Joseph Damico
12/22/2024
Applied Data Science
DSC680-T303 (2253-1)

Final Project: Predicting AQI Levels Using Geographic and Administrative Features

Business Problem

Air pollution remains a critical public health challenge. Among pollutants, fine particulate matter (PM2.5) is especially concerning due to its adverse health effects, including respiratory and cardiovascular diseases. This project aims to predict PM2.5 levels using geographic and administrative data, enabling the audience to make informed decisions for pollution control and public safety.

Background/History

PM2.5 refers to particulate matter with a diameter of less than 2.5 micrometers. Due to its small size, PM2.5 can penetrate deeply into the respiratory system causing significant health risks. In urban and industrialized areas PM2.5 levels often exceed safe thresholds necessitating robust monitoring and predictive capabilities.

Data Explanation

The dataset consists of PM2.5 monitoring data for 2022–2024 sourced from the U.S. Environmental Protection Agency (EPA). The data includes metrics such as PM2.5 levels, monitoring locations, and administrative details. After filtering and cleaning the dataset contains 550 records with key features used for analysis and modeling.

Data Dictionary:

- State Code, County Code, Site Num: Administrative identifiers for monitoring locations.
- Latitude, Longitude: Geographic coordinates of monitoring sites.
- Arithmetic Mean: Average PM2.5 levels for a given period.
- Parameter Name: Indicates the pollutant measured (PM2.5).
- Method Name: Specifies the monitoring method used.

Methods

The project involved three main steps:

1. Data Filtering and Cleaning: Filtering the data for PM2.5 (Parameter Code 88101 based of the EPA pdf accompanying the data) and approved methods based on advisory recommendations.

2. Exploratory Data Analysis (EDA): Generating insights into trends, distributions, and geographic variability in PM2.5 levels.

3. Modeling: Using a Random Forest Regressor to predict PM2.5 levels based on geographic and administrative features.

Analysis

EDA revealed significant variability in PM2.5 levels across states and counties. The distribution of PM2.5 levels is relatively consistent with a mean of 7.29. Top counties and states with high PM2.5 levels were identified.

Joseph Damico
12/22/2024
Applied Data Science
DSC680-T303 (2253-1)

Key Visualizations:

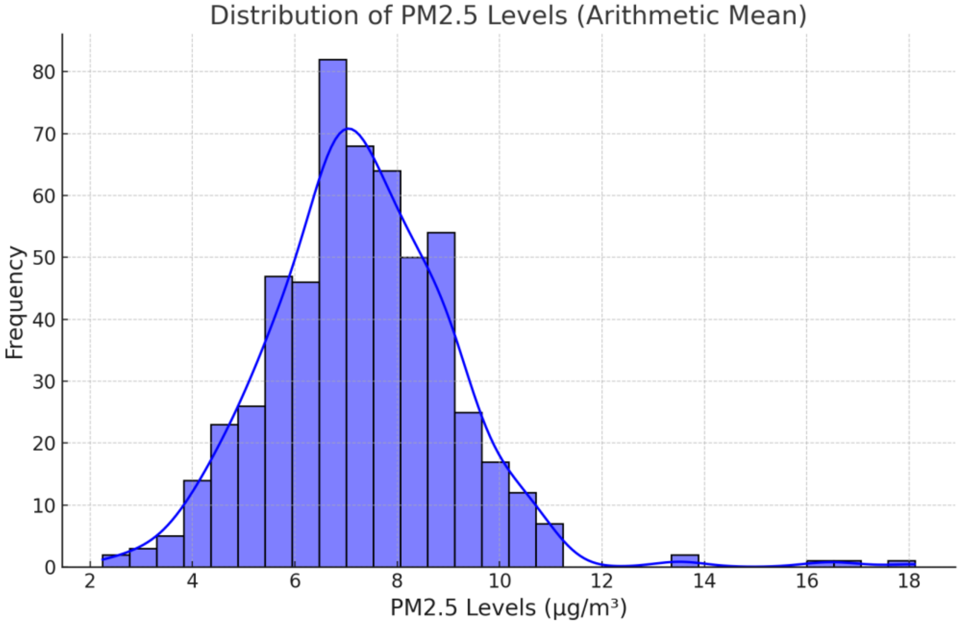

Figure 1: Distribution of PM2.5 Levels.

Joseph Damico
12/22/2024
Applied Data Science
DSC680-T303 (2253-1)

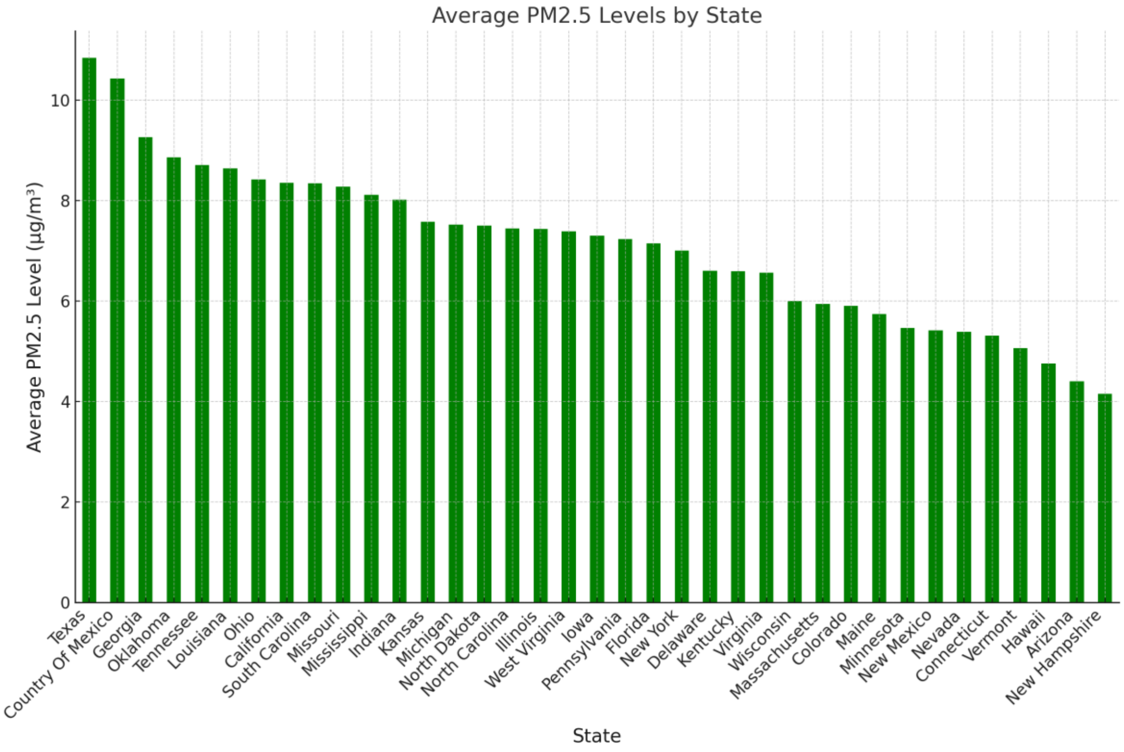

Figure 2: Average PM2.5 Levels by State.

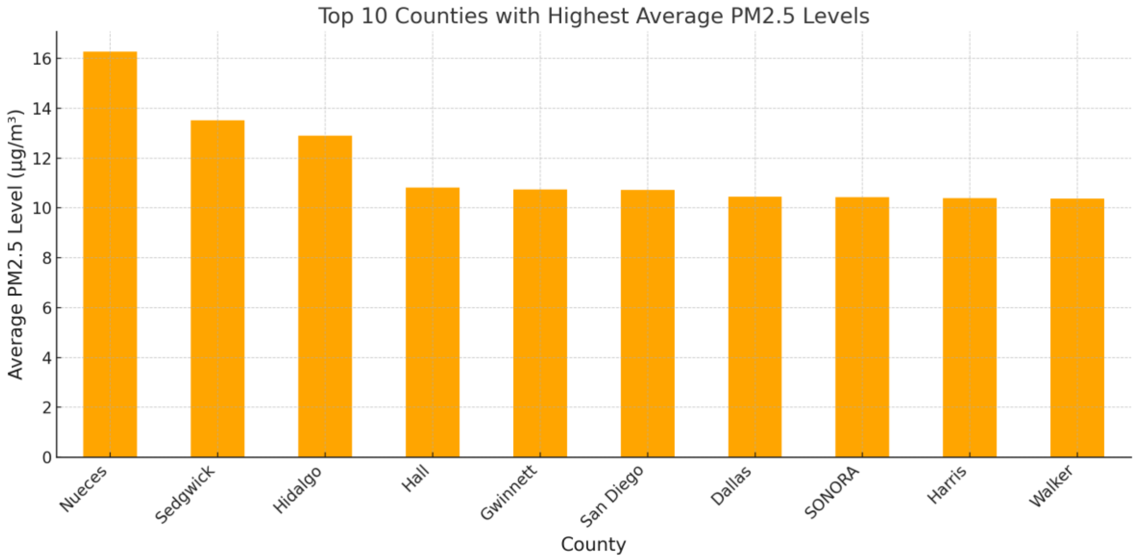

Figure 3: Top Counties with Highest Average PM2.5 Levels.

The Random Forest Regressor achieved the following metrics:

Root Mean Squared Error (RMSE): 1.01 µg/m³

R² Score: 0.58

Feature importance analysis revealed that geographic features (latitude and longitude) were the most significant predictors.

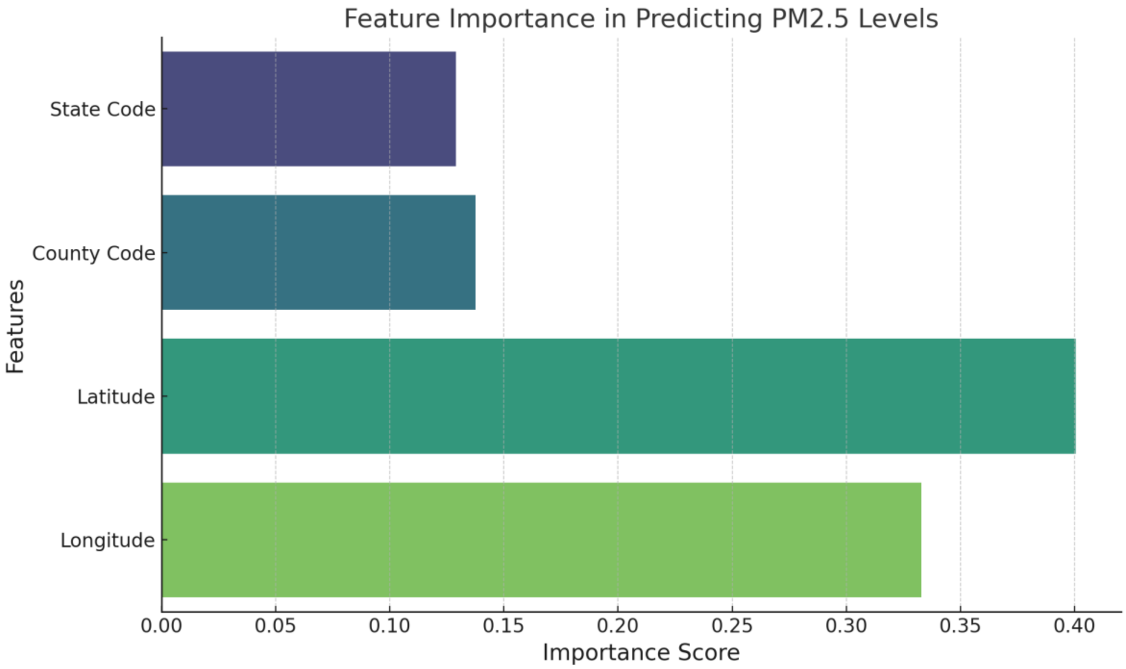

Figure 4: Feature Importance in Predicting PM2.5 Levels.

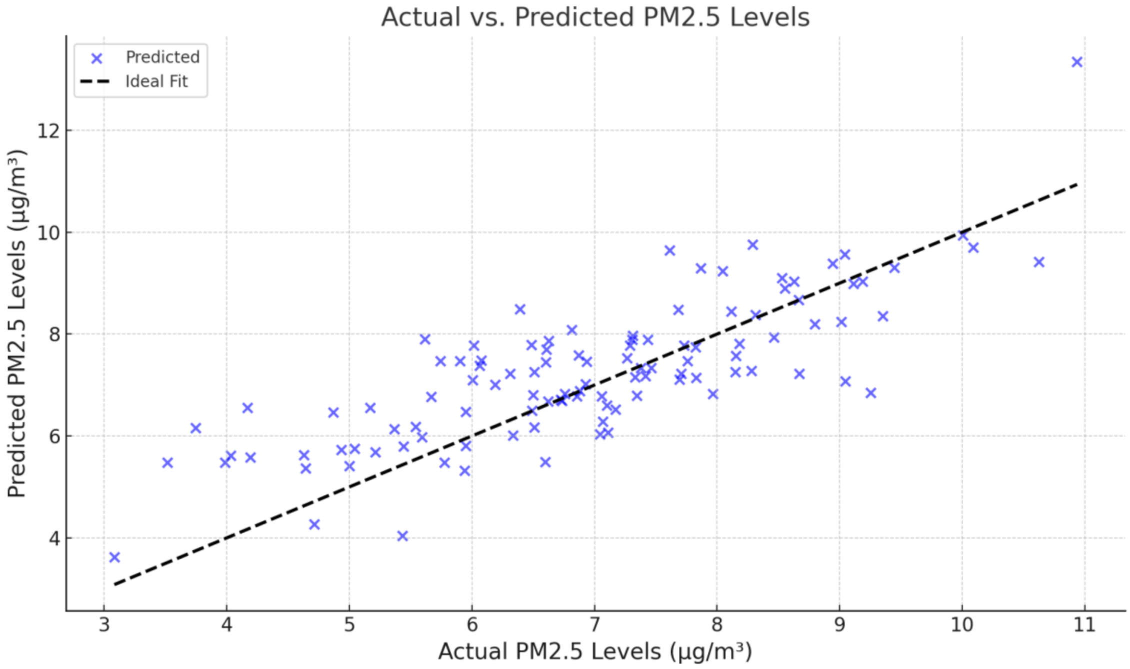

Figure 5: Actual vs. Predicted PM2.5 Levels.

Conclusion

The project demonstrated the viability of predicting PM2.5 levels using geographic and administrative features. Although the model achieved moderate predictive power the results highlight the importance of geographic variability in determining PM2.5 levels. I know this was a concern from the feedback on milestone 2 and it did prove to be difficult.

Assumptions

1. Geographic features (latitude and longitude) sufficiently capture spatial variability in PM2.5 levels.

2. Administrative codes (state and county) contribute meaningful information to predictions.
3. The filtered dataset is representative of real-world trends not just specific to the USA

Limitations

1. The model explains 58% of the variance in PM2.5 levels leaving room for improvement.
2. Sudden pollution events, such as wildfires, are not accounted for.
3. Data quality varies across regions potentially affecting results.

Challenges

1. Filtering the dataset based on advisory recommendations required manual verification of method names.
2. Geographic and administrative data alone may not fully capture the complexity of PM2.5 variability.
3. Ensuring consistency across multiple years of data presented integration challenges.

Recommendations

1. Incorporate additional features, such as industrial emissions and traffic data, to improve model accuracy.

2. Extend the analysis to include meteorological variables like temperature and wind speed.
3. Develop real-time prediction systems using streaming data for immediate applications.

Q&A

1. What motivated you to focus on PM2.5 for this project?
    a. PM2.5 is one of the most harmful air pollutants due to its ability to penetrate deeply into the respiratory system, causing severe health issues like asthma, lung cancer, and cardiovascular diseases. Its widespread impact, especially in urban and industrial areas, makes it a crucial target for data-driven analysis and prediction.
2. Why did you choose geographic and administrative features for prediction?
    a. Geographic features like latitude and longitude capture spatial variability, which is a key driver of PM2.5 levels. Administrative features such as state and county codes provide additional context such as regional regulations, industrial activity, and population density, which influence air quality.
3. How accurate was the model, and how can it be improved?
    a. The model achieved a Root Mean Squared Error (RMSE) of 1.01 and an R^2 score of 0.58, indicating moderate accuracy. Incorporating additional features, such as meteorological variables (e.g., temperature, wind

speed), industrial emissions, and traffic patterns, could improve the model's predictive power.

4. What were the biggest challenges you faced during the project?

    a. One major challenge was filtering the dataset based on advisory recommendations as this required detailed manual verification of method names. Another challenge was ensuring consistency across three years of data. Additionally, the absence of certain features like meteorological data limited the model's predictive accuracy.

5. Why did you use a Random Forest Regressor instead of another model?

    a. Random Forest is a robust non-linear model that handles complex interactions between features well. It also provides feature importance metrics which helped identify the most significant predictors of PM2.5 levels. Its flexibility and interpretability made it a good choice for this analysis.

6. What insights did you gain from the Exploratory Data Analysis (EDA)?

    a. EDA revealed that PM2.5 levels vary significantly across states and counties with some regions experiencing much higher averages. The distribution of PM2.5 levels showed a consistent clustering around the mean with a few outliers indicating pollution spikes. These insights highlight the importance of spatial variability in air quality management.

7. Why do latitude and longitude have higher importance than state or county codes?

    a. Latitude and longitude capture precise geographic locations which directly reflect environmental conditions influencing PM2.5 such as proximity to industrial areas, elevation, and wind patterns. State and county codes are broader administrative groupings that don’t capture localized variability as effectively.

8. How can this analysis be applied in real-world scenarios?

    a. This analysis can support policymakers in identifying high-risk areas and prioritizing interventions. Real-time predictions could be integrated into public dashboards to inform citizens about air quality risks. It could also guide urban planning and regulatory decisions to mitigate pollution in specific regions.

9. What were the limitations of the dataset and how did they affect the results?

    a. The dataset lacked features like meteorological variables and real-time data which are critical drivers of PM2.5 variability. Sudden pollution events such as wildfires or industrial accidents were not included. These limitations reduced the model's ability to fully explain the variance in PM2.5 levels.

10. What are the next steps for improving this project?

a. Future steps include incorporating additional data sources such as meteorological variables, traffic density, and industrial emissions, to improve predictive accuracy. Developing real-time prediction systems and integrating them into public health dashboards would enhance the project's impact and usability.

Joseph Damico
12/22/2024
Applied Data Science
DSC680-T303 (2253-1)

Appendix

Air Data: Air Quality Data Collected at Outdoor Monitors Across the US | US EPA. (2024, October 29). US EPA. https://www.epa.gov/outdoor-air-quality-data

Download Files | AirData | US EPA. (n.d.). https://aqs.epa.gov/aqsweb/airdata/download_files.html

National Oceanic and Atmospheric Administration. (n.d.). https://www.noaa.gov/

Home. (2024, November 28). https://www.who.int/