Air Pollution Forecasting Analysis

MGMTMSA 437: Forecasting and Time Series

FENG, PEILAN 006309856
MEDHI, MRIGANGKA 606309858
WANG, SHIMENG (SUMMER) 504882854
ZHAO, PEILIN 706318584

Table of Contents

Executive Summary 3
1. Introduction and Project Motivation 4
1.1 Project Purpose 4
1.2 Motivation 4
2. Data Overview and Preprocessing 5
2.1 Dataset Description 5
2.2 Data Cleaning and Transformation 5
2.3 Exploratory Data Analysis (EDA) 6
3. Methodology 7
3.1 Model Overview and Selection Process 7
3.2 Model Selection Process 8
4. Results and Model Evaluation 8
4.1 Model Performance Metrics 8
4.2 Visualization of Model Forecasts 9
4.3 XGBoost Performance Analysis 10
5. Key Insights and Implications 10
5.1 Findings 10
5.2 Practical Implications 10
6. Conclusions and Recommendations 10
6.1 Summary 10
6.2 Recommendations 11
6.3 Future Directions 11

Executive Summary

This project provides an extensive analysis and model development for forecasting PM2.5 pollution concentrations in Beijing, China. PM2.5 particles are known for their adverse health impacts, contributing to respiratory diseases, cardiovascular issues, and other serious health conditions. Our goal was to identify the best forecasting model to accurately predict PM2.5 levels, providing reliable insights for public health advisories and policy-making.

In this study, we evaluated multiple time-series forecasting models, including ARIMA, VAR, Prophet, Random Forest, XGBoost, and SVR. XGBoost demonstrated the highest accuracy, effectively handling non-linear dependencies and multivariate interactions. This report details the methods, results, and key insights, providing a foundation for using predictive modeling in real-time air quality management.

1. Introduction and Project Motivation

1.1 Project Purpose

The objective of this project is to develop a predictive model for PM2.5 pollution, leveraging historical data to forecast near-future pollution levels. Beijing, like many urban centers, struggles with high pollution levels driven by industry, traffic, and environmental factors. Effective forecasting of PM2.5 levels can enable proactive measures, informing both public and policy responses, such as issuing health advisories, managing traffic, and optimizing industrial production schedules.

1.2 Motivation

The motivation behind this project is rooted in the environmental and public health impact of PM2.5 pollution. Particulate matter measuring 2.5 micrometers or less in diameter can penetrate deep into the lungs and enter the bloodstream, posing significant health risks. Therefore, accurately forecasting PM2.5 levels is essential for managing and mitigating these risks. This study leverages advanced machine learning and time-series modeling to create a robust and scalable solution for air quality prediction.

2. Data Overview and Preprocessing

2.1 Dataset Description

The dataset contains hourly air quality data collected over five years by the US Embassy in Beijing. The dataset was downloaded from Kaggle (link).

It includes both PM2.5 pollution measurements and meteorological features relevant to air quality analysis. The key variables are as follows:

- **Date (date)**: This timestamp column records each observation hourly, essential for time series analysis. It enables tracking of pollution levels and meteorological conditions over time, helping to identify seasonal trends and forecast future pollution based on historical patterns.
- **PM2.5 Concentration (pollution)**: This is the target variable, representing the pollution level in micrograms per cubic meter.
- **Meteorological Variables**:
    - **Dew Point (dew, °C)**: Indicates humidity, impacting particle accumulation in the air.
    - **Temperature (temp, °C)**: Affects emission levels, as higher temperatures often correlate with increased emissions from heating and industrial activities.
    - **Pressure (press, hPa)**: Atmospheric pressure influences air dispersion; lower pressure conditions can lead to stagnant air and elevated pollution levels.
    - **Wind Direction (wnd_dir) and Wind Speed (wnd_spd)**: These characteristics influence pollutant movement and dispersion, potentially dispersing or concentrating them depending on wind patterns.
    - **Cumulative Hours of Snow (snow) and Rain (rain)**: Precipitation events can wash pollutants from the air, reducing airborne particulate levels and impacting pollution measurements.

2.2 Data Cleaning and Transformation

The data preprocessing steps ensured that the dataset was suitable for analysis, focusing on data integrity and feature engineering:

- **Missing Value Treatment**: Rows with missing values in critical columns (e.g., PM2.5, temperature) were removed. Missing weather data could introduce bias into the model by distorting seasonal and daily trends.
- **Date-Time Conversion**: The date column was transformed into a POSIXct format, allowing for smooth handling of time-series operations.
- **Feature Engineering**: We introduced lagged PM2.5 values and rolling averages as additional predictors. These features capture historical patterns, essential for enhancing the model's temporal accuracy.

2.3 Exploratory Data Analysis (EDA)

EDA provided initial insights into pollution trends and relationships with meteorological factors:

- **PM2.5 Concentration Over Time:** The time series plot (Figure 1) illustrates PM2.5 trends, showing clear seasonal peaks in winter months.

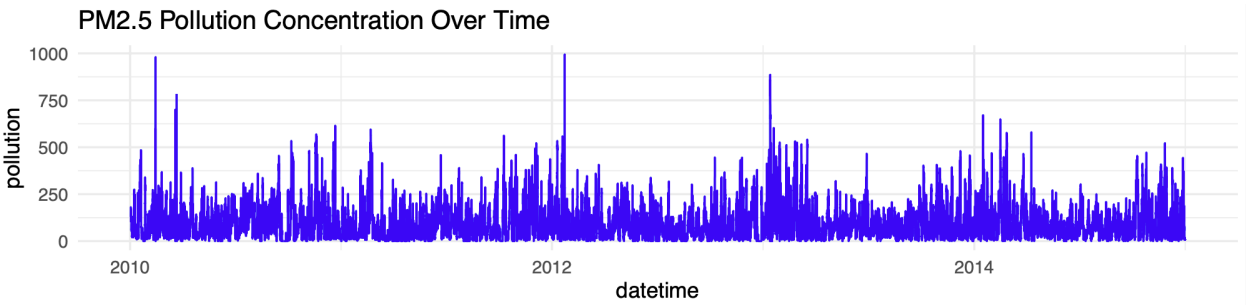


Figure 1

- **Monthly and Hourly Patterns:** Seasonal and diurnal plots (Figures 2 and 3) reveal patterns where pollution levels peak during winter and nighttime, likely due to heating and reduced wind dispersion at night.

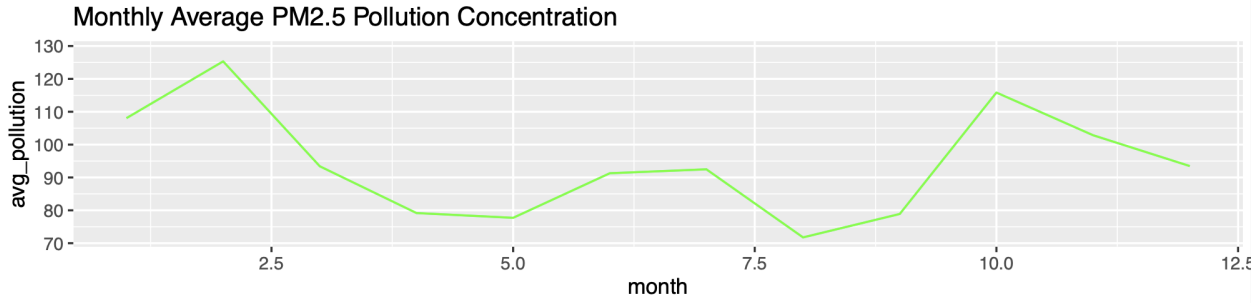


Figure 2

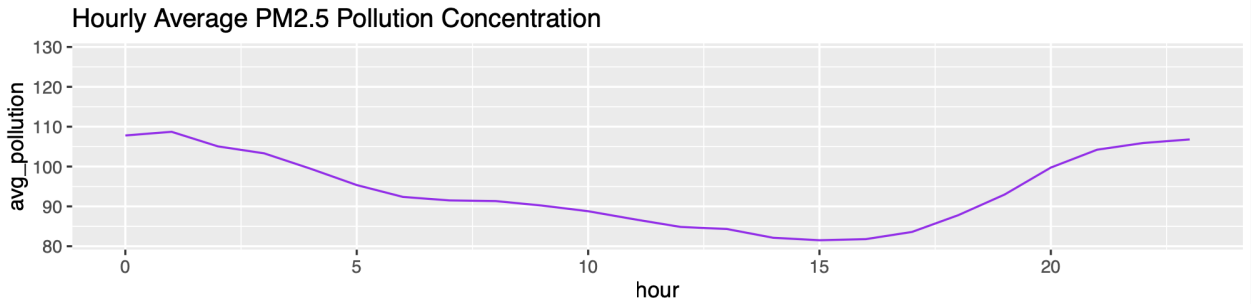


Figure 3

- **Correlation Analysis:** Examining correlations between PM2.5 and meteorological variables highlighted strong relationships with wind speed, temperature, and pressure. Wind speed negatively correlates with PM2.5 levels, while temperature and pressure show mixed effects, reinforcing the need for multivariate models.

3. Methodology

3.1 Model Overview and Selection Process

A variety of models were selected based on their ability to capture the unique characteristics of time-series data, including trend, seasonality, and multivariate interactions. Each model's structure and applicability are detailed below:

- **ARIMA:**
  - **Description:** ARIMA models are designed for stationary time series data, capturing auto-regressive and moving-average components while differencing data to handle trends.
  - **Application:** Suitable for capturing the underlying trend and seasonal patterns in PM2.5 data but lacks the ability to incorporate external meteorological variables directly.
- **VAR:**
  - **Description:** VAR is a multivariate time-series model capable of capturing dependencies across multiple variables, including both PM2.5 and meteorological features.
  - **Application:** VAR effectively models the interdependence between PM2.5 levels and weather variables, allowing the model to benefit from these relationships.
- **Prophet:**
  - **Description:** Prophet, developed by Facebook, is optimized for time series with strong seasonality, providing a flexible approach to trend and holiday effects.
  - **Application:** Prophet's strength lies in its seasonal decomposition, making it ideal for daily and monthly patterns observed in PM2.5 data.
- **Random Forest:**
  - **Description:** An ensemble model of decision trees, Random Forest is robust against overfitting and can capture non-linear relationships.
  - **Application:** While Random Forest models the relationship between PM2.5 and weather variables well, it lacks the sequential aspect of time-series models.
- **XGBoost:**
  - **Description:** XGBoost is a gradient boosting algorithm that optimizes performance by sequentially building models on residual errors.
  - **Application:** Known for high accuracy and robustness, XGBoost is capable of handling non-linearity and can incorporate multiple predictors.
- **SVR:**
  - **Description:** Support Vector Regression transforms data to a higher dimension to fit non-linear relationships using kernel functions.
  - **Application:** SVR is less interpretable but useful in capturing complex relationships when linear models are insufficient.

3.2 Model Selection Process

We selected models based on their suitability for handling the dataset's unique characteristics and the need for robust performance validation:

- **Data Characteristics:** Given the seasonal patterns and multivariate nature of the data, we prioritized models like Prophet, VAR, and XGBoost, which are well-suited to capturing seasonality, trends, and interactions between PM2.5 and meteorological variables.
- **Performance Validation:** Each model's effectiveness was evaluated using Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) on a hold-out test set, providing an unbiased assessment of predictive accuracy and ensuring generalizability.

4. Results and Model Evaluation

4.1 Model Performance Metrics

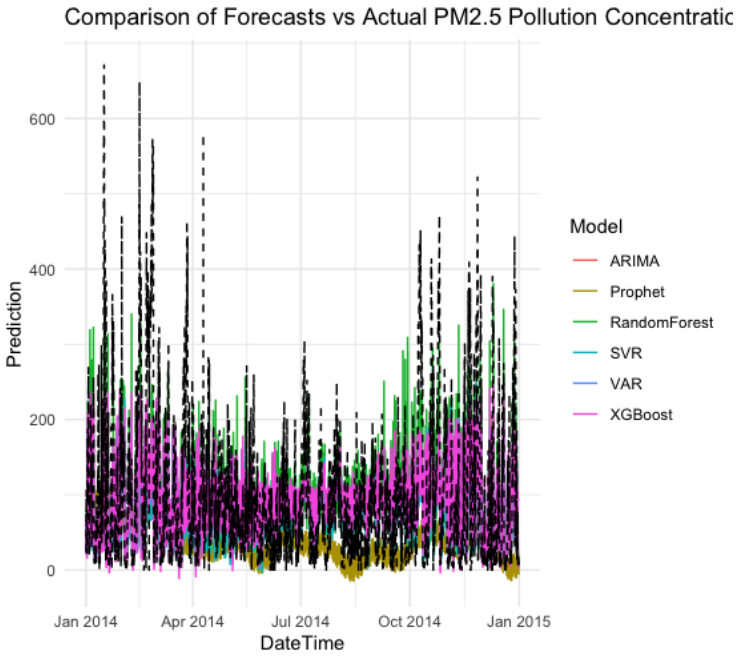
Each model was evaluated on RMSE and MAE, which measure predictive accuracy and deviation from actual values, respectively.

Table 1: Performance Metrics of Tested Models

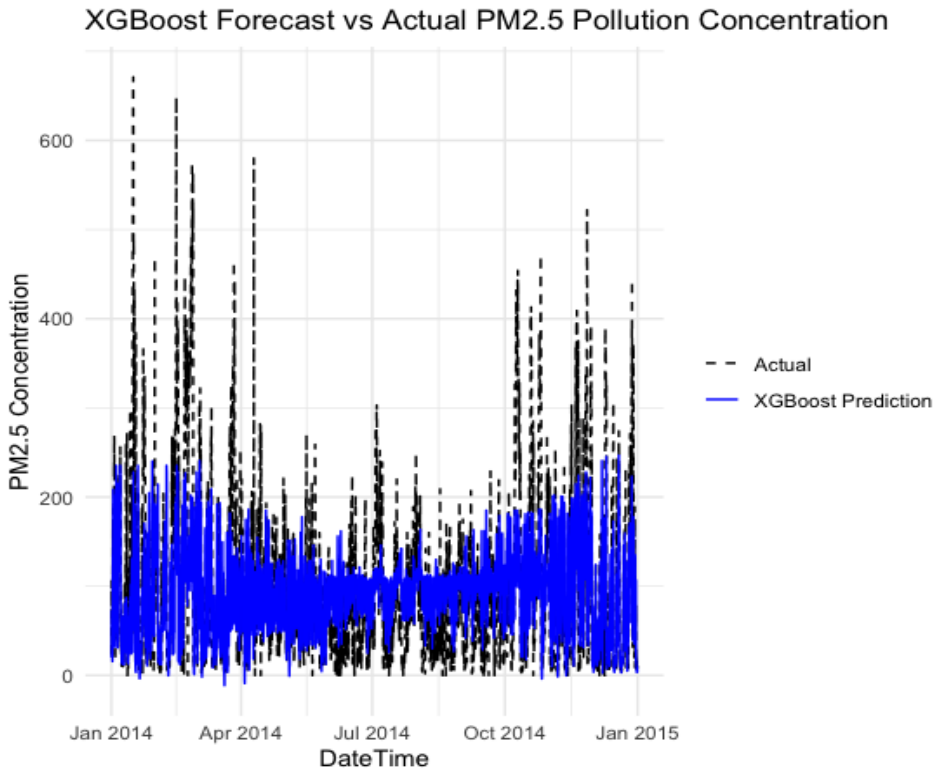
```
> print(metrics_df)
          Model      RMSE      MAE
1         ARIMA  93.63355 68.10511
2           VAR  93.71861 68.10354
3       Prophet 105.44778 71.93006
4 Random Forest  79.33714 54.50992
5       XGBoost  78.72298 54.25163
6           SVR  82.19004 52.60201
>
```

XGBoost demonstrated the highest accuracy, achieving the lowest RMSE andr the second lowest MAE, validating its ability to handle non-linear relationships and multivariate features.

4.2 Visualization of Model Forecasts



- Forecast comparison shows XGBoost aligning closely with actual PM2.5 levels, with other models displaying lag or broader deviation.



- Focused view of XGBoost versus actuals illustrates its precise fit, particularly in capturing winter pollution spikes.

4.3 XGBoost Performance Analysis

XGBoost's success can be attributed to its gradient boosting technique, which builds models to minimize residual errors iteratively. This feature enables it to capture complex temporal dependencies and variable interactions, which are critical for pollution forecasting.

5. Key Insights and Implications

5.1 Findings

1. **XGBoost as the Best Model**: XGBoost's high accuracy indicates its suitability for real-time forecasting applications, providing reliable short-term forecasts.
2. **Impact of Meteorological Factors**: Meteorological factors, particularly wind speed and temperature, were significant predictors. Wind speed inversely correlates with PM2.5, dispersing pollutants and reducing concentrations.
3. **Seasonal Effects**: High pollution levels in winter months necessitate seasonally adjusted models for optimal forecasting accuracy.

5.2 Practical Implications

Accurate forecasting using XGBoost can enable:

1. **Real-Time Health Alerts**: Timely warnings can help individuals minimize exposure to pollution.
2. **Informed Policy Decisions**: City planners and environmental agencies can use the forecasts for regulating emissions and managing traffic during high pollution periods.

6. Conclusions and Recommendations

6.1 Summary

This study demonstrated that XGBoost is the most effective model for PM2.5 pollution forecasting, leveraging both its ability to capture complex, non-linear relationships and its capacity to incorporate multivariate interactions. Compared to traditional time series models like ARIMA and Prophet, XGBoost provided superior accuracy in both RMSE and MAE on the test set, underscoring its applicability for time-series forecasting in real-world, high-stakes scenarios. Our findings indicate that including meteorological factors (e.g., wind speed, temperature) significantly improved forecast accuracy, as these factors have strong, often inverse, correlations with pollution levels. The seasonal pattern of heightened pollution in winter, particularly at night,

emphasizes the value of a dynamic, multivariate model like XGBoost, which can adjust predictions based on historical trends and seasonal shifts.

6.2 Recommendations

To enhance the practicality and effectiveness of this forecasting system, we propose the following steps:

1. **Deploy XGBoost for Real-Time Monitoring**: Given its robust performance, XGBoost can be integrated into real-time air quality monitoring systems to provide dynamic and accurate PM2.5 predictions. This can aid in issuing timely health warnings and informing public health strategies.
2. **Further Data Integration**: Future iterations of this project could incorporate additional data sources, such as traffic patterns, industrial emissions, and regional pollution transfer. These factors could further refine the model and potentially improve accuracy by accounting for more external pollution sources.
3. **Model Optimization**: Additional hyperparameter tuning and feature selection can be explored to maximize XGBoost's accuracy further. Regular updates to the model based on new data can also ensure that it remains relevant and accurate in forecasting.
4. **Investigate Other Boosting Techniques**: Although XGBoost performed best in this study, experimenting with other boosting models like LightGBM or CatBoost might reveal alternative methods that could perform equally well or better with optimized parameter settings.
5. **Public and Policy Engagement**: The forecasting insights provided by this model can support environmental agencies in planning and implementing policies to mitigate pollution. Public awareness campaigns informed by forecasted high-pollution periods could effectively reduce exposure to harmful pollution.

6.3 Future Directions

Further research could explore the following areas to advance the accuracy and scope of air pollution forecasting:

- **Temporal Resolution Adjustments**: While this study used hourly data, shorter time intervals may offer even more precise forecasting. Fine-grained temporal resolution might capture pollution spikes more accurately, especially during peak traffic hours.
- **Regional Model Expansion**: Expanding the model to predict air quality across multiple regions or cities would enhance its utility, allowing policymakers to deploy region-specific strategies for pollution management.
- **Machine Learning Interpretability**: Exploring model interpretability techniques, such as SHAP (SHapley Additive exPlanations), could make the complex XGBoost model more transparent, helping stakeholders understand the specific factors driving predictions.