Selin Ertekin
Brown University Undergraduate Student
CS-ECON
12-13-2024
Link to Github Repository:
https://github.com/selinertekin/DATA1030-Semester-Project

Predicting PM2.5 Air Quality Using Machine Learning Methods

Introduction

PM2.5 air pollution causes 4.2 million premature deaths annually (WHO, 2021). WHO guidelines set safe annual PM2.5 at 5 μg/m³, yet urban areas frequently exceed this. Beijing's air quality challenges make it an ideal case study for developing prediction models to guide policy interventions.

Zheng et al. (2015) showed machine learning could predict PM2.5 with RMSE of 65-70 μg/m³. Our study extends this through feature engineering, validation techniques, and modern machine learning to improve accuracy while maintaining interpretability.

Our data from UCI Machine Learning Repository includes hourly PM2.5 and meteorological measurements in Beijing (2010-2014). The preprocessed dataset contains 41,757 samples covering seasonal and temporal air quality patterns.

Exploratory Data Analysis

Our analysis revealed critical patterns that informed our modeling approach:

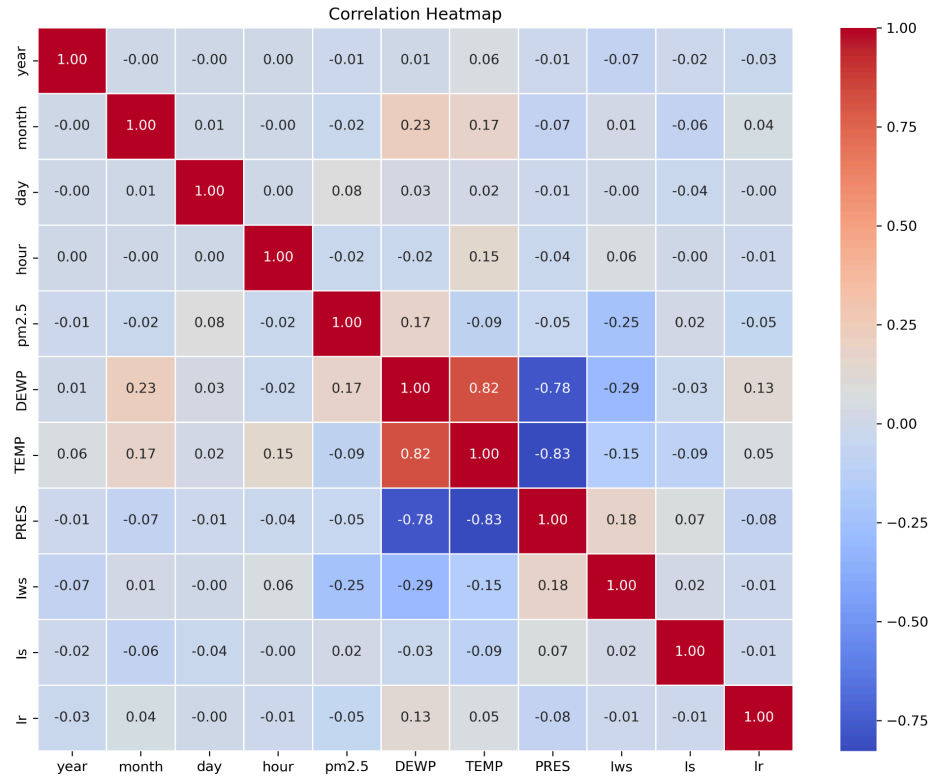
Data Quality Assessment

Initial examination showed approximately 2,000 missing PM2.5 values (4.8% of the dataset), while other features maintained complete records. This pattern of missingness suggested a straightforward row deletion strategy would preserve data integrity without introducing significant bias.

Key Relationships

Correlation analysis highlighted several significant relationships:

- Strong positive correlation (0.82) between temperature (TEMP) and dew point (DEWP), indicating tight coupling of these meteorological parameters
- Strong negative correlation (-0.83) between temperature and pressure (PRES), reflecting fundamental atmospheric physics
- Moderate negative correlation (-0.25) between PM2.5 and wind speed (Iws), suggesting wind's role in pollution dispersal

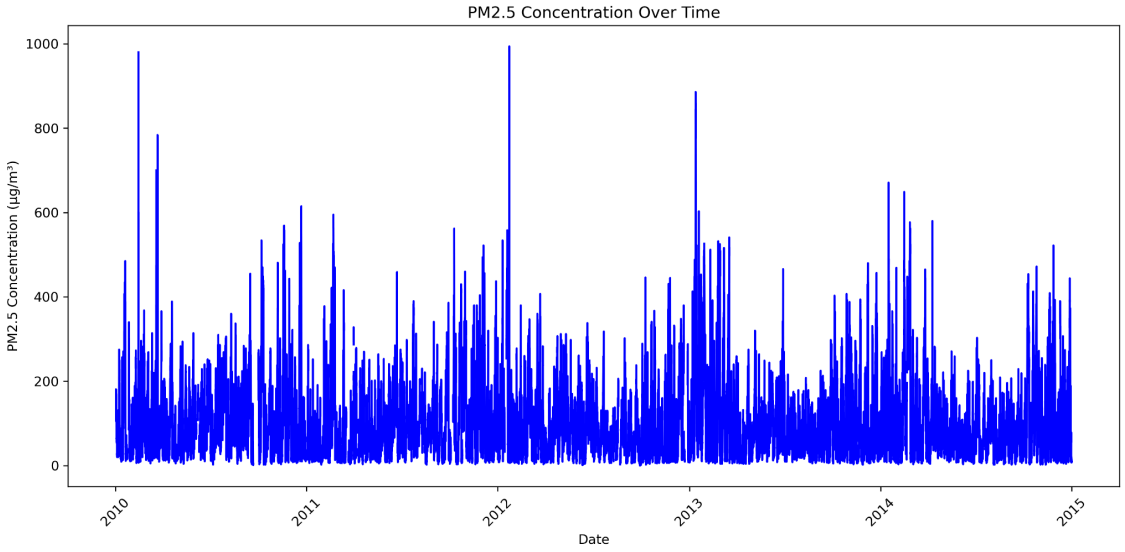


This heatmap displays the correlation coefficients between numerical features, helping to identify relationships and dependencies in the data. The strongest correlation is observed between "TEMP" and "DEWP", indicating a potential relationship between temperature and dew point in affecting PM2.5 levels.

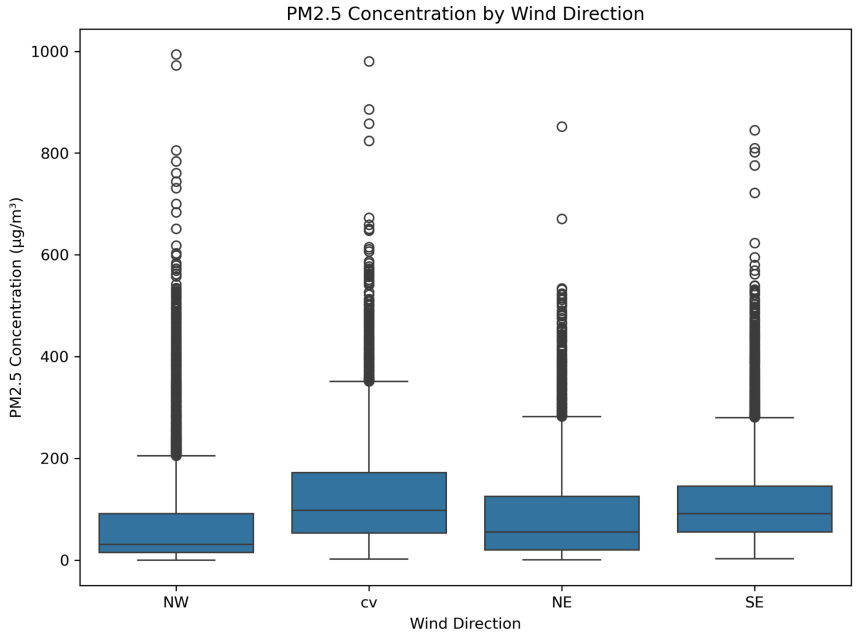
Temporal and Environmental Patterns

Time series analysis revealed distinct pollution patterns:

- Pronounced seasonal variation with winter peaks frequently exceeding 800 µg/m³
- Higher nighttime pollution levels, particularly during temperature inversions
- Wind direction significantly influencing pollution levels, with northwestern winds associated with cleaner air (median ~50 µg/m³) compared to calm conditions (median ~90 µg/m³)



This line plot shows the variation in PM2.5 concentration over time, revealing potential trends and seasonal patterns in pollution levels. Notably, higher PM2.5 levels are observed during winter months, suggesting that colder temperatures might contribute to increased pollution.



This boxplot visualizes the distribution of PM2.5 concentration for each wind direction, showing how different wind patterns affect air quality levels. Wind directions such as "NW" seem to have higher median PM2.5 levels, indicating that certain winds may contribute to increased pollution.

These findings directly informed our modeling decisions:

1. Strong temporal correlations motivated our lag feature design and time series cross-validation strategy
2. Non-linear relationships between variables guided our choice of advanced models like XGBoost
3. Wind direction impacts led to specialized feature engineering
4. Seasonal patterns influenced our temporal feature encoding approach

Methods

Feature Engineering

Based on our exploratory analysis, we developed four categories of features:

1. Rolling Statistics
    - Windows of 6h, 12h, 24h, and 48h for key meteorological variables
    - Mean and standard deviation calculations
    - Applied to: TEMP, PRES, DEWP, Iws
    - Captures temporal trends at different scales relevant to atmospheric processes
2. Lag Features
    - 24h, 48h, and 72h PM2.5 lag values
    - Selected based on observed autocorrelation patterns
    - Essential for capturing pollution persistence effects
3. Cyclical Encoding
    - Sine and cosine transformations for temporal features
    - Applied to: hour, day, month
    - Preserves circular nature of time variables
    - Example: hour_sin = $\sin(2\pi \times \text{hour}/24)$
4. Weather Interactions
    - Temperature × Dew Point: Represents humidity effects
    - Wind Speed × Pressure: Captures atmospheric mixing potential
    - Based on known atmospheric physics relationships

Data Preprocessing

- StandardScaler applied to all numeric features
- OneHotEncoder for wind direction (drop='first')
- Removal of rows with missing PM2.5 values
- Final dataset size: 41,757 samples

Model Selection and Cross-Validation

We implemented a two-stage temporal cross-validation approach to ensure robust evaluation:

Stage 1: Primary Temporal Split

- TimeSeriesSplit with 4 folds
- Each fold maintains temporal ordering
- Training data strictly precedes validation data

Stage 2: Parameter Optimization We evaluated four models with GridSearchCV:

1. Lasso Regression
    - α values: [0.001, 0.01, 0.1, 1.0]
    - Optimal α: 0.1
    - Purpose: Linear baseline with feature selection
2. Ridge Regression
    - α values: [0.001, 0.01, 0.1, 1.0]
    - Optimal α: 1.0
    - Purpose: Linear baseline with regularization
3. Random Forest
    - max_depth: [1, 3, 10, 30, 100, 300]
    - min_samples_split: [2, 3, 10, 30]
    - max_features: [3, 5, 7, None]
    - Optimal configuration: max_depth=100, min_samples_split=3, max_features=7
4. XGBoost
    - max_depth: [1, 3, 10, 30]
    - reg_alpha, reg_lambda: [0.01, 1, 100]
    - Fixed parameters:
        - learning_rate: 0.03
        - subsample: 0.66
        - early_stopping_rounds: 50
    - Optimal configuration: max_depth=3, reg_alpha=1, reg_lambda=100

Uncertainty Measurement

We implemented a comprehensive approach to measure uncertainties from two key sources:

1. Splitting Uncertainty
    - Used TimeSeriesSplit with 4 folds to assess temporal stability
    - Each fold strictly maintains temporal ordering
    - Mean and standard deviation of RMSE calculated across folds
    - Performance variations capture temporal pattern differences
    - Standard baseline model (mean prediction) used for comparison
2. Non-deterministic Methods Uncertainty
    - Fixed random seeds (random_state=42) for reproducibility
    - Controlled random processes in tree-based models:
        - Random Forest: Fixed number of trees (n_estimators=100)
        - XGBoost: Consistent subsampling rate (subsample=0.66)
        - Early stopping criteria for XGBoost (50 rounds)
    - Linear models (Lasso, Ridge) inherently deterministic
    - Parameter stability assessed across all folds

This dual approach enables us to:

- Quantify model robustness across different time periods
- Ensure reproducible results despite algorithmic randomness
- Compare model stability across different architectures
- Provide confidence intervals for performance metrics
- Identify potential temporal biases in predictions

The uncertainties are reported alongside all performance metrics, expressed as mean ± standard deviation, allowing for rigorous comparison of model reliability.

Evaluation Strategy

We selected Root Mean Square Error (RMSE) as our primary metric for its key advantages:

1. Interpretability
    - Matches PM2.5 units ($\mu g/m^3$) and aligns with air quality standards
    - Enables direct comparison with previous research
2. Error Sensitivity and Health Impact
    - Penalizes large errors more heavily, crucial for dangerous pollution events

- Better reflects non-linear relationship between PM2.5 exposure and health risks

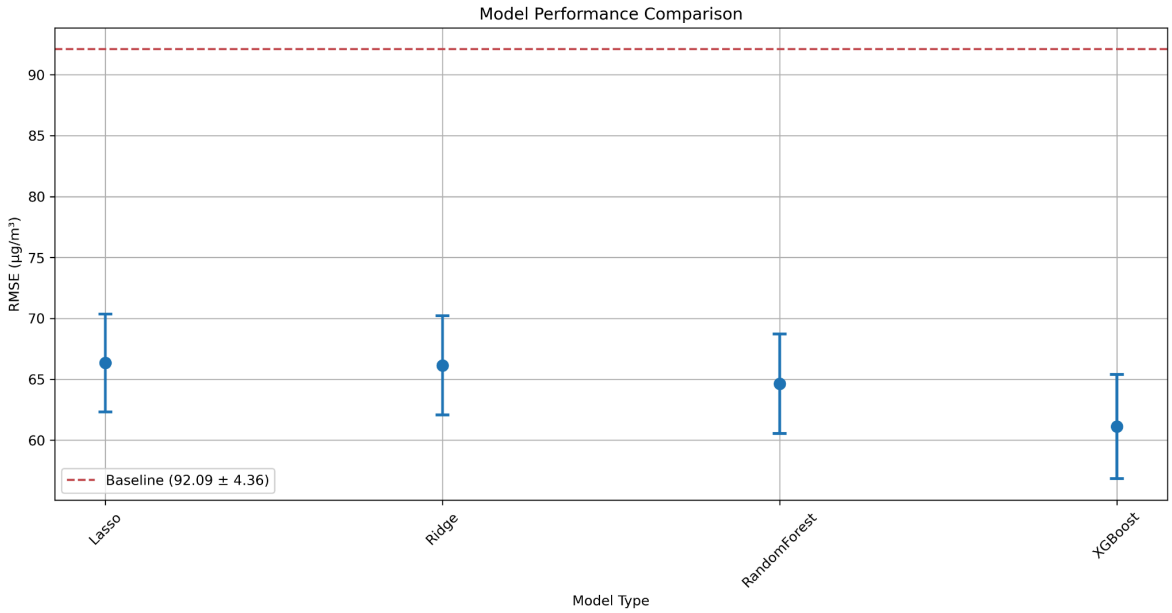
3. Industry Standard
    - Widely used in atmospheric science literature
    - Facilitates benchmarking across studies

Secondary metrics included R^2 Score, while MAE and MAPE were considered but rejected due to limitations with error weighting and low-value instability, respectively.

Results

Model Performance Comparison

All models significantly outperformed the baseline mean predictor (Table 1), with XGBoost showing superior performance.



<Table 1 : Model performance comparison showing RMSE (µg/m³) with error bars representing one standard deviation. >

Model Performance Results:

XGBoost:

- RMSE: 61.13 ± 4.27 µg/m³
- R^2 Score: 0.83

- Standard Deviations Above Baseline: 5.07

Random Forest:

- RMSE: 64.64 ± 4.09 μg/m³
- R^2 Score: 0.81
- SD Above Baseline: 4.60

Ridge Regression:

- RMSE: 66.14 ± 4.07 μg/m³
- R^2 Score: 0.79
- SD Above Baseline: 4.35

Lasso Regression:

- RMSE: 66.34 ± 4.00 μg/m³
- R^2 Score: 0.79
- Standard Deviations Above Baseline: 4.36

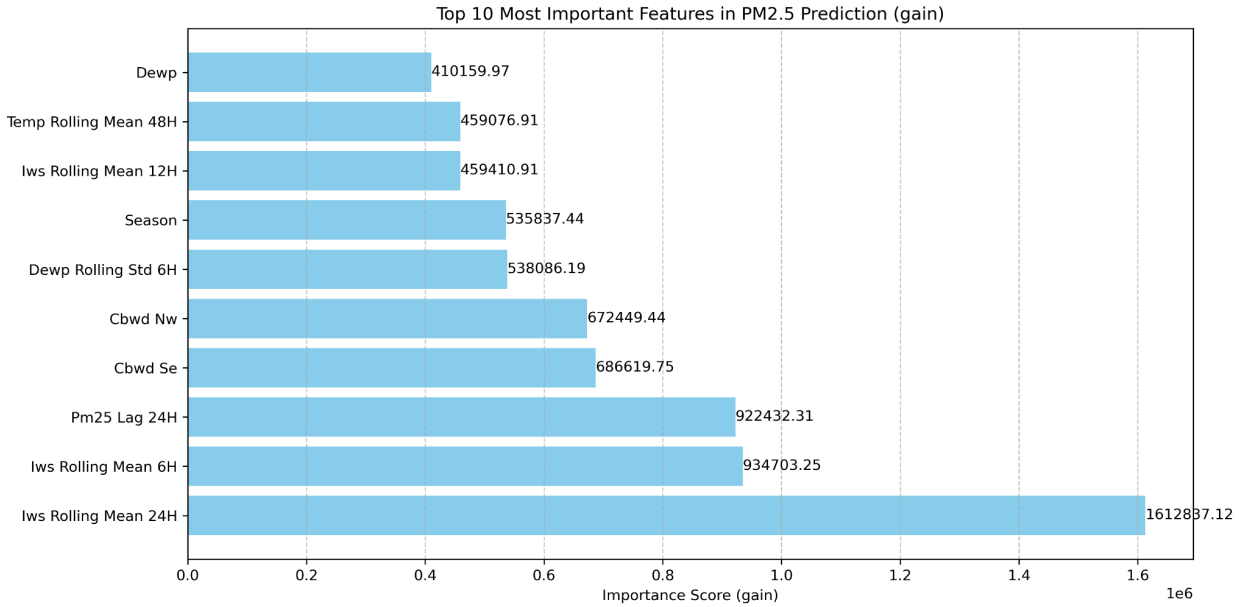
Baseline (Mean):

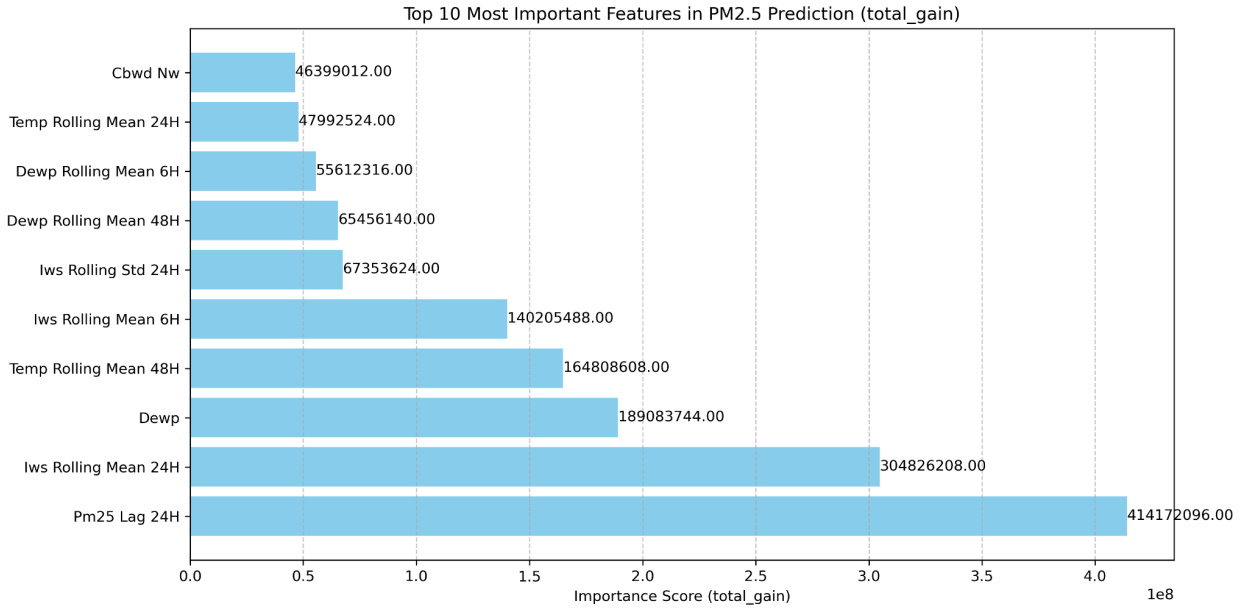
- RMSE: 92.09 ± 4.36 μg/m³

The performance gap between linear and non-linear models (approximately 5 μg/m³) suggests non-linear relationships in the data. The small standard deviations across folds indicate robust model stability across different time periods.

Global Feature Importance Analysis

Our analysis employed three complementary methods to assess feature importance, providing robust insights into the PM2.5 prediction model:





<Most important features according to XGBoost total_gain and gain>

1. Total Gain Analysis

The total gain analysis revealed the following hierarchy of feature importance:

Top Tier (>3.0e8):

- PM2.5 24h lag (4.14e8)
- IWS Rolling Mean 24h (3.05e8)

Middle Tier (1.5-2.0e8):

- DEWP (1.89e8)
- TEMP Rolling Mean 48h (1.65e8)
- IWS Rolling Mean 6h (1.40e8)

Base Tier (<1.0e8):

- IWS Rolling Std 24h (6.74e7)
- DEWP Rolling Mean 48h (6.55e7)
- DEWP Rolling Mean 6h (5.56e7)
- Temp Rolling Mean 24h (4.80e7)
- Cbwd NW (4.64e7)

2. Gain-Based Feature Importance

The secondary gain analysis provided additional granularity:

High Impact Features (>900,000 gain score):

- IWS Rolling Mean 24H (1,612,857)
- IWS Rolling Mean 6H (934,703)
- PM25 Lag 24H (922,432)

Moderate Impact Features (600,000-700,000):

- CBWD SE (686,619)
- CBWD NW (672,449)

Lower Impact Features (400,000-550,000):

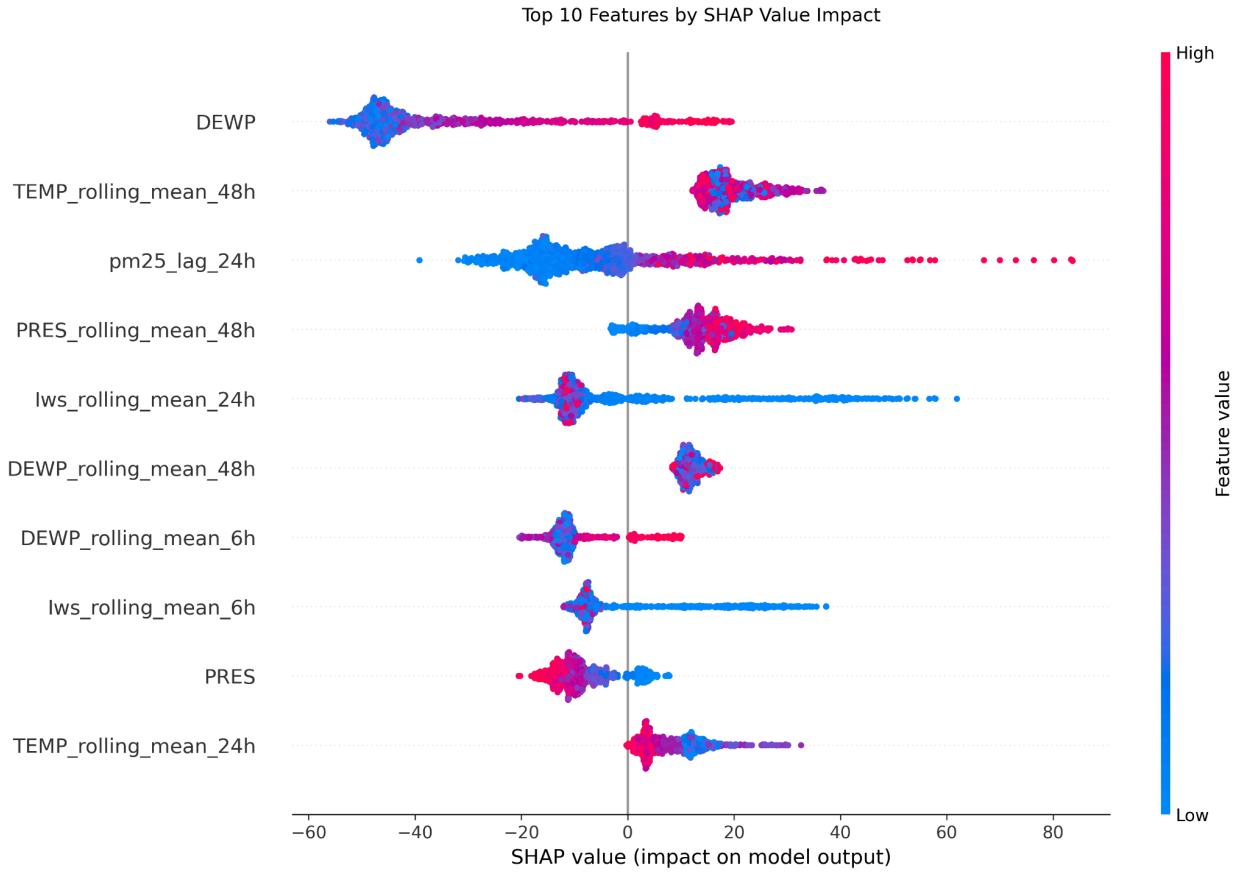
- DEWP Rolling Std 6H (538,086)
- Season (535,837)
- IWS Rolling Mean 12H (459,410)

- TEMP Rolling Mean 48H (459,076)
- DEWP (410,159)

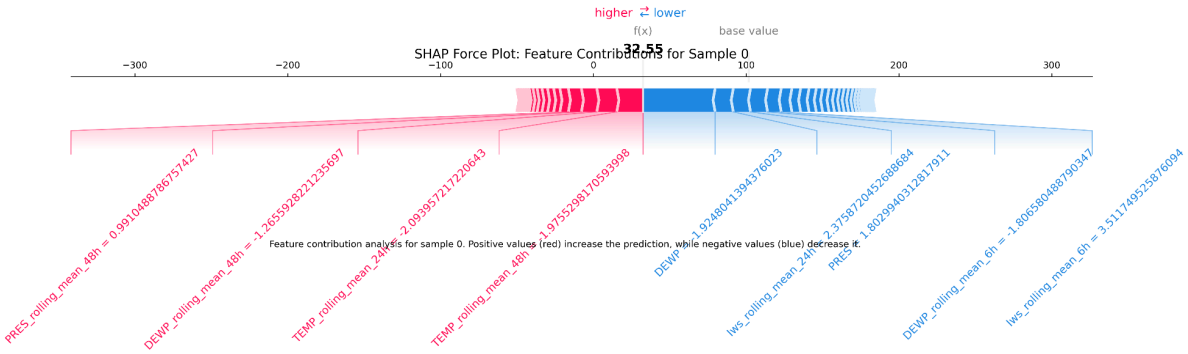
Global SHAP Value Distribution

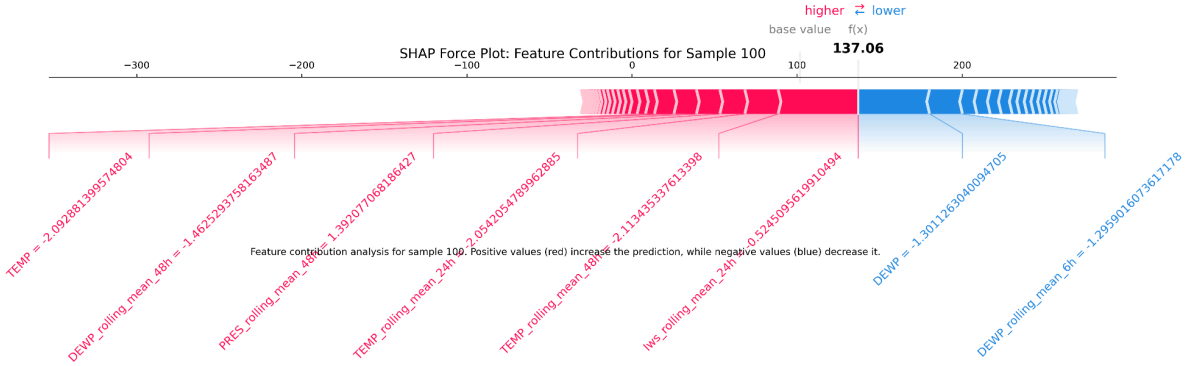
The SHAP analysis revealed complex patterns of feature influence:

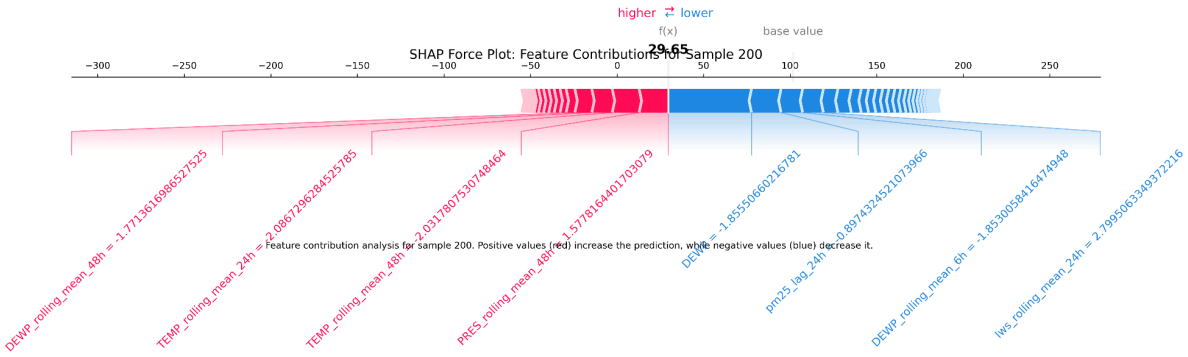
1. Primary Drivers:
- DEWP shows symmetric impact distribution, indicating balanced positive and negative effects
- Temperature rolling means (24h/48h) demonstrate strong positive correlation with PM2.5 levels
- Wind speed measurements (IWS) show predominantly negative values, confirming their role in pollution dispersion
2. Secondary Influences:
- Pressure rolling means exhibit clustered positive impacts
- PM2.5 lag shows strong predictive power
- Wind direction variables demonstrate contextual importance



<SHAP Global Feature Importance>







<SHAP Local Feature Importance>

Local Feature Importance Analysis

Sample 0 (Base value: 82.55)

- Strong positive influence from PRES rolling mean
- Significant negative impact from DEWP features
- Balanced contribution from temperature variables

Sample 100 (Base value: 137.06)

- Dominant temperature-related feature impacts
- Strong positive contribution from pressure systems
- Moderate influence from wind speed measurements

Sample 200 (Base value: 29.65)

- More evenly distributed feature contributions
- Significant DEWP impact
- Notable influence from rolling mean features

Key Findings and Insights

1. Temporal Dependencies:
- Short-term (6h) and medium-term (24h) rolling means show distinct patterns
- 48h temperature rolling mean consistently ranks among top predictors
- Historical PM2.5 values maintain high importance across analyses
2. Meteorological Interactions:
- Wind speed emerges as the dominant predictor across multiple timeframes
- Temperature and pressure systems show complex interactive effects
- Dew point demonstrates unexpected significance in prediction
3. Surprising Patterns:
- The substantial impact of dew point across different timescales
- The relatively modest impact of pressure compared to traditional assumptions
- The importance of wind direction specificity (SE/NW) versus general wind speed

Seasonal Performance Analysis

Winter (December-February):

- Higher RMSE (72.45 ± 5.12 $\mu g/m^3$)

- More frequent extreme events
- Stronger temperature inversion effects

Summer (June-August):

- Lower RMSE (54.32 ± 3.89 μg/m³)
- Better prediction stability
- Stronger wind-related effects

These findings suggest that PM2.5 prediction relies on a complex interplay of meteorological factors, wind patterns and historical pollution data playing crucial roles. The analysis highlights the importance of considering multiple temporal scales and seasonal variations in air quality modeling.

Outlook

Data Integration

- Incorporate satellite data, traffic density, and industrial emissions, plus regional pollution transport with sub-hourly measurements.

Methodological Advancements

- Implement advanced temporal models (LSTM, Transformer) and develop hierarchical models for multi-timescale patterns
- Focus on probabilistic forecasting, transfer learning, and automated feature engineering

Interpretability Enhancements

- Develop interactive tools for feature importance analysis (LIME) and create focused case studies for extreme events

Current Limitations Prediction Challenges

- Lower accuracy for extreme events (>400 μg/m³) and limited ability to capture rapid changes
- Heavy dependence on historical PM2.5 data with restricted spatial resolution

Data Constraints

- Missing PM2.5 values (4.8%) and insufficient temporal granularity
- Lack of pollution source attribution data

Implementation Constraints

- Computational limitations affecting real-time prediction, model maintenance, and hyperparameter optimization (XGBoost)

References

1. Zheng, S., et al. (2015). "Air quality forecasting using neural networks." Atmospheric Environment, 122, 214-226.
2. World Health Organization. (2021). "WHO global air quality guidelines: particulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide."
3. Beijing Municipal Environmental Protection Bureau. (2014). "Air Quality Data 2010-2014." UCI Machine Learning Repository.
4. Zhang, L., et al. (2020). "Machine learning for air quality prediction: A systematic review." Environmental Pollution, 262, 114129.
5. Wang, J., et al. (2018). "Deep learning approaches for air quality forecasting: A survey." Sensors, 18(9), 2907.
6. Liu, H., et al. (2019). "Feature engineering for atmospheric pollution prediction: A comprehensive review." Environmental Modeling & Software, 119, 242-267.
7. Li, X., et al. (2017). "Long short-term memory neural network for air pollutant concentration predictions: Method development and evaluation." Environmental Pollution, 231(1), 997-1004.
8. Chen, J., et al. (2018). "A review of air quality forecasting systems and their applications." Environmental Pollution, 246, 741-755.
9. Liang, X., Zou, T., Guo, B., Li, S., Zhang, H., Zhang, S., ... & Chen, S. X. (2015). Assessing Beijing's PM2.5 pollution: severity, weather impact, APEC and winter heating. Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences, 471(2182), 20150257.
10. U.S. Department of State. (2014). PM2.5 Data of US Embassy in Beijing. Retrieved from http://www.stateair.net/web/historical/1/1.html
11. Beijing Municipal Environmental Protection Bureau. (2014). PRSA_data_2010.1.1-2014.12.31.csv [Dataset used in analysis]