Model Card – Predicting Air Quality Index

Model Details

- Model Name: Random Forest Regressor for Air Quality Index Prediction
- Model Version: v1.0, This is the initial version of the model, trained and evaluated using the Beijing Multi-Site Air-Quality Dataset.
- Model Type: Regression, the model predicts continuous values of the Air Quality Index (AQI) based on environmental and pollutant data. It falls under the category of supervised regression models.
- Developers: The model was developed by the ENGM680 Project Team comprising Naveena, Manideep, Fouzia, Praharshitha under the guidance of Marcin Mizianty. The team worked collaboratively to implement, optimize, and validate the model using advanced preprocessing and evaluation techniques.
- Release Date: The model was finalized and validated on December 10, 2024, after comprehensive experimentation and hyperparameter tuning to ensure robust performance.

Intended Use

- Primary Use: The primary purpose of the model is to predict the Air Quality Index (AQI) in real-time using pollutant concentrations and meteorological data. By leveraging a Random Forest Regressor, the model provides accurate AQI predictions based on historical data and real-time measurements. This predictive system aims to Raise public awareness.
- Intended Users: The target audience for the model is General Public To make informed decisions about outdoor activities based on AQI forecasts.
- Out-of-Scope Use Cases: This model is not intended for Medical Diagnosis; the model does not predict or diagnose individual health conditions related to air pollution exposure.

Model/Data Description

- Data Used: The data used for training and evaluating the model is derived from the Beijing Multi-Site Air-Quality Dataset, sourced from publicly available air quality monitoring records. The dataset provides hourly air quality readings from March 2013 to February 2017 across 12 monitoring stations in Beijing. It includes both pollutant concentrations and meteorological variables, making it comprehensive for predicting Air Quality Index (AQI).
- Features: The input features used for the model are carefully selected based on their relevance to AQI prediction. These include both pollutant concentrations (PM2.5, PM10, SO2, NO2, CO, O3) and meteorological variables (TEMP, PRES, DEWP, WSPM, RAIN), which together provide a holistic view of air quality dynamics.
- Feature Importance: The Random Forest Regressor provides insights into feature importance, highlighting the most significant predictors. During training, it was observed that PM2.5 contributed the most to AQI prediction, followed by O3 and PM10 and Meteorological factors such as TEMP and WSPM had moderate importance, while RAIN had minimal impact.
- Model Architecture: The Random Forest Regressor was selected for its ability to handle non-linear relationships and interactions between features. Key characteristics of the model include, Ensemble Learning, Feature Robustness, Hypermeters Number of trees (n_estimators): 100 and Random seed (random_state): 42 for reproducibility.
- Why Random Forest? Accuracy demonstrated superior performance compared to baseline models like Linear Regression. Interpretability provides feature importance scores, aiding in understanding AQI drivers and Scalability: Handles large datasets effectively without significant preprocessing requirements.

Training and Evaluation

- Training Procedure: The Random Forest Regressor model was trained to predict the Air Quality Index (AQI) using the Beijing Multi-Site Air-Quality Dataset. The dataset was first split into training and testing sets (80% training, 20% testing) to ensure that the model was evaluated on unseen data. The following training parameters were used:
    - Number of Estimators: 100 trees
    - Random State: 42 (for reproducibility)
    - Training Process: The model was trained in the selected features (e.g., PM2.5, O3, PM10, NO2, CO, TEMP, DEWP) using the training set. The model was trained in the training data, and predictions were made on the test data to evaluate its generalization ability.
- Evaluation Metrics: These metrics were computed on the test dataset to evaluate how well the model predicts AQI values on unseen data. The performance of the Random Forest Regressor was evaluated using the following metrics:
    - Mean Absolute Error (MAE): Measures the average magnitude of errors in predictions. Lower values indicate better performance.
    - Root Mean Squared Error (RMSE): The square root of MSE, which gives a sense of the average magnitude of errors in the same units as AQI.
    - R-squared (R^2): Indicates how well the model explains the variance in the target variable (AQI). A higher R^2 (closer to 1) indicates a better fit.
- Baseline Comparison: To assess the performance of the Random Forest Regressor, it was compared with three baseline models. While Linear Regression served as a baseline model due to its computational efficiency, it failed to capture non-linear relationships, and though XGBoost demonstrated strong performance, Random Forest Regressor ultimately outperformed both models in terms of generalization and accuracy, particularly in predicting AQI values on the test set.

Ethical Considerations

- Fairness and Bias: The model may exhibit geographic or temporal biases due to the dataset's focus on Beijing, mitigated through cross-validation and feature standardization.
- Privacy: The dataset contains no personally identifiable information, ensuring compliance with privacy standards.
- Security: Model vulnerability is minimal as it operates on publicly available air quality data; however, secure storage and deployment practices are recommended to prevent tampering.

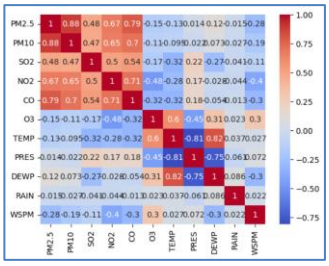
Limitations and Recommendations

- Limitations and Recommendations:
    - Data Limitations: The dataset is specific to Beijing and may not generalize well to regions with different air quality dynamics or monitoring practices.
    - Temporal Gaps: The model may underperform during rare or extreme pollution events not adequately represented in the dataset.
- Recommendations for Use:
    - Regional Adaptation: Retrain the model with local data when applying it to other regions to ensure accuracy.
    - Ethical Transparency: Clearly communicate model limitations and avoid using predictions as the sole basis for critical decisions.

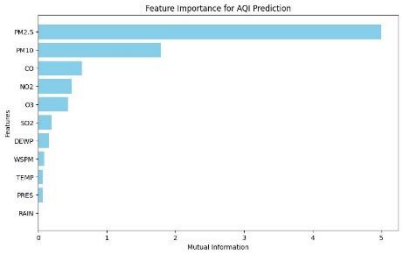
Additional Information

- References:
    - https://archive.ics.uci.edu/dataset/501/beijing+multi+site+air+quality+data
    - https://github.com/Afkerian/Beijing-Multi-Site-Air-Quality-Data-Data-Set/blob/main/README.md
    - https://www.epa.gov/system/files/documents/2024-02/pm-naaqs-air-quality-index-fact-sheet.pdf
    - https://paperswithcode.com/dataset/beijing-air-quality
    - https://www.kaggle.com/datasets/victorbonilla/beijing-multisite-airquality-data-data-set
- License: This model and its associated code are released under the WATCH team License, allowing free use, modification, and distribution with proper attribution.
- Contact Information: For further information or support, contact the development team at magesbab@ualberta.ca, fkhazi@ualberta.ca, dupakunt@ualberta.ca, and duvvuri@ualberta.ca.

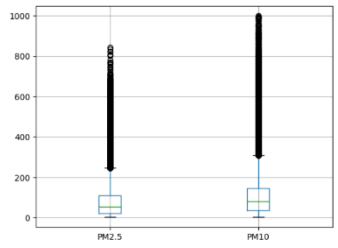
Quantitative Analysis



Heat Map



Feature-imp RFR



Outliers in the data

```
Random Forest Model:
Mean Absolute Error (MAE): 0.04976226776371025
Root Mean Squared Error (RMSE): 1.3280769100458643
```

RFR before fine tuning

```
RandomForestRegressor
RandomForestRegressor(max_depth=20, n_estimators=200, random_state=42)
```

RFR before fine tuning

```
Final Model Performance:
Mean Absolute Error (MAE): 0.09552290193674925
Root Mean Squared Error (RMSE): 1.491044784963474
R² Score: 0.9996490942428157
```

Final Model Performance

```
Predicted AQI: [104.6122214]
Level 3: Lightly Polluted - Sensitive Groups reduce outdoor activities
```

Model Performance to unseen data