Traffic and Air Quality in Toronto: Implications for TTC Strategy

Dario Greco, Lovera Lokeswara, Sophia Feng, Sanika Poojary

2026-02-12

Table of Contents

1 Abstract

1.1 Business Context

In 2024, the Toronto Transit Commission (TTC) made a formal commitment to sustainability and environmental performance through its organization-wide Innovation and Sustainability Program (TTC 2024). As Data Analysts supporting the TTC's Planning and Strategy, this analysis was conducted to inform the Director of Sustainability and Environmental Strategy, whose role is to oversee environmental performance targets, emissions reduction initiatives, and data-driven climate policy decisions across the organization. This analysis aims to evaluate whether road traffic meaningfully contributes to urban air pollution in Toronto, providing evidence to inform the TTC's sustainability strategy, funding advocacy, and long-term transit planning decisions.

1.2 Executive Summary

This study integrates traffic volume, air quality (NO_2), and meteorological data from February 2022 to December 2024 across five monitoring stations across Ontario to help evaluate the relationship between road traffic and urban air pollution in Toronto. Following data cleaning, imputation, and feature engineering, both exploratory analyses and predictive modeling approaches including XGBoost, Random Forest, Generalized Additive Models (GAM), and a stacking ensemble were implemented using time-series cross-validation. Model comparisons assessed performance with and without traffic-related features to determine their incremental predictive value. Results indicate that while traffic contributes to variation in NO_2, seasonal and meteorological factors play a more dominant role, and nonlinear models provide better predictive performance in capturing these complex relationships.

2 Methodology

2.1 Data Collection and Integration

We integrated three independent data sources: air quality measurements, traffic volumes, and weather conditions into a unified dataset spanning February 2022 to December 2024 across five urban monitoring stations (i.e. Sarnia, Ottawa, Sault St. Marie, Downtown Toronto, Toronto West). Due to the limited data available for Toronto stations, we analyzed three stations outside Toronto to better understand the true underlying relationship between air quality and traffic across regions in Ontario.

2.1.1 Data Formatting

Air quality data was cleaned by extracting station identifiers and aggregating hourly measurements into daily averages. Invalid inputs were replaced with missing values for later imputation.

Traffic data was reshaped from wide to long format and geographic coordinates were mapped to station identifiers. When multiple sensors covered the same location, counts were averaged for that station-date combination.

Weather data were standardized by normalizing variable names and units across different source files. Wind direction was converted from tens of degrees to standard degrees, and dates were normalized to a consistent format.

2.1.2 Data Integration

Weather, traffic and air quality stations across regions we matched manually based on their proximity to each other. The three datasets were merged temporally on (Station, Date) using outer joins, which retained all available observations even when individual data sources had gaps for specific station-date combinations. This approach maximized information retention across the full study period. The integrated dataset was then filtered to the consistent date range of February 2022 to December 2024.

2.1.3 Handling Missing Data

Our data sources contain gaps from potential outages and maintenance. Rather than discarding incomplete records, we employed IterativeImputer from Sklearn to impute missing values as a function of the other known variables. For each station independently, the algorithm modeled dependencies between measurements (e.g., how traffic and weather patterns relate to NO_2 concentrations) and used these relationships to fill gaps. Post-imputation validation ensured all values were physically realistic: negative values for concentrations and traffic were clipped to zero, wind direction was bounded to 0–360°, and traffic counts were rounded to integers.

2.1.4 Output

Five station-specific cleaned datasets were generated with complete time series, ready for analysis and modeling. For detailed definitions of all variables included in these datasets, refer to the Data Dictionary.

2.2 Data Analysis

The following analysis explores the relationship between traffic volume and NO_2 concentrations across monitoring stations. Correlation matrices indicate that traffic-related variables generally exhibit a negative association with NO_2, which contradicts what is expected, since NO_2 is a part of TRAP (Traffic Related Air Pollutant). Meteorological variables such as wind speed, temperature, and precipitation exhibit varying correlations on air quality.

Seasonal trend analysis further reveals systematic temporal variation in NO_2 levels, with higher concentrations typically observed during colder months (i.e., October to April) and lower levels during warmer periods (i.e., May to September). This seasonal pattern aligns with established atmospheric behavior, where weaker sunlight, lower temperatures, and more stable air conditions in winter reduce photochemical removal and trap pollutants near the surface, while stronger solar radiation and enhanced atmospheric mixing in summer promote dispersion and chemical transformation of NO_2, resulting in lower observed concentrations (Bai et al. 2025). Together, these

exploratory findings motivate the inclusion of traffic, temporal, and meteorological features in the predictive modeling framework.

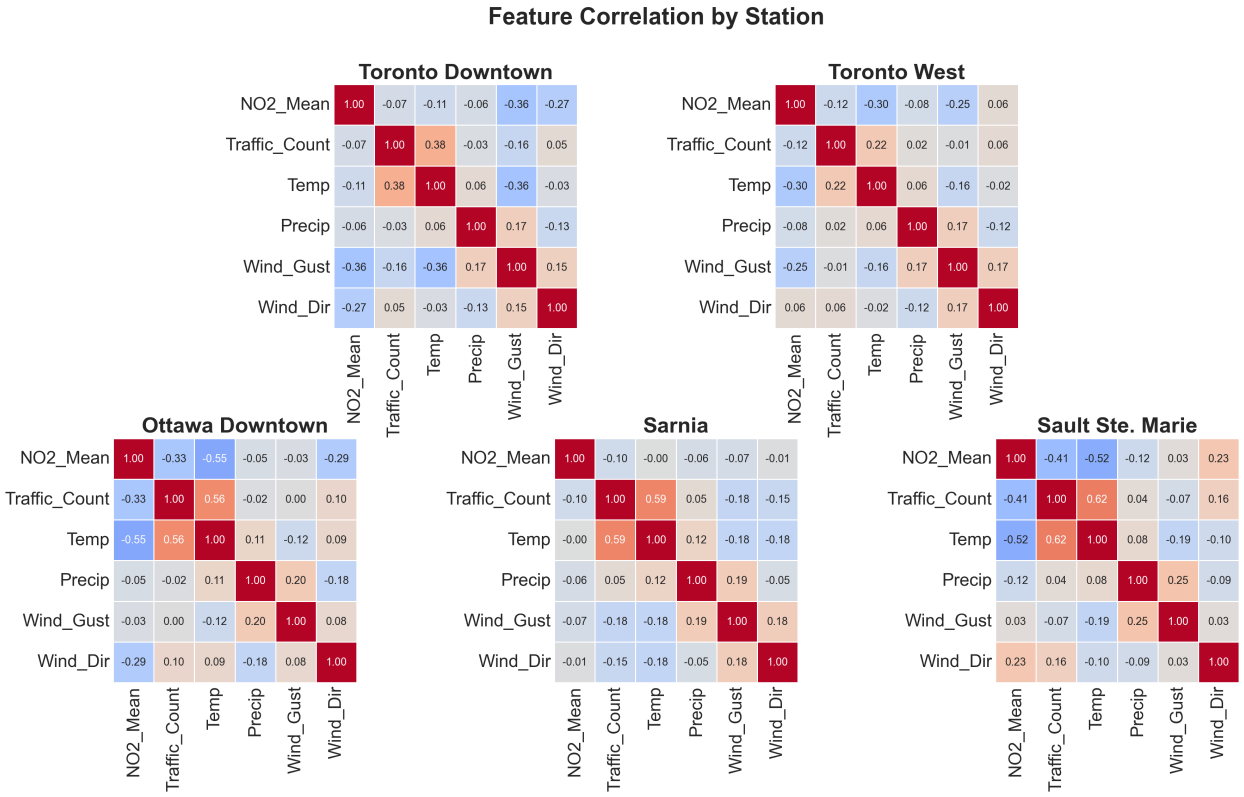


Figure 1: Correlation Matrices by Station

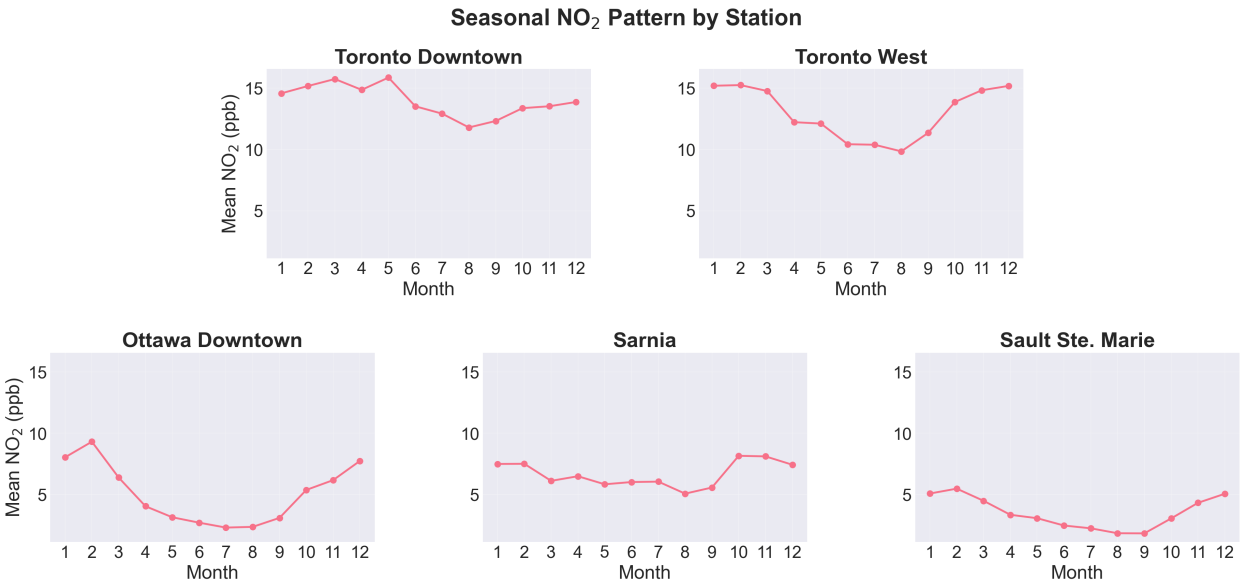


Figure 2: Seasonal NO(_2) Patterns by Station

2.3 Model Architecture

2.3.1 Feature Engineering

To better capture temporal structure and provide informative predictors, several calendar-based features are derived from the date variable. These include the day of the week (0–6), month (1–12), a binary weekend indicator (Is_Weekend), and a seasonal index (1–4) inferred from the month. Such temporal features enable the model to learn recurring weekly, seasonal, and economic patterns commonly observed in air-quality dynamics.

Autoregressive characteristics of NO_2 are incorporated through lagged and rolling-window features. Specifically, NO2_Lag_1 and NO2_Lag_7 represent NO_2 concentrations from the previous day and the previous week, respectively, while NO2_Rolling_7 denotes the 7-day moving average of NO_2.

In addition, Temp_Rolling_7 captures smoothed temperature trends over the same period. Wind direction and magnitude are encoded using continuous vector components defined as Wind_NS = Wind_Gust $\cdot$ cos(Wind_Dir) and Wind_EW = Wind_Gust $\cdot$ sin(Wind_Dir), providing a physically meaningful representation of wind behavior suitable for regression-based learning.

2.3.2 Datasets Used

To assess whether incorporating traffic information improves predictive performance, two modeling scenarios were constructed: one including traffic-related features and one excluding them. This design enables a direct comparison of model accuracy under differing feature availability while keeping all other modeling components consistent.

- **Scenario A (with traffic features).** This setting incorporates predictors describing traffic dynamics and their interactions with meteorological conditions. Specifically, Traffic_Lag_1 denotes the previous day's traffic count, Traffic_Rolling_7 the 7-day moving average, Traffic_x_Wind the interaction between traffic volume and wind gust, and Temp_x_Traffic the interaction between temperature and traffic. The full feature set is listed in Appendix A.[1]
- **Scenario B (without traffic features).** This setting uses the same predictors outlined in Scenario A but excludes all traffic related predictors. The complete feature set is provided in Appendix A.[2]

2.3.3 k-fold Cross-Validation (expanding window)

Because the dataset is time series in nature, the cross-validation strategy must preserve temporal ordering rather than rely on random shuffling and partition the training dataset into k equal parts (Lysy 2026). Two common approaches that respect temporal structure are the sliding window and the expanding window methods. In this study, we adopt the expanding window approach to ensure that each training fold contains only observations that occur prior to the corresponding validation period, as illustrated in Figure Figure 3.

[1] See Appendix A
[2] See Appendix A

The expanding window procedure is applied using five folds on the training portion of the data only. In each fold, the model is trained on an earlier time interval and validated on the immediately following contiguous time block. This design prevents temporal leakage, where information from future observations could otherwise influence past predictions, thereby providing a more realistic estimate of out-of-sample performance.

Figure 3: Expanding-window time-series cross-validation diagram

2.3.4 Model Definition

All models are evaluated using a test dataset that corresponds to the most recent 20% of the dataset, ensuring that performance reflects true out-of-sample prediction in a time-series setting.

- **XGBoost** XGBoost is a gradient-boosted tree model that can capture complex patterns in data, including nonlinear relationships, interactions between variables, and threshold effects common in environmental data. The top 10 most important features for each station are shown below.

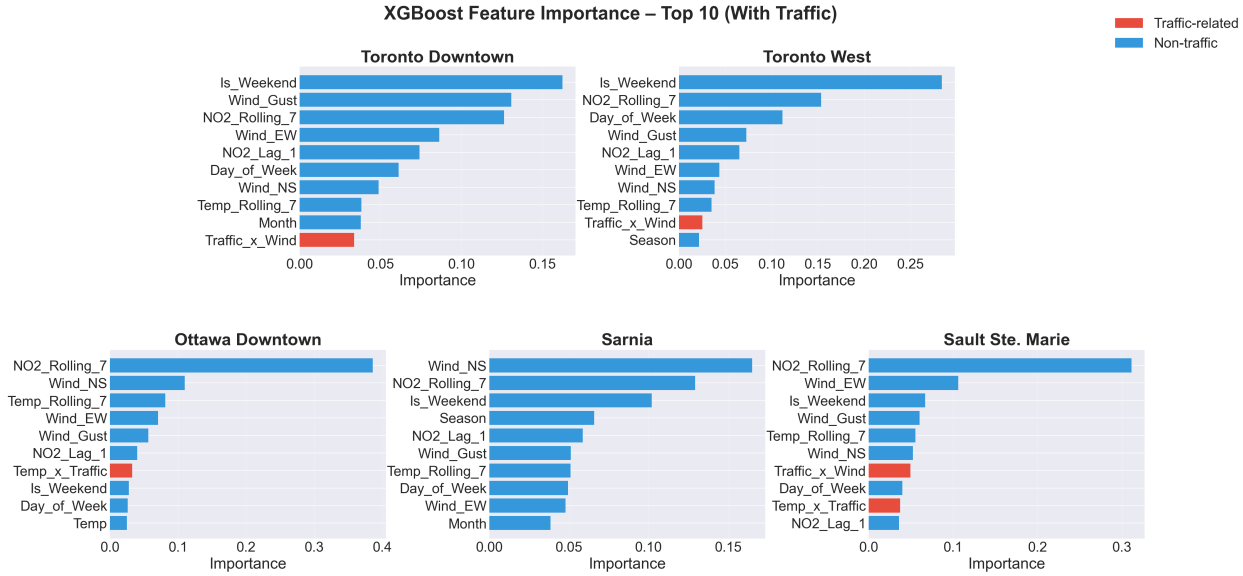


Figure 4: XGBoost Feature Importance With Traffic

- **Random Forest** Random Forest is an ensemble of bagged decision trees which provides a nonlinear benchmark and performs reliably with heterogeneous feature types, making it well suited for comparative evaluation.
- **GAM Model** The generalized additive model (GAM) captures smooth, nonlinear relationships between air quality and key predictors such as weather and time-based factors (i.e., Day_of_Week, Month, Is_Weekend, Season) by modeling the response variable as a sum of smooth functions of its predictors. Its main advantage is interpretability, allowing us to understand how each factor individually influences NO_2 levels.
- **Stacking Ensemble** The stacking architecture combines multiple learners to improve generalization, using XGBoost, Random Forest, and Linear Regression as base learners and Ridge Regression to get a final combined model. This ensemble strategy integrates the strengths of nonlinear tree-based models with a stable linear combiner, often leading to improved predictive performance.

2.3.5 Model Performance and Selection

Model performance is evaluated using RMSE, which penalizes large prediction errors and serves as the primary metric for model selection. The comparative performance of all models across monitoring stations is illustrated in the figures below.

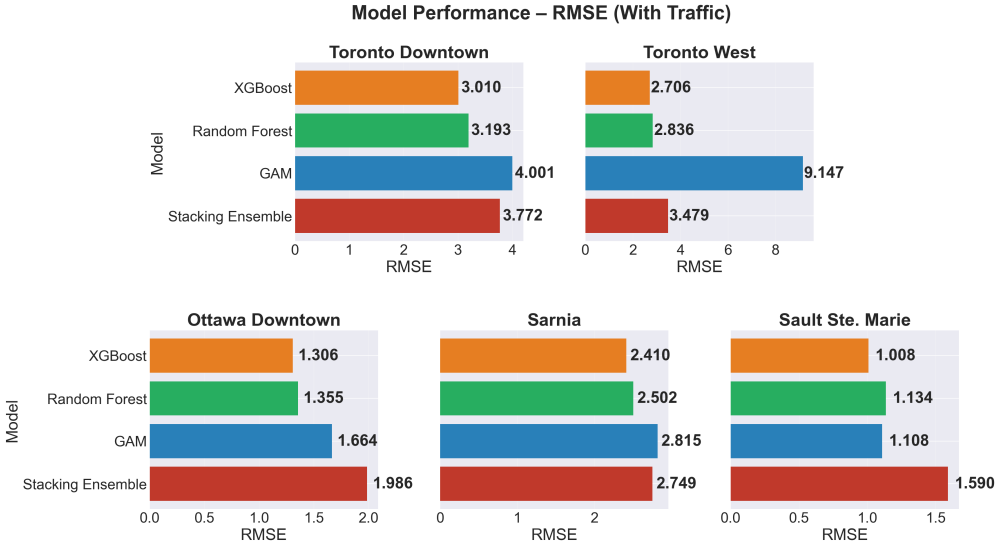


Figure 5: Model Performance with Traffic

Across both scenarios, XGBoost consistently demonstrates the strongest predictive accuracy, achieving the lowest RMSE at most stations. Notably, the performance across models, including and excluding traffic-related features, was similar. Additional diagnostic visualizations, including cross-validation versus test RMSE and residual distributions, are provided in the Appendix C to support the analysis. Feature importance plots are also included to aid interpretability.

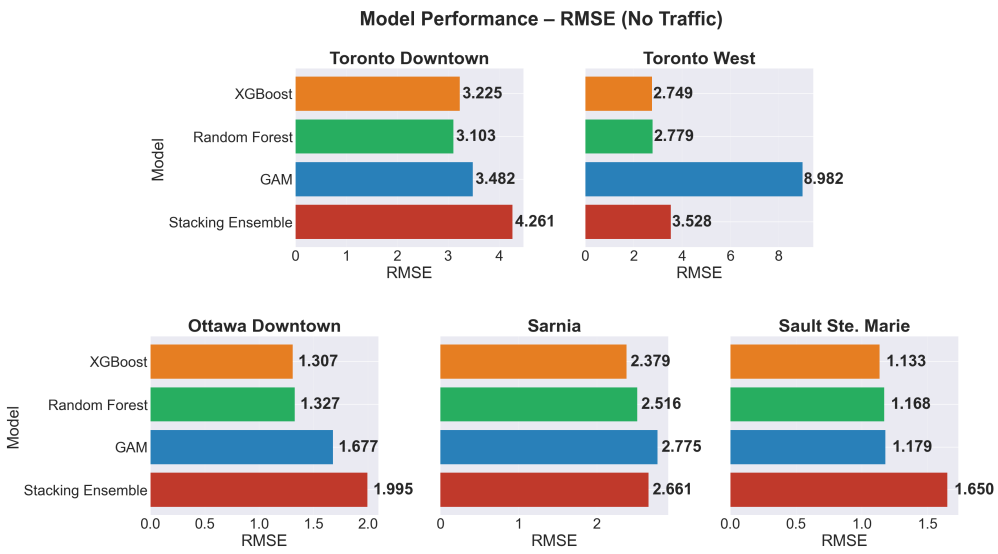


Figure 6: Model Performance without Traffic

3 Results

3.1 Analysis Result

Overall, the results of this analysis indicate that while there is a relationship between traffic volume and air quality in Toronto, this relationship is weak. During the initial analysis phase, the correlation matrices suggested a weak negative relationship between traffic and NO_2. However, when seasonal patterns and meteorological variables were considered, these covariates consistently demonstrated high feature importance across multiple model types and cities.

Specifically, Is_Weekend, NO2_Rolling_7 and Wind_Gust consistently appear to be the features with the strongest importance. Is_Weekend offers an interesting perspective on NO_2 behaviour, though traffic is potentially lower on weekends, other major polluters like factories and mills also may have lower activity.

These findings suggest that weather conditions and existing pollution levels play a more substantial role in determining air quality than traffic volume. As a result, although this analysis provides exploratory insights, it does not allow for definitive answers to several of the originally proposed research questions due to data and methodological limitations. First, while preliminary correlations suggest a weak negative association between traffic and air quality, the question of whether there is a clear positive or negative association between traffic and air quality cannot be conclusively determined because the observed relationships are highly sensitive to seasonal and meteorological controls. Similarly, the question of whether increased traffic volume is associated with higher concentrations of air pollutants in Toronto cannot be answered with certainty, as background pollution levels and atmospheric conditions appear to have stronger influence than traffic volume alone.

While nonlinear models such as XGBoost demonstrated improved predictive performance, this does not provide definitive causal evidence of nonlinear effects between traffic volume/air quality levels and rather, it suggests model flexibility better captures complex patterns in the data. Finally,

because the dataset is restricted to a limited time frame, it is not possible to determine whether the relationship between traffic and air quality changes over time beyond seasonal fluctuations. Consequently, the findings should be interpreted as exploratory rather than conclusive with respect to the original research questions.

3.2 Limitations and Future Directions

A major limitation of this analysis was the spatial discrepancy between the data measurement stations. For instance, the meteorological monitoring station used for Toronto West was approximately 7.89 km away from the air quality station, while the traffic station was roughly a kilometer away from the air quality station. A similar spatial mismatch affected the other locations, introducing potential bias when inferring a relationship between traffic at Point A and air quality at Point B. Another significant limitation was the temporal resolution of our data. Although air quality measurements were available at an hourly level, traffic data was only available as a daily aggregate. By consolidating the air quality data to a daily level to match the traffic data, we lost important hourly information.

In the future, to conduct a more valuable analysis of the research questions, prioritizing the use of collocated hourly air quality, meteorological, and traffic monitoring stations is recommended. For example, Kendrick et al. (al. 2015) demonstrates that the relationship between air quality and traffic is significantly affected by the way that data is temporally aggregated. While our current analysis on publicly available data suggests only a weak association, accessing higher resolution data would allow for a more accurate characterization of the impact of traffic on urban air quality.

3.3 Conclusions

For the TTC and the Department of Sustainability, the findings of this analysis provide a preliminary analysis of the relationship between traffic and air quality. The evidence indicates that while traffic is associated with NO_2 , it is not the dominant driver within the limits of the available data. Instead, seasonal patterns, weather conditions, and existing pollution levels consistently exert stronger influence on observed air quality outcomes.

Overall, the results of this analysis showcase that current publicly available Toronto data is not sufficient to conclusively deduce the relationship between traffic and air quality. As a result, these findings support the need for further investments in monitoring stations of air quality, traffic, and weather data that are all close in proximity and of high resolution. These investments could consequently be used to better establish the relationship between traffic and air quality and hence be used to support the TTC's sustainability strategy, capital planning decisions, and advocacy for climate-focused transit funding.

4 References

al., Kendrick et. 2015. "Diurnal and Seasonal Variations of NO, NO_2 and $PM_{2.5}$ Mass as a Function of Traffic Volumes Alongside an Urban Arterial." *Atmospheric Environment* 122: 133–41. https://doi.org/10.1016/j.atmosenv.2015.09.019.

Bai, Xuehui, Yi Wang, Lu Gui, Minghui Tao, and Mingyu Zeng. 2025. "Comparing the Influences on NO2 Changes in Terms of Inter-Annual and Seasonal Variations in Different Regions of China: Meteorological and Anthropogenic Contributions." *Remote Sensing* 17 (1): 121. https://doi.org/10.3390/rs17010121.

Lysy, Martin. 2026. "Key Concepts in Statistical Learning."

TTC. 2024. "TTC Innovation and Sustainability Strategy 2024-2028." https://www.ttc.ca/about-the-ttc/Moving-toward-a-sustainable-future/Innovation-and-Sustainability-Strategy.

5 Data Sources

The data was obtained by web scraping the following links:

- **Air Quality:** Hourly NO_2 concentrations from monitoring stations. Data Link
- **Traffic:** Daily vehicle counts from permanent detection systems. Data Link
- **Weather:** Daily meteorological data (temperature, precipitation, wind speed and direction). Data Link

6 Appendix

6.1 Appendix A — Feature Set

6.1.1 Model with Traffic

feature_cols = [“Traffic_Count”, “Temp”, “Precip”, “Wind_Gust”, “Day_of_Week”, “Month”, “Is_Weekend”, “Season”, “NO2_Lag_1”, “NO2_Lag_7”, “Traffic_Lag_1”, “NO2_Rolling_7”, “Traffic_Rolling_7”, “Temp_Rolling_7”, “Traffic_x_Wind”, “Temp_x_Traffic”, “Wind_NS”, “Wind_EW”]

6.1.2 Model without Traffic

feature_cols = [“Temp”, “Precip”, “Wind_Gust”, “Day_of_Week”, “Month”, “Is_Weekend”, “Season”, “NO2_Lag_1”, “NO2_Lag_7”, “NO2_Rolling_7”, “Temp_Rolling_7”, “Wind_NS”, “Wind_EW”]

6.2 Appendix B — Model Hyperparameter

6.2.1 XGBoost

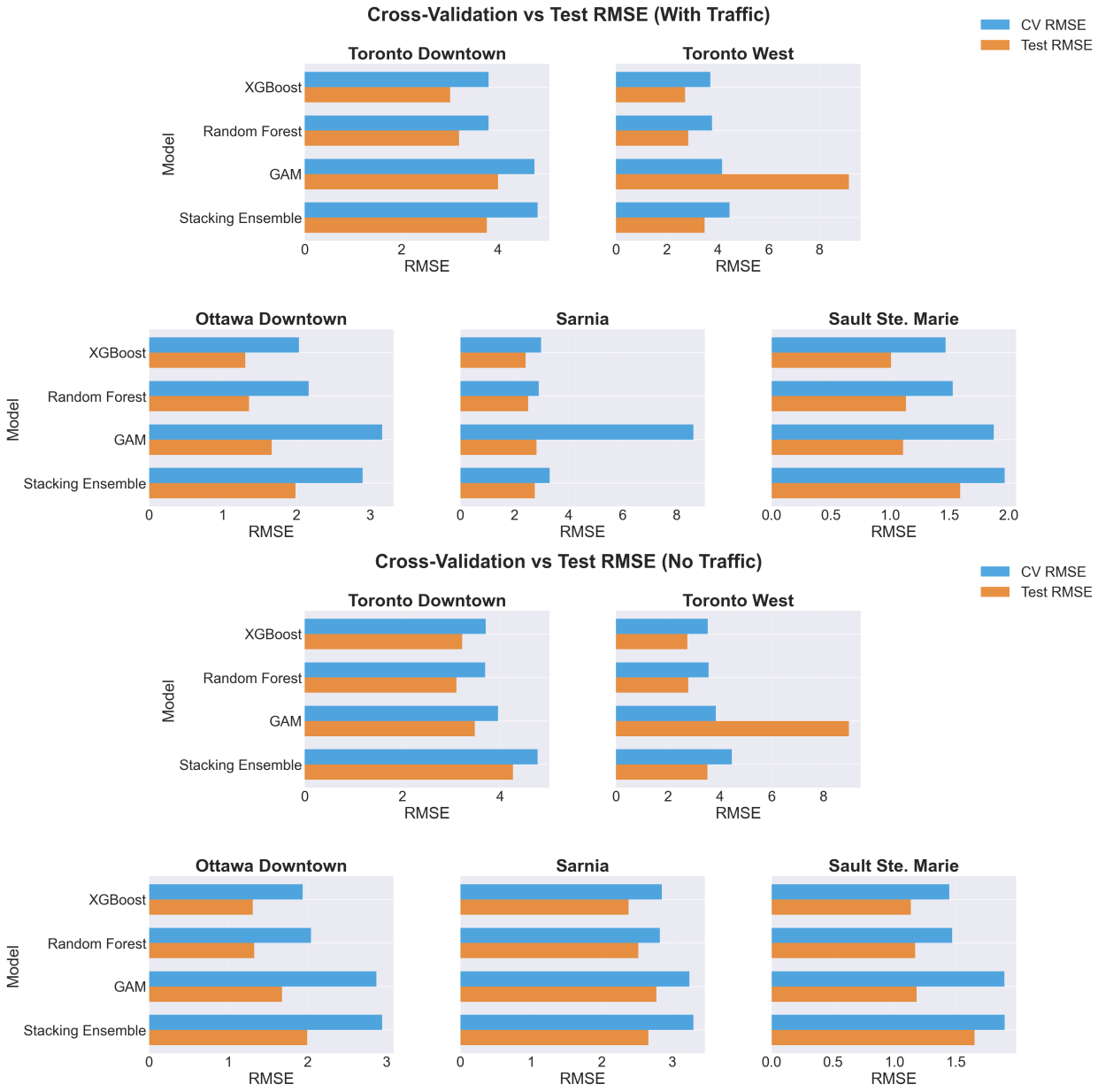
The XGBoost model is configured with the following key hyperparameters: XGBRegressor(n_estimators=200, learning_rate=0.05, max_depth=6, min_child_weight=3, subsample=0.8, colsample_bytree=0.8, random_state=42, n_jobs=-1, verbosity=0).

6.2.2 Random Forest

The Random Forest baseline is configured as RandomForestRegressor(n_estimators=200, max_depth=12, min_samples_split=5, min_samples_leaf=2, max_features=“sqrt”, random_state=42, n_jobs=-1, verbose=0).

6.3 Appendix C — Model Performance Evaluation

6.3.1 Cross



6.3.2 Residual Distribution