Grace Berg
Institute at Brown for Environment and Society
Repository: https://github.com/gberggg/data1030_finalproject_new.git

Developing a ML calibration scheme for Breathe Providence low-cost NO_2 measurements

Introduction

Brown University's Breathe Providence project has developed a network of 25 low-cost air sensors (LCS) distributed throughout Providence, RI. The project's goal is to characterize neighborhood-scale pollution dynamics and disentangle the impacts of local versus transported sources. Each sensor measures a suite of common air pollutants, including particulate matter ($PM_{2.5}$) and several gaseous pollutants (CO, O_3, NO, and NO_2). While LCS offer advantages in cost, size, and deployment density, they face significant data quality challenges. Electrochemical gas sensors, used by Breathe Providence, are known to be affected by meteorological conditions (e.g. temperature and humidity), cross-sensitivities to pollutants other than the target gas, and measurement drift. These sensor-unique biases can be mitigated through calibration schemes, which often rely on the relationship of one or more co-located low-cost sensors to reference-grade measurements.

Previous work (**Table 1**) has investigated the use of ML methods for calibrating electrochemical LCS NO_2 measurements, including multilinear regression [Levy Zamora et al., 2022; Frederickson et al., 2022], random forest [Zimmerman et al., 2018], and support vector machine [Bigi et al., 2018]. There remains debate over the use of complex ML models versus simpler, physically interpretable ones, as complex models can underperform in field calibrations [Winter et al., 2024 (in review); Russell et al., 2022]. This project aims to develop and compare both complex and interpretable ML calibration models for Breathe Providence's NO_2 measurements, focusing on co-located calibration as a foundation for future field calibration efforts.

	Location	Method	Days	Temporal resolution (minutes)	Average concentration (ppb)	R^2	RMSE (ppb)
Winter et al. (in review)	Bay Area, CA	Multiple linear regression	365	60	6.4	0.82	2.26
Levy Zamora et al. (2022)	Baltimore, MD	Multiple linear regression	365	60	8.1	0.77	3.6
van Ratingen et al. (2021)	Netherlands	Multivariate regression	248	60	NR	0.69 - 0.84	5.32 - 10.04
Bigi et al., (2018)	Switzerland	Support Vector Machine	120	60	NR	0.74	4.6
Zimmerman et al. (2018)	Pittsburgh, PA	Random Forest	41	15	270	0.91	NR

Table 1. *Sample of recent Alphasense NO_2 low-cost calibration work.*

The target variable is the hourly reference NO_2 measurement from the East Providence site operated by the RI Department of Health, accessed via the Air Quality System API (US EPA). Features include co-located Breathe Providence sensor measurements: NO_2, NO, and O_3 concentrations (voltage, Alphasense NO2-B43F, NO-B4, and Ox-B43) and meteorological variables (temperature in °C and relative humidity in %, Adafruit BME280). LCS data, accessed via the Berkeley Environmental Air-quality & CO_2 Network API, are recorded every few seconds and aggregated to hourly averages. The

dataset spans 2023. Although time-stamped, this is not a time series problem; the data are treated as IID, with all features and the target being continuous.

Exploratory Data Analysis

EDA revealed key insights about the target variable and the interplay of different sources of bias. For example, **Figure 1** below shows the positive skew of reference NO_2. Most values are clustered between 0 and 5 ppb, with a tail reaching a maximum value of 38.8 ppb. This skew has implications for splitting the data; the split may need to be stratified in order to ensure the full target variable distribution is represented in each set.

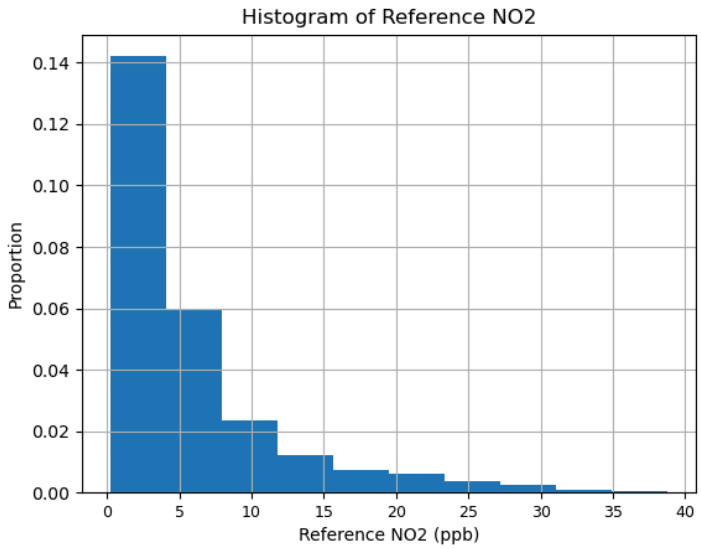

Figure 1*. Distribution of the target variable expressed as a proportion of the total.*

Changes in temperature affect both ambient NO_2 levels and LCS bias. **Figure 2a** shows NO_2 concentrations are higher and more variable at low temperatures, likely due to seasonal factors: changes in emissions (e.g., wintertime wood burning), atmospheric chemistry (e.g., higher summertime O_3 is an NO_2 sink), and meteorology (e.g., lower winter boundary layer traps pollutants). **Figure 2b** shows that the LCS has sensor-unique, temperature-driven biases introduced at > 30°C.

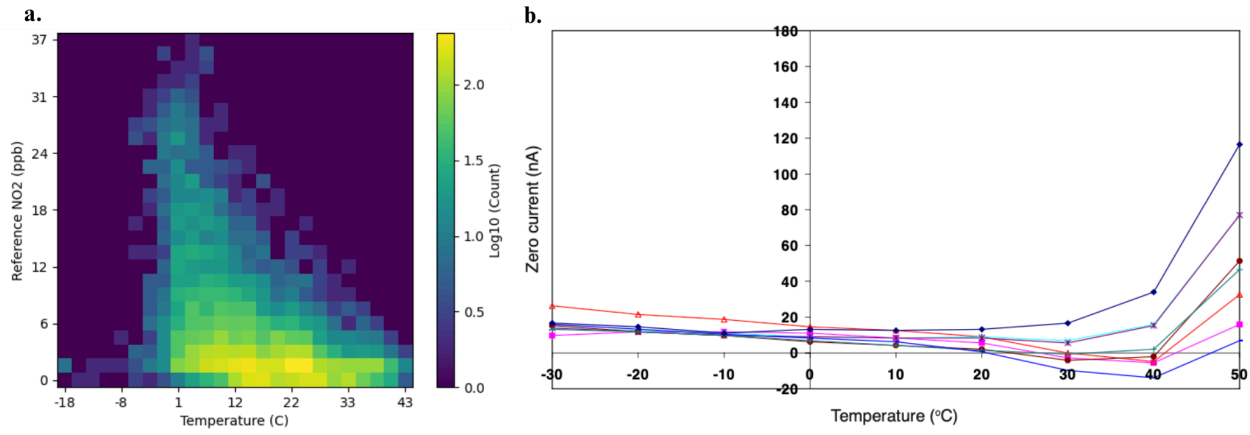

Figure 2a. *Heat map (log scale) of the target variable and temperature.* ***2b.*** *Zero temperature dependence of the Alphasense NO2-B43F low-cost sensor, sourced from the sensor manual. Each color represents a different sensor.*

At low temperatures, there is a strong positive linear relationship between reference and LCS NO_2; this holds true even at low concentrations. **Figure 3** demonstrates that LCS performance degrades at high temperatures regardless of the ambient NO_2 concentration, though at high temperatures lower concentrations are much more likely. In 2023, 13.4% of measurements were made at temperatures greater than °C. With this in mind, model performance at high temperatures will be evaluated.

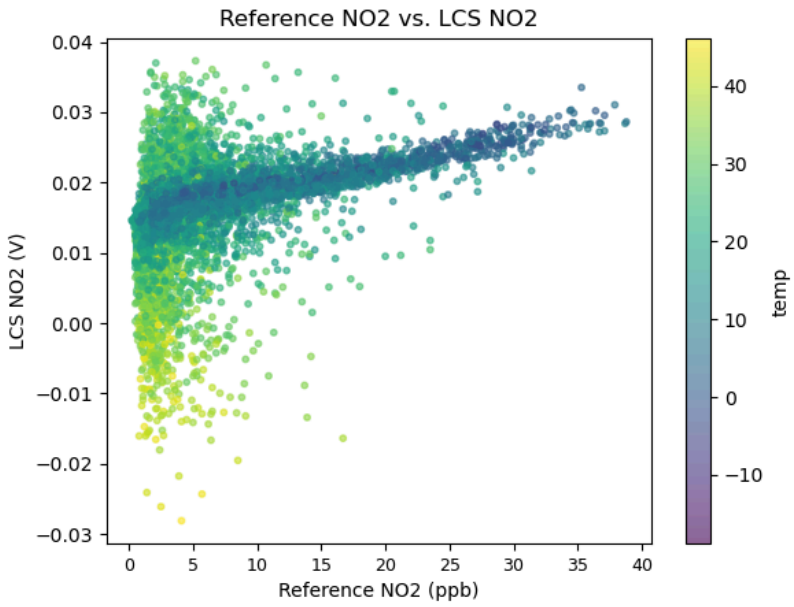

Figure 3. Reference NO_2 (ppb) versus low-cost NO_2 (V), colored by temperature.

Methods

Throughout the study period, the Breathe Providence low-cost sensor occasionally experienced power outages or was unplugged. This resulted in 1.6% of data to be missing across all features. Because methods such as reduced-feature models are not suitable, missing data were interpolated while the dataset was still ordered by date. Three interpolation methods were tested—linear, quadratic, and spline—and assessed as a measure of uncertainty. In addition, a column was added to count sensor deployment hours to account for drift over time.

The dataset was stratified on decile bins of the continuous target variable using StratifiedKFold , splitting 20% for testing and the remaining 80% into 5 training-validation folds. A preprocessor was built that uses PolynomialFeatures to generate interaction-only terms. Previous work has found that interaction terms can be significant features in linear models [Levy Zamora et al. ,2022]. Lastly, StandardScalar was used to scale all features. After feature engineering and scaling, there were 63 total features, each with a mean of 0 and standard deviation of 1.

Five models were tested per interpolation method: two linear (Ordinary Least Squares and Ridge Regression) and three non-linear (Random Forest, XGBoost, and Nearest Neighbors). Random states (5 total) were fixed, and model performance uncertainty was assessed via mean and standard deviation across splits. When applicable, maximum iterations were set to 1,000,000. Root mean squared error (RMSE), commonly used to evaluate LCS calibration methodologies, was used as the evaluation metric. In addition, the coefficient of determination (R^2) was calculated for the test set. GridSearchCV was used to tune hyperparameters.

	OLS	Ridge	Random Forest	XGBoost	Nearest Neighbors
Parameters	N/A	alpha	max_depth*	max_depth; n_estimators;** reg_alpha	n_neighbors; weights
Range	N/A	[1e-5,**1e-4**,1e-3]	[30, 50, **100**]	[**5,** 15, 30]; [10, 100, **500**]; [1e-1, **1e0**, 1e1]	[2, **3**, 4, 20]; [**'distance'**,'uniform']

Table 2. *Parameters and parameter ranges tuned for each model.* ***Bold*** *values indicate best parameters for linear interpolation. *max_depth and **n_estimators intentionally limited to 100 and 500, respectively, to reduce complexity.*

Results

Table 3 shows model performance for the full test set and temperatures >30°C ("High T"), with XGBoost performing the best. XGBoost achieved 49 standard deviations above the baseline, while OLS (worst model) performance achieved 41. Ridge Regression slightly outperformed OLS, but tuning revealed minimal improvement with small alphas (<0.001), indicating regularization had little effect. **Figure 4a** highlights the performance of Ridge and XGBoost.

a. Linear interpolation

	OLS	Ridge	Random Forest	XGBoost	Nearest Neighbors	Baseline
Overall RMSE	1.98 ± 0.05	1.98 ± 0.05	1.44 ± 0.03	1.20 ± 0.01	1.98 ± 0.03	6.12 ± 0.10
Overall R^2	0.89 ± 0.01	0.89 ± 0.01	0.94 ± 0.00	0.96 ± 0.00	0.90 ± 0.00	0.0
High T RMSE	2.04 ± 0.13	2.03 ± 0.13	1.45 ± 0.06	1.32 ± 0.05	1.75 ± 0.21	3.82 ± 0.05
High T R^2	-0.22 ± 0.32	-0.22 ± 0.31	0.39 ± 0.10	0.49 ± 0.10	0.13 ± 0.08	0.0

b. Quadratic interpolation

	OLS	Ridge	Random Forest	XGBoost	Nearest Neighbors	Baseline
Overall RMSE	1.99 ± 0.05	1.99 ± 0.05	1.45 ± 0.04	1.19 ± 0.02	1.96 ± 0.03	6.12 ± 0.10
Overall R^2	0.89 ± 0.01	0.89 ± 0.01	0.94 ± 0.00	0.96 ± 0.00	0.90 ± 0.00	0.0
High T RMSE	1.99 ± 0.13	1.99 ± 0.13	1.47 ± 0.05	1.33 ± 0.02	1.75 ± 0.21	3.82 ± 0.05
High T R^2	-0.17 ± 0.31	-0.17 ± 0.30	0.37 ± 0.13	0.48 ± 0.14	0.13 ± 0.08	0.0

c. Spline interpolation

	OLS	Ridge	Random Forest	XGBoost	Nearest Neighbors	Baseline
Overall RMSE	2.10 ± 0.07	2.10 ± 0.07	1.51 ± 0.02	1.30 ± 0.02	2.05 ± 0.04	6.12 ± 0.10
Overall R^2	0.88 ± 0.01	0.88 ± 0.01	0.94 ± 0.00	0.95 ± 0.00	0.89 ± 0.00	0.0
High T RMSE	2.04 ± 0.13	2.03 ± 0.13	1.45 ± 0.05	1.30 ± 0.02	1.74 ± 0.21	3.82 ± 0.05
High T R^2	-0.23 ± 0.32	-0.22 ± 0.30	0.39 ± 0.11	0.50 ± 0.12	0.14 ± 0.06	0.0

Table 3a. *Linear interpolation results.* ***3b.*** *Quadratic interpolation results.* ***3c.*** *Spline interpolation results.*

All models performed worse at high temperatures. While RMSE remained better than the baseline and similar to overall RMSE, the *relative* RMSE worsened significantly because NO_2 concentrations are typically low at high temperatures. Here, $\mathbf{R^2}$ becomes more informative: linear models, in particular, saw large $\mathbf{R^2}$ drops, performing worse than the baseline. This aligns with EDA findings, which revealed

nonlinear biases at high temperatures. **Figure 4b** illustrates Ridge's struggle to capture these biases compared to XGBoost.

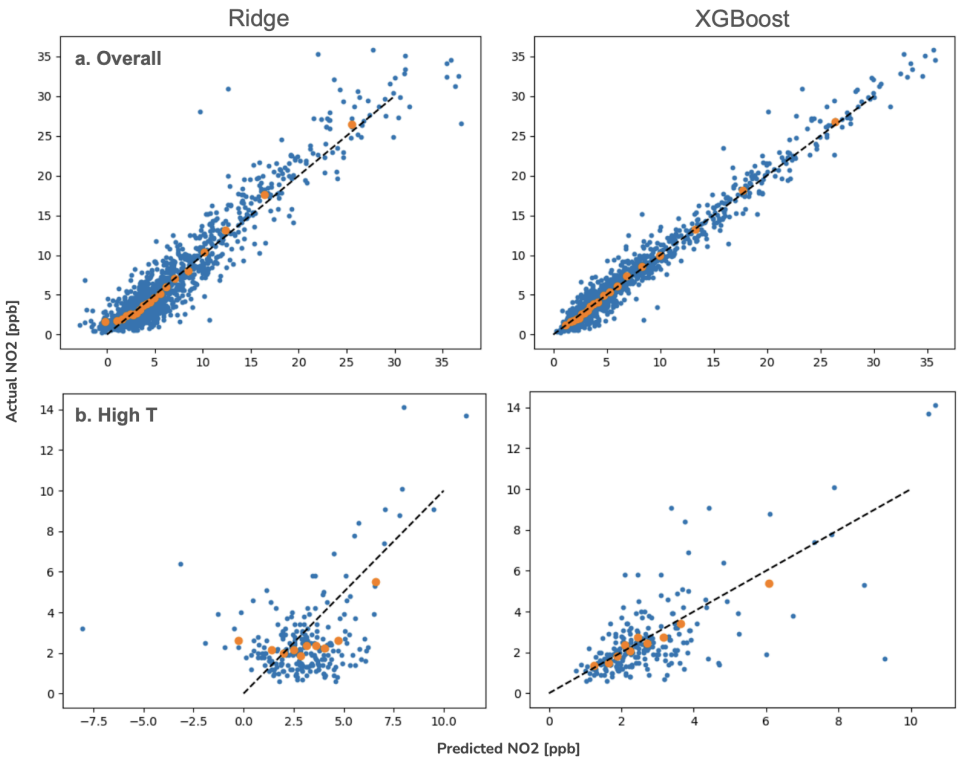

Figure 4. *Scatter plots of Ridge (best linear model) and XGBoost (best complex model) performance against target variable. Orange dots represent aggregated results grouped as 5% bins.*

Model rankings remained consistent across interpolation methods (**Figure 5**), though performance varied slightly. Linear and quadratic interpolation showed similar results, while spline interpolation significantly underperformed in the XGBoost model.

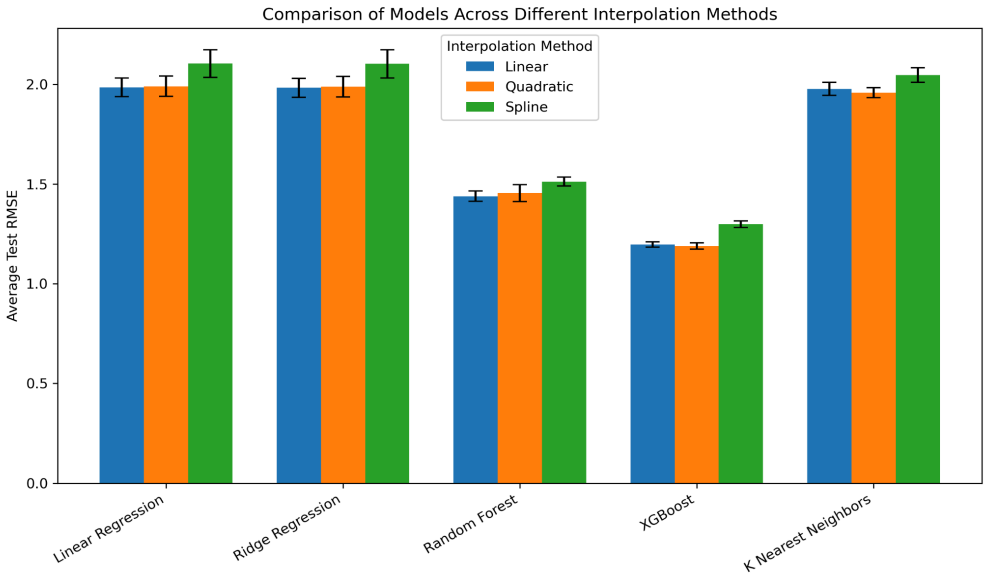

Figure 5. *Comparison of overall model performance across interpolation methods.*

Relative feature importances were consistent across interpolation methods; for simplicity, only figures from the linearly interpolated data are shown. There were distinct differences between global feature importances for Ridge regression (**Figure 6a and 6b**) and for XGBoost (**Figure 7a, 7b, and 7c**). More complex interaction terms had greater importance for Ridge. One surprise was that the raw NO_2 term did not show up as a top 10 feature, which means that the model's optimization algorithm identified ozone and temperature interactions as better predictors for reference NO_2 than the LCS' own raw NO_2 measurement. Future work could explore the frequency and significance of the original six features within high-weighted interaction terms for linear models.

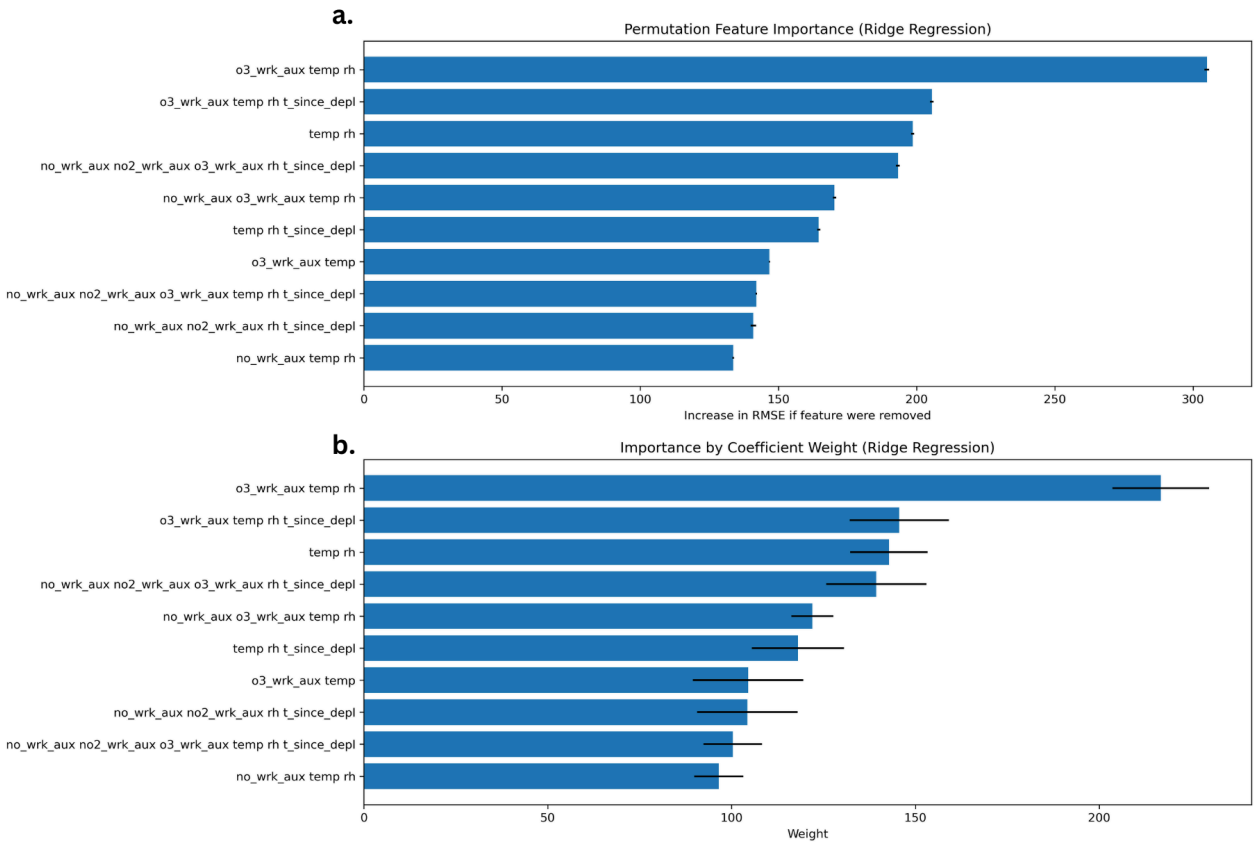

Figure 6a. *Mean absolute coefficient weights for Ridge model.* ***6b.*** *Mean permutation feature importance for Ridge model.*

The feature importances for XGBoost were more expected: the top term across multiple global feature importance methods was an interaction term between the raw NO_2 and NO measurements, typically followed by raw O_3 or raw NO_2. This indicates that the cross sensitivity of the NO_2 sensor to NO changes meaningfully with concentration. One surprise was that O_3 was a slightly more important feature than temperature; meaning, the O_3 cross sensitivity is more important than the temperature bias (though high O_3 and high temperatures frequently co-occur). Overall, because feature interactions are accounted for in the structure of tree models, it makes sense that the auto-generated interactions would be less important compared to Ridge.

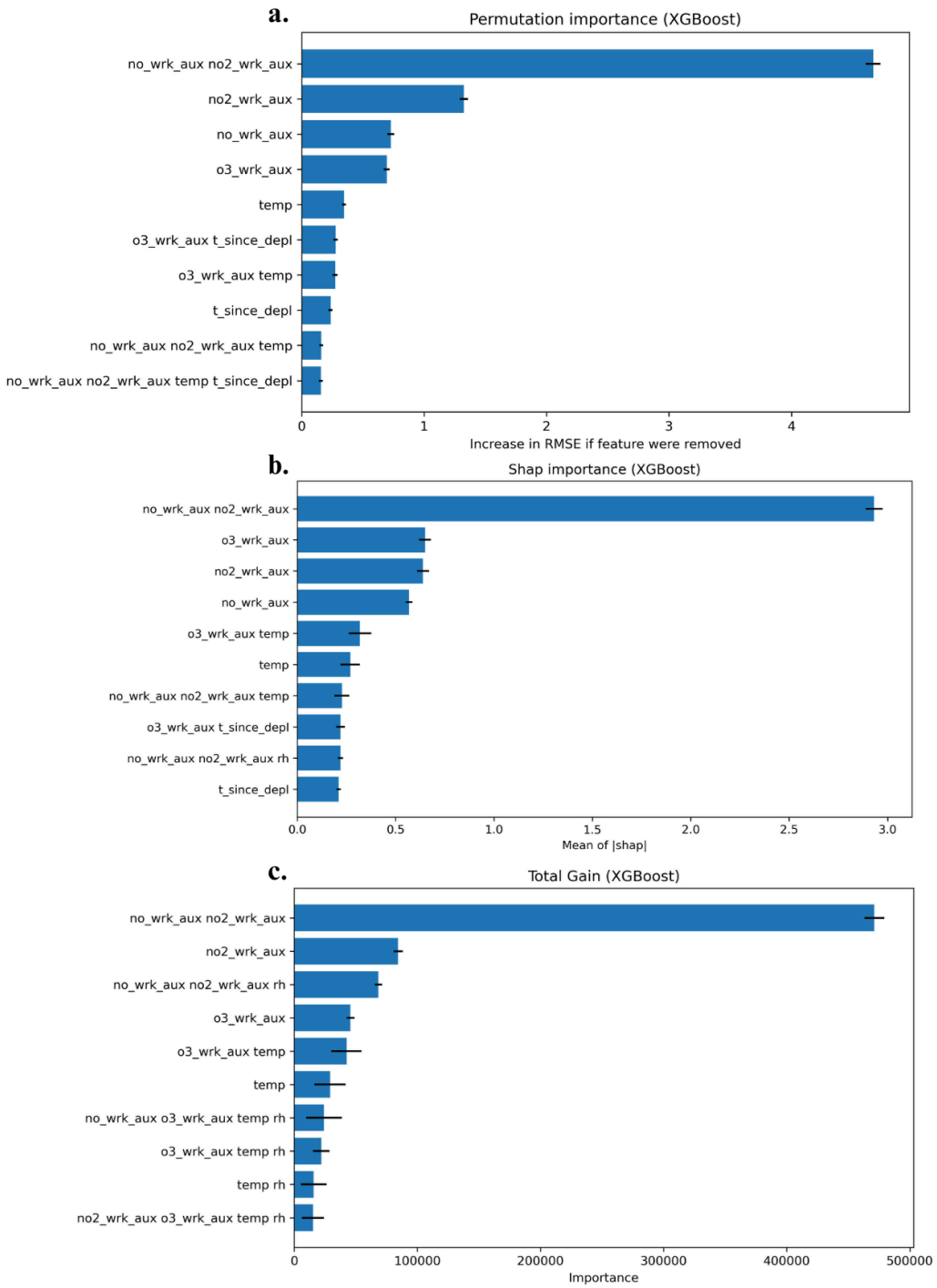

Figure 7a. *Mean permutation feature importance for XGBoost.* ***7b.*** *Mean global shap feature importance for XGBoost.* ***7c.*** *Mean total gain for XGBoost.*

An exploration of shap local feature importances (**Figure 8**) reveals more detail as well as slight differences between local and global importances. O_3 typically increases predicted NO_2 values, while NO_2 and the NO_2 - NO interaction term typically decreases the prediction.

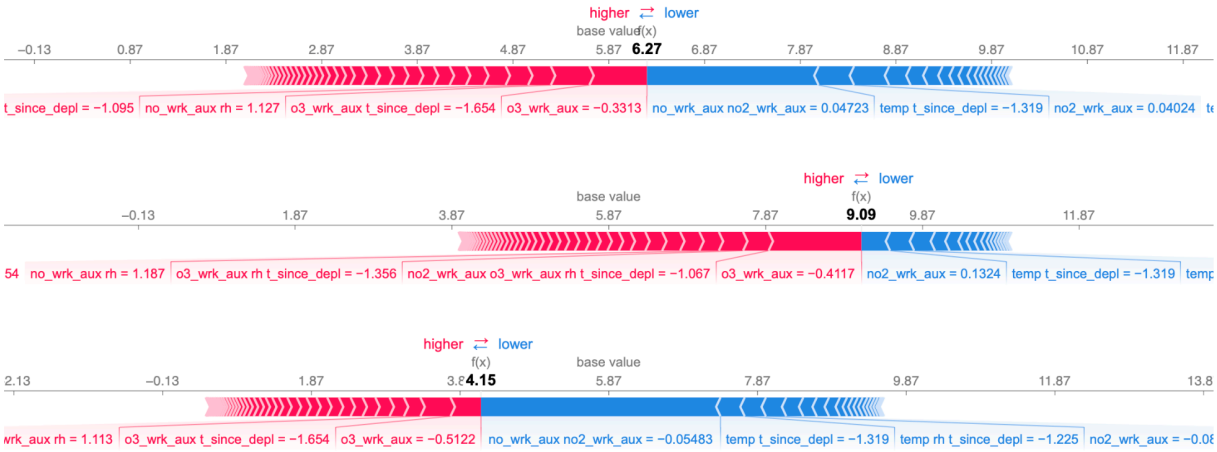

Figure 8. *Local shap importances for points 0, 1, and 2 of the first XGBoost model.*

Outlook

Ultimately, this work will form part of a larger field sensor calibration process. I plan to evaluate both the linear and more complex models' performance in this process, which will likely result in needing to simplify the tree-based models. I plan to retrain them without pre-generated interaction terms and experiment with further reducing XGBoost's n_estimators and Random Forest's max_depth. Future work will also assess model performance across seasons to understand the impact of seasonal variability and explore developing season-specific models. One weakness of my methodology is that the t_since_depl term, which I added to account for drift, may instead function as a proxy for time-of-year in the models. I would like to investigate this further and see if there is a better way to account for drift.

References

Bigi, A., Mueller, M., Grange, S. K., Ghermandi, G., & Hueglin, C. (2018). Performance of NO, NO_2 low-cost sensors and three calibration approaches within a real-world application. Atmospheric Measurement Techniques, 11(6), 3717–3735. https://doi.org/10.5194/amt-11-3717-2018

Frederickson, L. B., Sidaraviciute, R., Schmidt, J. A., Hertel, O., & Johnson, M. S. (2022). Are dense networks of low-cost nodes really useful for monitoring air pollution? A case study in Staffordshire. Atmospheric Chemistry and Physics, 22(22), 13949–13965. https://doi.org/10.5194/acp-22-13949-2022

Zamora, M. L., Buehler, C., Lei, H., Datta, A., Xiong, F., Gentner, D. R., & Koehler, K. (2022). Evaluating the Performance of Using Low-Cost Sensors to Calibrate for Cross-Sensitivities in a Multipollutant Network. ACS ES&T engineering, 2(5), 780–793. https://doi.org/10.1021/acsestengg.1c00367

Russell HS, Frederickson LB, Kwiatkowski S, Emygdio APM, Kumar P, Schmidt JA, Hertel O, Johnson MS. (2022). Enhanced Ambient Sensing Environment—A New Method for Calibrating Low-Cost Gas Sensors. Sensors, 22(19,:7238. https://doi.org/10.3390/s22197238

van Ratingen, S., Vonk, J., Blokhuis, C., Wesseling, J., Tielemans, E., & Weijers, E. (2021). Seasonal influence on the performance of low-cost NO_2 sensor calibrations. Sensors, 21(23), 7919. https://doi.org/10.3390/s21237919

Winter, A. R., Zhu, Y., Asimow, N. G., Patel, M. Y., & Cohen, R. C. (in review). A scalable calibration method for enhanced accuracy in dense air quality monitoring networks. Unpublished manuscript, Department of Chemistry, University of California Berkeley.

Zimmerman, N., Presto, A. A., Kumar, S. P. N., Gu, J., Hauryliuk, A., Robinson, E. S., Robinson, A. L., & Subramanian, R. (2018). A machine learning calibration model using random forests to improve sensor performance for lower-cost air quality monitoring. Atmospheric Measurement Techniques, 11(1), 291–313. https://doi.org/10.5194/amt-11-291-2018