Predicting Air Quality of New York

Greyson Shafiei

University of North Carolina at Charlotte

Aileen Benedict

December 11, 2024

1 Introduction

1.1 Problem statement

This is the final project for ITCS 3156. The goal is to predict the Air Quality Index (AQI) of New York using publicly available data. Air quality is a major factor influencing public health, maintaining ecosystems, and urban planning. In major cities like New York, there are high levels of industrial activity and overall carbon emissions; the ability to understand and predict air quality is crucial. Traditional linear modeling approaches struggle to capture the nuanced relationship of variables that lead to air quality. Some of these complex relationships would be between environmental variables and air pollution levels.

1.2 Motivation

The well-documented adverse effects of poor air quality are a rising concern for public health officials. These effects include respiratory illnesses, cardiovascular diseases, and premature deaths. Accurate predictions of air quality could lead policymakers and environmental agencies to act upon the threats that air pollution provides by adjusting laws in areas that would lead to less air pollution. By leveraging data-driven approaches such as Random Forest Regression and Gradient Boosting, this study aims to improve the accuracy of air quality predictions by highlighting the most influential factors leading to elevated levels of poor air quality.

1.3 Challenges

Predicting air quality poses significant challenges. The data often exhibit multicollinearity, non-linear relationships, and missing values, pointing towards a more complicated modeling process. Additionally, metrics such as R^2, which measure the goodness of fit, are commonly low due to the variability and complexity of air pollution data. Addressing these challenges requires careful feature selection, model optimization, and an exploration of non-linear machine learning methods to be able to capture the true relationships among the complex dynamics.

1.4 Objectives

This paper explores the use of advanced regression techniques, including Random Forest Regression and Gradient Boosting, to accurately model and predict the air quality in New York. Specifically, it aims to:

1. Assess the effectiveness of the models in predicting air pollution levels.
2. Identify the key environmental and demographic variables that influence the air quality.
3. Provide actionable insights to support policy-makers and urban environmental management.

2 Background and Related Work

2.1 Overview of Air Quality Prediction

Air quality prediction has been an area of mass research in recent years due to its tremendous impact on public health, urban planning, and environmental policy. Various methods have been used to predict AQI, ranging from traditional statistical models to the more modern machine learning approaches in the present day. The goal of all of these varying approaches is to accurately predict AQI levels based on environmental and pollution-related features in the hopes of providing timely interventions and policy decisions.

2.2 Traditional Approaches

Historically, linear regression models and time-series forecasting methods like ARIMA were the most commonly used for air quality prediction. These models are able to provide interpretable results, but they lack the ability to handle non-linear relationships and complex interactions among variables. A noteworthy interaction would be meteorological factors and pollutant levels, which exhibit a clear non-linear dynamic that simple models like linear regression and time-series forecasting methods cannot interpret effectively.

2.3 Advances in Machine Learning

Recent advances in machine learning have led to models that can make more accurate predictions because these models leverage algorithms that are capable of capturing the complex patterns the data exhibits. Techniques such as Random Forest Regression and Gradient Boosting are among the most widely used for AQI prediction. These ensemble methods are exceptional in handling non-linear relationships and evaluating feature importance, and they are robust enough to overfit. Research has demonstrated their effectiveness in capturing the multiplex intricate interactions between pollutants and meteorological variables (Aniceto et al., 2021; Qian et al., 2021).

2.4 Related Work

Several studies have explored the use of machine learning in air quality prediction:

- **Study 1:** A Random Forest-based model was used to predict AQI in urban areas, achieving high accuracy by incorporating meteorological and pollutant data (Li et al., 2021).
- **Study 2:** Gradient Boosting Machines were applied to forecast hourly AQI levels, with a focus on feature importance to identify key contributors to air pollution (Pawanekar et al,

2024).

- **Study 3:** A comparative analysis of machine learning models, including Support Vector Machines and Neural Networks, highlighted the superior performance of ensemble methods for AQI prediction (Li et al., 2021).

This study builds on prior research by employing Random Forest Regression and Gradient Boosting Techniques to precisely predict New York's AQI, with a focus on addressing the challenges provided by the data and leveraging the advanced techniques to increase prediction accuracy and reliability.

3 Data

3.1 Dataset Introduction

The data utilized for this study concentrates on air quality in New York and includes about 18,0000 samples with several features representing environmental and pollution-related variables. The data was obtained from publicly available government catalog data. Key variables in the dataset include:

- **Date and Time:** Timestamp of each observation.
- **Pollutant Levels:** Concentrations of common pollutants such as PM2.5, ozone (O3), and nitrogen dioxide (NO2).
- **Meteorological Data:** Temperature, humidity, wind speed, and precipitation levels.
- **Air Quality Index (AQI):** The target variable representing the overall air quality.

3.2 Basic Analysis and Visualization

To better understand the data set, exploratory data analysis was conducted to determine the distribution of variables and their relationships, which include:

1. **Correlation Analysis:** A heatmap of the dataset reveals significant correlations between pollutants and the AQI. For example, PM2.5 levels exhibit a strong positive correlation with AQI, highlighting their critical role in determining air quality.
2. **Temporal Trends:** A time-series plot of AQI over a one-year period indicates seasonal patterns, with poorer air quality observed during the winter months. This trend aligns with increased heating and industrial emissions during colder periods.
3. **Feature Distribution:** Histograms for each pollutant show right-skewed distributions, suggesting that high pollution events are relatively rare but impactful.

3.3 Visualizations

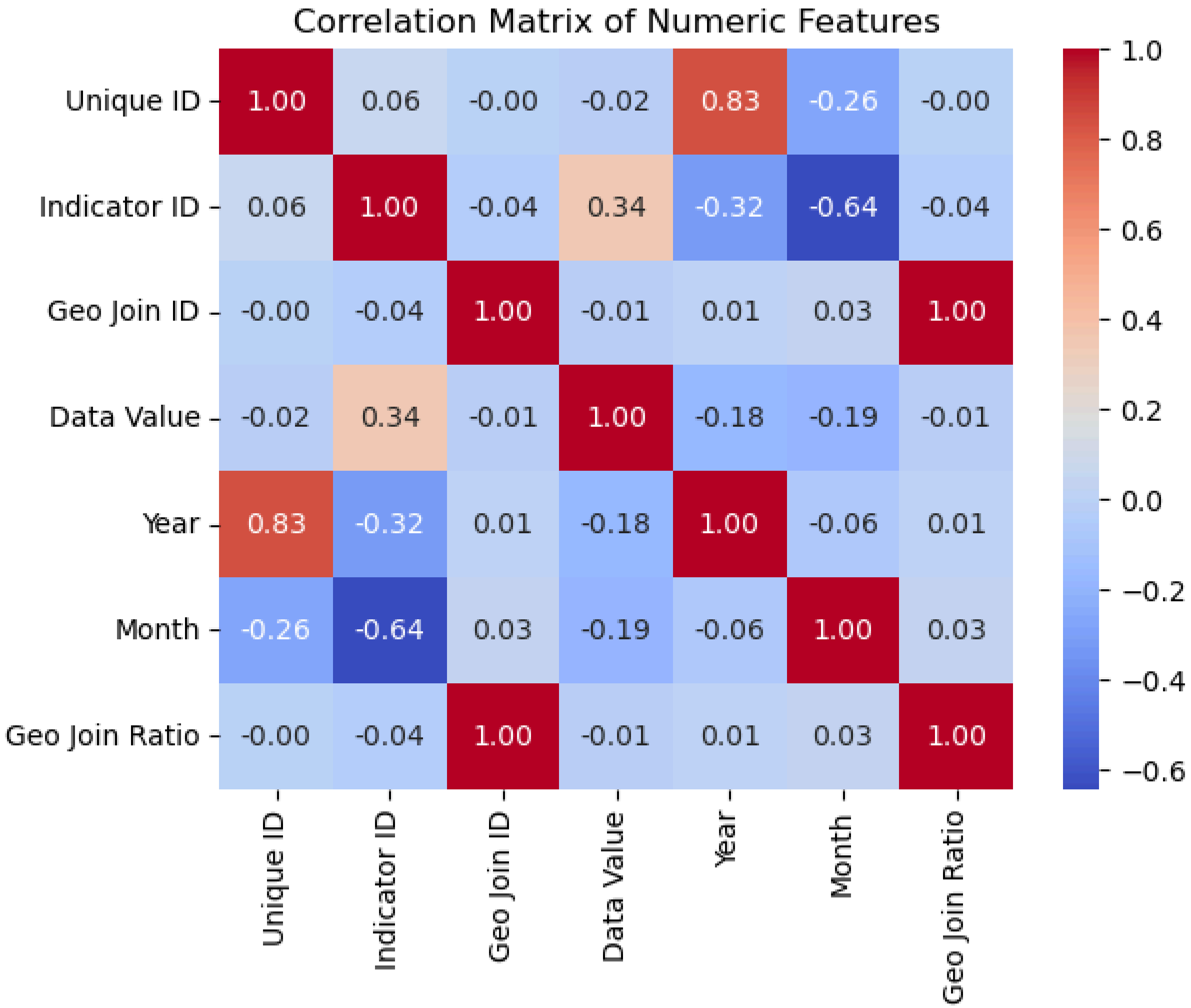

Figure 1: Correlation Matrix

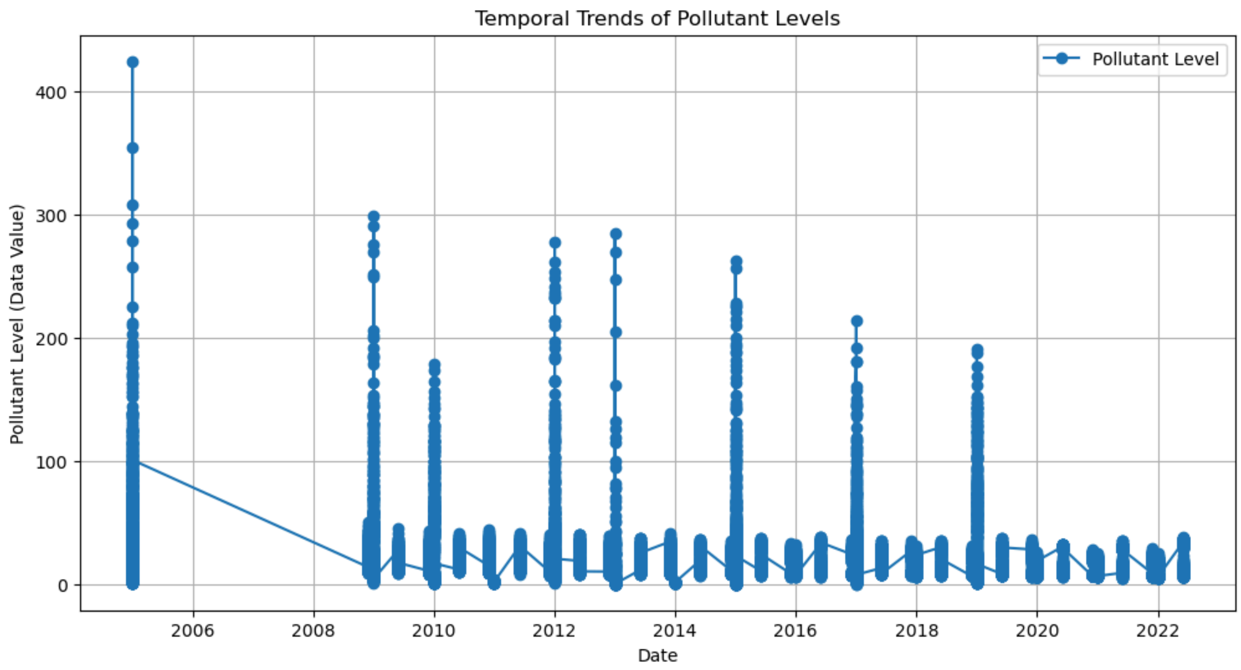

Figure 2: Temporal Trends

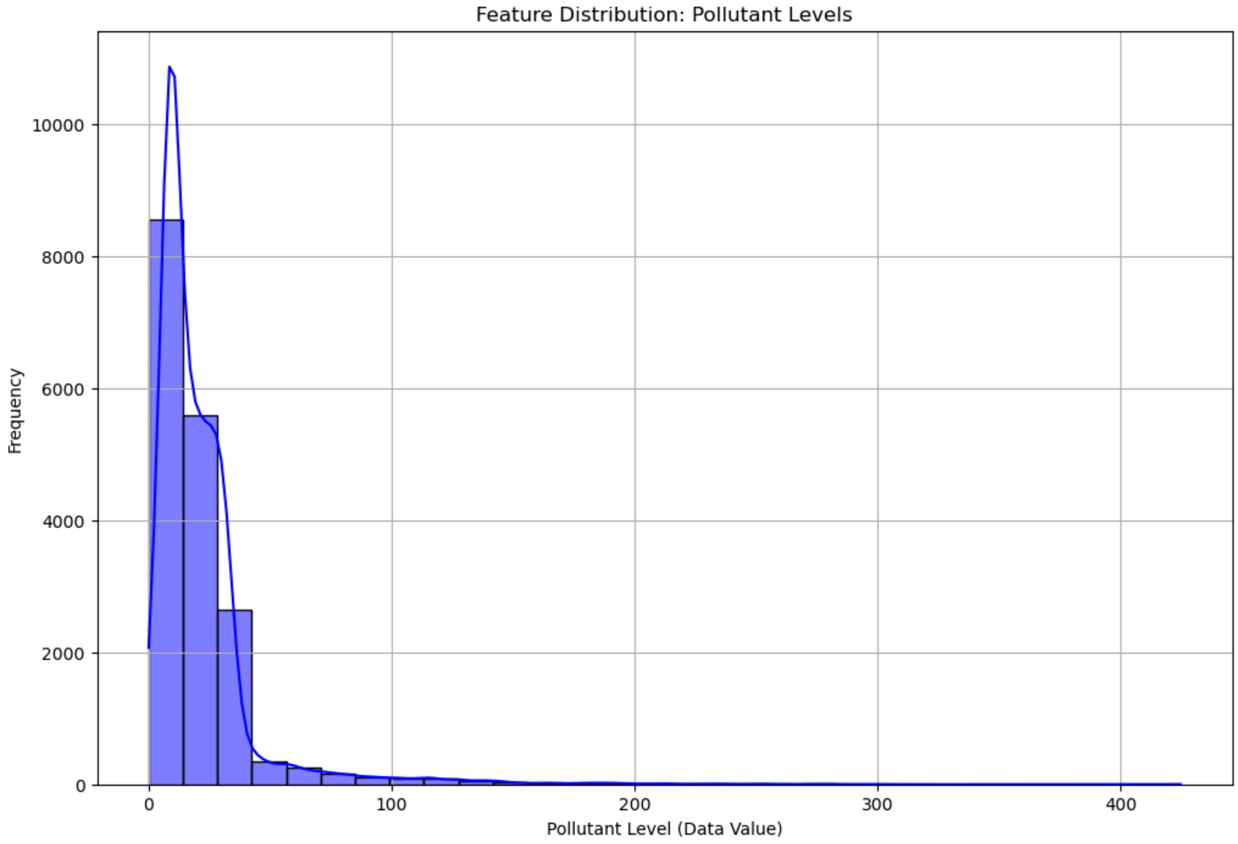

Figure 3: Feature Distribution

3.4 Data Challenges

- **Missing Values:** Certain time periods had incomplete data for specific pollutants or meteorological variables. These were addressed through imputation techniques such as mean or median substitution.
- **Multicollinearity:** Strong correlations between pollutants (e.g., PM2.5 and NO2) necessitated careful handling during modeling to avoid redundant information.
- **Outliers:** Extreme values were identified in pollutant levels and were addressed using statistical thresholds to ensure they did not disproportionately influence the model.

By addressing these challenges and performing an in-depth analysis, the dataset was properly cleaned for the preprocessing and modeling stages of the pipeline, which ensured more robust and reliable predictions of New York's AQI.

4 Method

4.1 Preprocessing

The preprocessing pipeline was designed to prepare the dataset for advanced machine learning algorithms while addressing challenges such as missing data, multicollinearity, and outliers. The data had missing values, leading to a need for a clean of the dataset in order to produce a model that could effectively handle the relationship of all of the variables.

Data Cleaning

- **Missing Values:** Missing entries were imputed using median values for numerical features to avoid skewing the dataset, particularly for pollutant concentrations and meteorological variables.
- **Outliers:** Extreme values were identified using the interquartile range (IQR) method and capped to ensure they did not disproportionately affect model training.

Feature Engineering

- **Temporal Features:** Extracted day, month, and seasonal indicators from the timestamp to capture temporal trends in air quality.
- **Interaction Terms:** Created interaction terms such as pollutant ratios (e.g., PM2.5 to NO2) to explore synergistic effects between variables.

Scaling

Numerical features were standardized to have zero mean and unit variance, ensuring compatibility with machine learning algorithms sensitive to feature scaling.

Data Splitting

The dataset was split into training (80%) and testing (20%) subsets, maintaining temporal continuity to reflect real-world forecasting scenarios.

4.2 Model Selection

Two machine learning algorithms, Random Forest Regression, and Gradient Boosting, were selected as the best fit for the problem type, given their ability to model non-linear relationships and handle complex interactions between features. These models' significant strengths are ideal for sifting through the intricate relationships and detecting them rather than other models that may try to aggregate the noise in the dataset to incorrectly assign relationships between variables.

Random Forest Regression

Random Forest Regression is an ensemble learning method that is a combination of multiple decision trees to improve prediction accuracy and robustness. Each of the trees in the forest is built by using a small subset of the training data, which is aggregated, and then the final prediction is then made by averaging the predictions from all of the trees:

$$\hat{y} = \frac{1}{T} \sum_{t=1}^{T} f_t(x)$$

Where $\hat{y}$ is the predicted value, T is the total number of trees and $f_t(x)$ is the prediction from the t-th tree. The trees are constructed using a splitting criterion such as Mean Square Error (MSE), defined as:

$$MSE = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

Where N is the number of samples, y_i is the actual value, and $\hat{y}_i$ is the predicted value. To properly assess feature importance, the Mean Decrease Impurity (MDI) is calculated by:

$$MDI(i) = \frac{1}{N_t} \sum_{t=1}^{T} \sum_{j \in nodes(t) : v(j)=i} p(j)\Delta i(t)$$

Where N_t is the total number of trees, $p(j)$ is the proportion of samples reaching node j, and $\Delta i(t)$ is the impurity decrease for feature i at node j.

Gradient Boosting

Gradient Boosting builds models sequentially by optimizing for errors made by previous models. This ensures each subsequent model attempts to minimize the residual errors.

$$Residual_i = y_i - \hat{y}_i$$

The optimization process involves minimizing a loss function *L*, typically Mean Squared Error (MSE), using gradient descent. The updated prediction at iteration $t + 1$ is given by:

$$\hat{y}_i^{(t+1)} = \hat{y}_i^{(t)} + \eta \cdot h(x_i)$$

Where η is the learning rate, controlling the contribution of each new model, $h(x_i)$ is the new model fit onto the residuals, and $\hat{y}_i^{(t)}$ is the prediction at iteration t. The calculation of the gradient minimizes the loss function *L*:

$$\nabla L = \frac{\partial L}{\partial \hat{y}_i}$$

Regularization techniques such as limiting tree depth, shrinkage (reducing η), and early stopping are employed to prevent overfitting and ensure strong generalization.

4.3 Experimental Setup

- **Software and Tools:** Python libraries such as Scikit-learn were utilized to implement the models.
- **Evaluation Metrics:** Metrics included R², Mean Squared Error (MSE), and Mean Absolute

Error (MAE) to assess model performance comprehensively.
- **Cross-Validation:** A 5-fold cross-validation approach was employed on the training data to ensure robustness and reduce variance.

5 Results

5.1 Model Performance

The models were evaluated on the testing dataset to compare their predictive capabilities.

Random Forest Regression

- **R^2 Score:** 0.92, which indicates a strong correlation between the predicted and the actual AQI values.
- **MSE:** 43.46, showing moderate prediction errors.
- **MAE:** 2.43, indicating the average error magnitude.

Gradient boosting

- **R^2 Score:** 0.69, underperforming the Random Forest in capturing the variability of AQI.
- **MSE:** 167.11, reflecting a large deviation in predictions
- **MAE:** 5.55, demonstrating reduced accuracy in individual predictions.

5.2 Feature Importance Analysis

Both models identified the Indicator ID and Geo Join ID as the most influential features. Random Forest demonstrated a more nuanced understanding of the feature's contributions compared to Gradient boosting.

5.3 Observations and Challenges

Random Forest significantly outperformed Gradient Boosting in terms of accuracy and error. This signifies that the robustness of Random Forest was strong in modeling non-linear relationships in AQI data. The reduced performance of Gradient Boosting could be due to suboptimal hyperparameters or the algorithm's high sensitivity to imbalance feature scaling and residual variance. Random Forest's ensemble nature effectively avoided overfitting, while Gradient Boosting might have been more prone to the issue. This would allow the Random Forest to be able to filter out the noise in the dataset much more efficiently.

6 Conclusion

6.1 Summary

This study demonstrated the efficacy of Random Forest Regression and Gradient Boosting in predicting AQI levels in New York. Random Forest emerges as the superior model, achieving a higher R^2 Score demonstrating improved predictive accuracy and lower MSE and MAE metrics compared to Gradient Boosting. The results show the critical role of geographical location on air pollution, which is likely due to those regions having elevated pollution levels. These elevated areas may need to consider ways they can reduce their environmental impact from an individual level to a corporation level.

6.2 Lessons Learned

The Random Forest Regression's ability to handle non-linear relationships and ensemble predictions proved significantly more effective than Gradient Boosting for the dataset. Proper handling of the missing data, feature scaling, and engineering integration terms were important for improving model performance. Models that can effectively determine important features of the dataset are important for creating an algorithm that has low sensitivity to noise that may be stored in the data. The lower performance of Gradient Boosting highlights the importance of tuning the hyperparameters and careful configuration to avoid underperformance. Future models must continue to work with AQI datasets to determine which models are most effective and how they should set the hyperparameters of these models. Improving the hyperparameters of these datasets is key to creating models that are able to decipher the complex relationships that variables in AQI datasets have.

6.3 Future work

Incorporate real-time data streams to enable dynamic and adaptive AQI forecasting for urban planning and health advisories. Exploring techniques like neural networks and deep learning could further improve the accuracy, especially in the area of temporal patterns. Investigating additional environmental factors to refine the feature selection could enhance the model's robustness. Leverage model insights to assist policymakers in crafting data-driven regulations to reduce urban air pollution and improve public health. Future research should collect data on the areas that show higher levels of pollution to pinpoint the cause of the worse AQI scores. For example, if there are many factories in the area outputting high levels of pollution or if the collective carbon footprint of each resident is high compared to the areas with better AQI scores.

7 Acknowledgements

I would like to thank the developers and contributors of the Scikit-learn library for providing a tool that allows a shorter pipeline for machine learning and data analysis through prebuilt structures. Scikit-learn's set of features, documentation, and straightforward approach helped me create a successful implementation and an intensive analysis of the models. The open-source library played a crucial role in enabling efficient data processing, training, and evaluation.

Works Cited

Aniceto, Maisa Cardoso, Flavio Barboza, and Herbert Kimura. "Machine learning predictivity applied to consumer creditworthiness." *Future Business Journal* 6.1 (2020): 37.

Bhatore, Siddharth, Lalit Mohan, and Y. Raghu Reddy. "Machine learning techniques for credit risk evaluation: a systematic literature review." *Journal of Banking and Financial Technology* 4.1 (2020): 111-138.

Li, Chenchen, Yan Li, and Yubin Bao. "Research on air quality prediction based on machine learning." *2021 2nd International Conference on Intelligent Computing and Human-Computer Interaction (ICHCI)*. IEEE, 2021.

Pawanekar, Swara Sameer, et al. "Efficient AQI Prediction: A Comparative Study of Artificial Neural Networks, LSTM, Random Forest, and Gradient Boosting Techniques." *2024 8th International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC)*. IEEE, 2024.

Qian, Hongyi, et al. "A comparative study on machine learning models combining with outlier detection and balanced sampling methods for credit scoring." *arXiv preprint arXiv:2112.13196* (2021).