in a highly polluted environment (and generally for air quality data), we should not expect a high number of anomaly occurrences unless there are extreme and unexpected air quality conditions. Given this, the contamination hyperparameter, which specifies the expected percentage of anomalies, was set to a value of 0.01 to only target the most rare cases in the data set.

2.3 Feature Importance

Using Isolation Forest, an unsupervised learning model, prevents us from using traditional feature importance extraction methods; however, SHAP (SHapley Additive exPlanations) is one way we can learn more about how our features contribute to the anomaly detection results.

SHAP is a method that uses Shapley values from cooperative game theory to provide insight as to why the algorithm flagged certain data points as anomalies. We can use the SHAP summary plot to find which features attribute the most to the detected anomaly.

2.4 Plot Interpretation

Using the results from the SHAP summary plot, we can identify the three most influential features and generate time series plots to see if the anomalies detected line up with the gas levels on the same day. This will help both in validating model functionality and uncovering insights such as any time periods of high frequency anomalies that can attribute to sensor degradation.

Additionally, we can generate comparative scatter plots of each combination of the three most influential gasses with the anomalies to further verify performance by comparing clusters.

3 RESULTS

3.1 Isolation Score Distribution

After running the Isolation Forest algorithm and computing the isolation scores, we see from the distribution of the scores on the histogram that the model is working as expected. In Figure 1, the scores skewed to the right, and we see a lower number of anomalies that scored below zero, indicating that the majority of our data points are considered "normal" conditions.

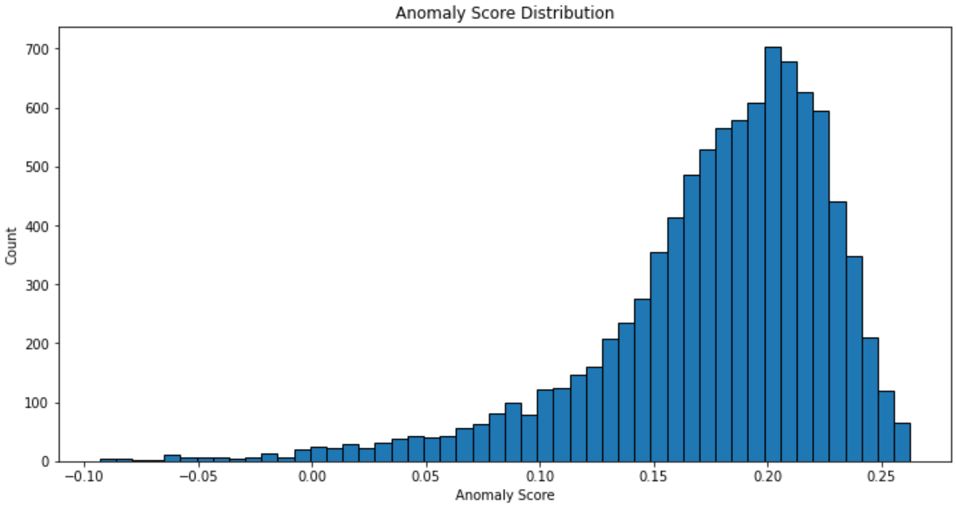

Figure 1 **-** Anomaly score distribution from Isolation Forest algorithm

3.2 SHAP and Feature Importance

Using SHAP, we can use the summary plot to determine which features contribute most to the anomalies detected. From Figure 2, we see that CO(GT), NOx(GT), and NO2(GT) have the most influence.

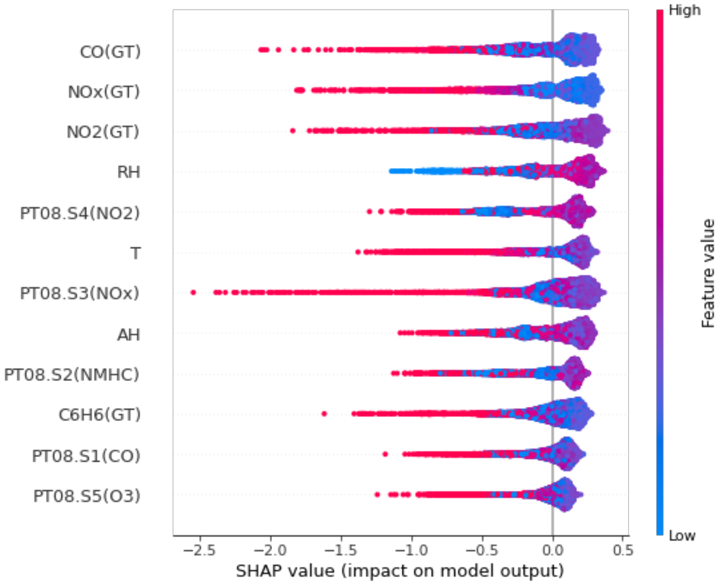

Figure 2 - Anomaly score distribution from Isolation Forest algorithm

3.3 Time Series Visualizations

Knowing that CO(GT), NOx(GT), and NO2(GT) had the most influence over the anomalies detected, we can plot the gas levels over time for each of the gases and mark the time at which the anomalies occurred over it to see if there are spikes in the gas levels as well.

Examining Figures 3, 4, and 5, we see that, for the most part, the spikes in each of the three gasses line up with the detected anomalies. The lower volume anomalies that were detected are irrelevant as we are more interested in the high-volume cases that will attribute the most to sensor degradation. The most interesting insight that we gain from these three charts is that majority of these anomalies occur from October 2004 to January 2005. This is likely this is due to geological events at the time such as the effusive eruption of Mount Etna from September 2004 to March 2005 (Aiuppa et al., 2006), which can contribute to increased levels of CO and NO2 (Romanias et al., 2020).

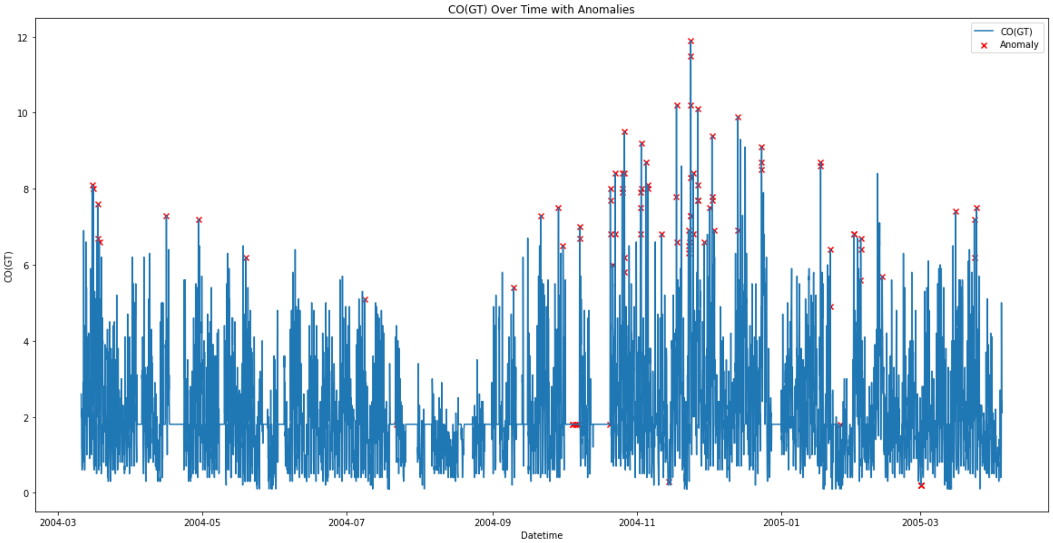

Figure 3 – CO(GT) plotted over time with Anomalies marked over

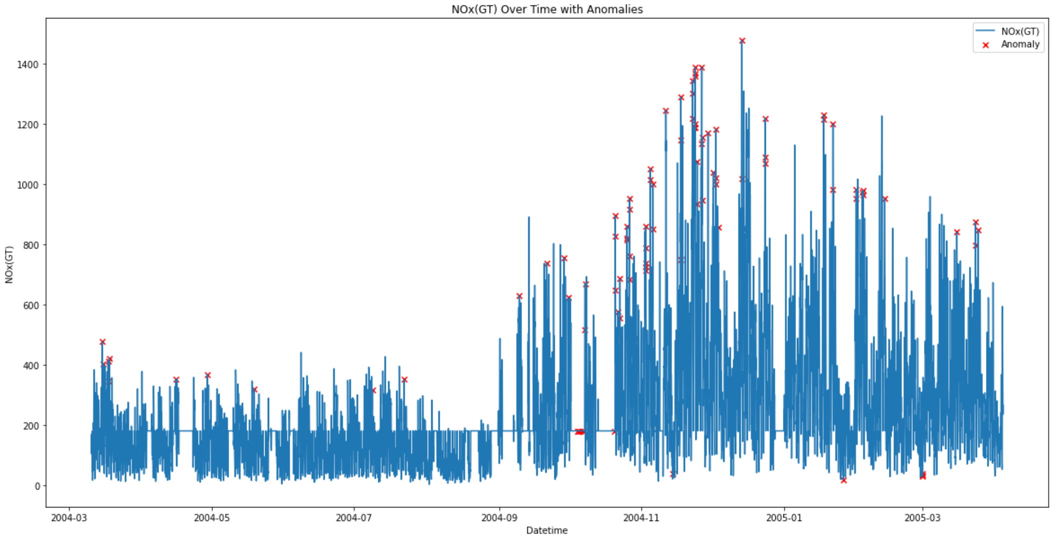

Figure 4 – NOx(GT) plotted over time with Anomalies marked over

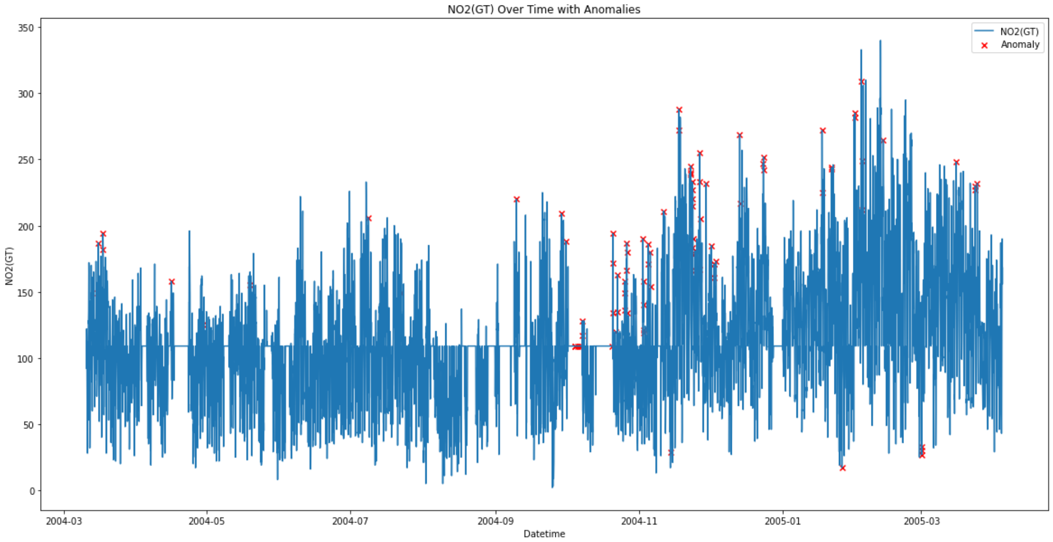

Figure 5 – NO2(GT) plotted over time with Anomalies marked over

3.4 Comparative Scatter Plots

To further validate the results from the isolation forest model, we can compare CO(GT), NOx(GT), and NO2(GT) against each other to see where the anomalies appear relative to what were identified as "normal" data points.

Examining charts 6, 7, and 8, we see some good separation between the normal observations (red) and anomalous observations (blue). This further validates that the Isolation Forest algorithm was able to identify the anomalies effectively since majority of our blue points are separated from the cluster of red points.

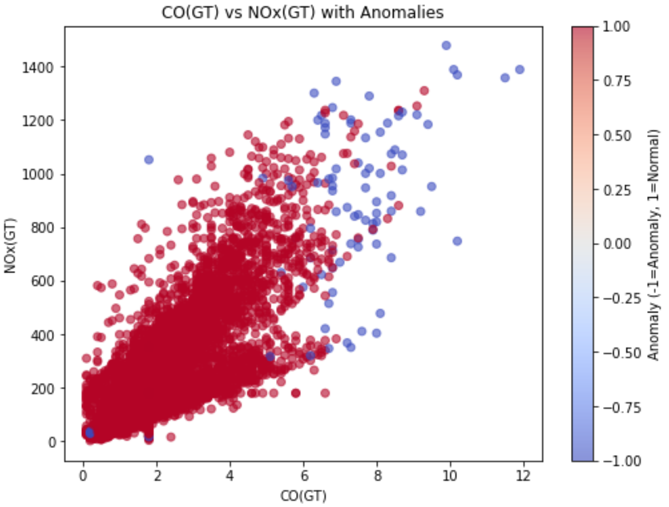

Figure 6 – CO(GT) vs NOx(GT), with Anomalies as blue points and normal observations as red points

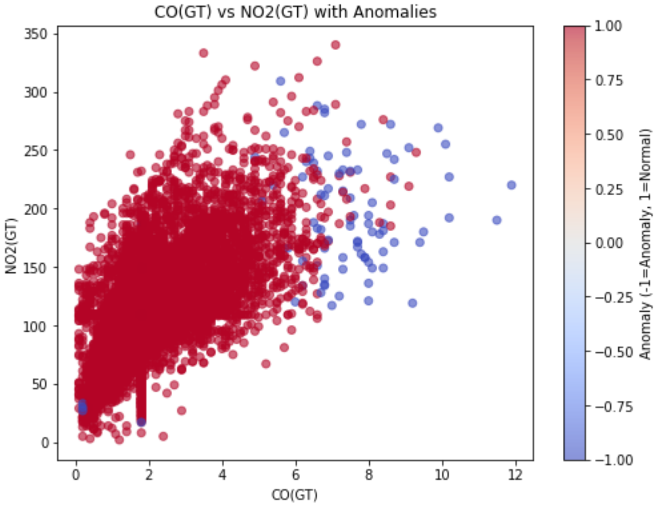

Figure 7 – CO(GT) vs NO2(GT), with Anomalies as blue points and normal observations as red points

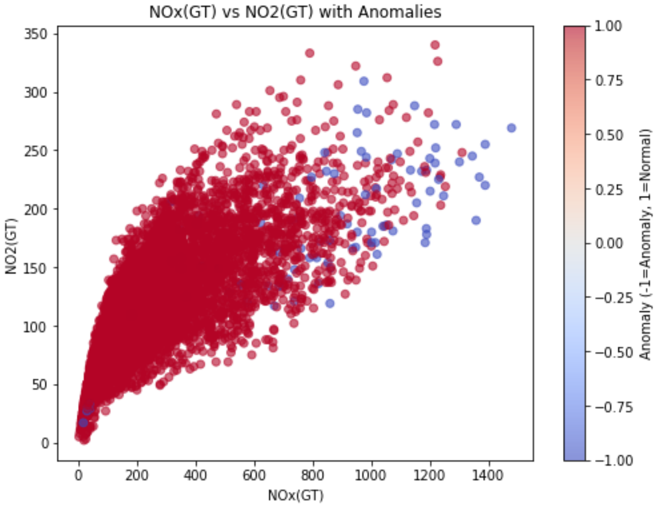

Figure 8 – NOx(GT) vs NO2(GT), with Anomalies as blue points and normal observations as red points

4 DISCUSSION AND FUTURE WORK

Overall, the algorithm worked as intended and we were able to identify period of high frequency of anomalies. In practice, this analysis can be useful in providing early warning signals to environmental analyst by flagging abnormal behavior to indicate that the air quality sensors are likely to experience degraded quality due to increased exposure to abnormal environmental conditions. Domain experts can help by setting degradation thresholds to determine when a sensor will actually need to be replaced (i.e. duration of increased anomalous activity).

For future improvements, it would be helpful to integrate weather conditions into this analysis as well. Given the time frame of the data set, it's difficult to tell if the sensor was already degraded early in the data collection due to other weather conditions. It would also be interesting to see how factors such as precipitation and wind speeds, for example, influence the frequency of anomalies detected. Additionally, having a longer period of data to analyze would greatly improve the interpretability and insights gathered from this analysis.

5 REFERENCES

1. De Vito, S., Massera, E., Piga, M., Martinotto, L., & Di Francia, G. (2008). On field calibration of an electronic nose for benzene estimation in an urban pollution monitoring scenario. Sensors and Actuators B: Chemical, 129(2), 750–757. https://doi.org/10.1016/j.snb.2007.09.060
2. How to detect when an air quality sensor's detection limit has shifted | Clarity. (2025). Clarity.io. https://www.clarity.io/blog/how-to-detect-when-an-air-quality-sensors-detection-limit-has-shifted-and-needs-recalibration-or-servicing
3. Cortes, D. (n.d.). Overview. R-Packages. https://cran.r-project.org/web/packages/isotree/vignettes/An_Introduction_to_Isolation_Forests.html
4. Lundberg, S. (2018). An introduction to explainable AI with Shapley values. https://shap.readthedocs.io/en/latest/example_notebooks/overviews/An%20introduction%20to%20explainable%20AI%20with%20Shapley%20values.html
5. Aiuppa, A., Federico, C., Giudice, G., Gurrieri, S., Liuzzo, M., Shinohara, H., Favara, R., & Valenza, M. (2006). Rates of carbon dioxide plume degassing from Mount Etna volcano. Journal of Geophysical Research, 111(B9). https://doi.org/10.1029/2006jb004307
6. Romanias, M. N., Ren, Y., Benoit Grosselin, Véronique Daële, Abdelwahid Mellouki, Pavla Dagsson-Waldhauserova, & Thevenet, F. (2020). Reactive uptake of NO2 on volcanic particles: A possible source of HONO in the atmosphere. Journal of Environmental Sciences, 95, 155–164. https://doi.org/10.1016/j.jes.2020.03.042