frontier tech leaders

A new generation of tech specialists

frontier tech leaders

Predicting PM2.5/AQI for Chiang Rai (Group 11)

Presented by
Myo Zin Thant
&
Swan Htut Oakkar Aung

Frontier tech leaders

Outline

- **Concept note and implementation plan:**
  - *Background*: The air pollution crisis in Chiang Rai.
  - *Objectives*: Forecasting PM2.5 using Machine Learning.
  - *SDG Relation*: Alignment with SDGs 3, 11, and 13.
- **Data**
  - *Data Collection:* APIs (OpenAQ & Open-Meteo).
  - *EDA & Feature Engineering*: Handling time-series data.
- **Model Selection and Training**
  - *Evaluation*: Comparing Linear Regression, Random Forest, & XGBoost.
  - *Refinement*: Hyperparameter tuning & hybrid modeling.
- **Results**
  - Final Model Performance (RMSE, R Squared) & 7-Day Forecast.
- **Deployment**
  - System Architecture & Streamlit Web Dashboard.
- **Future Work**
  - Automated retraining & scaling for mobile apps.

Frontier tech leaders

Background

- **Brief project overview**
  - Predict next-day PM2.5 and AQI for Chiang Rai.
  - Use meteorological and historical pollution data.
  - The final deliverable is a public web dashboard.
  - The core model is the high-performing XGBoost Regressor.

- **Provide brief background**
  - Chiang Rai faces severe air pollution from seasonal burning.
  - This creates high concentrations of hazardous PM2.5.
  - Previous studies validated using weather data to predict pollution.
  - We use advanced non-linear models to improve this prediction.

- **Importance of the problem being solved**
  - Pollution poses a direct threat to public health (SDG 3).
  - Forecasting provides early warnings for the community.
  - The data informs local air quality management (SDG 11).
  - The tool helps people take protective measures.

Frontier tech leaders

Objectives

- **Forecast PM2.5**: Accurately predict the next-day PM2.5 concentration.
- **Predict AQI**: Translate the forecast into an Air Quality Index (AQI) category.
- **Use Key Data**: Utilize meteorological and historical pollution data as inputs.
- **Evaluate Algorithms**: Compare Linear Regression, Random Forest, and XGBoost models.
- **Select Best Model**: Identify the best-performing model (XGBoost) for refinement.
- **Deploy Public Tool**: Create a lightweight web dashboard for public access.

Frontier tech leaders

SDG Relation

- **SDG 3 (Good Health & Well-Being)**: Provides early warnings to minimize pollutant exposure.
- **SDG 11 (Sustainable Cities & Communities)**: Offers data to inform local air quality management strategies.
- **SDG 13 (Climate Action)**: Highlights the link between seasonal activities (burning) and pollution events.
- **The project** translates high accuracy prediction into actionable public value.

frontier tech leaders

Data

Data

- **Source(s) of the dataset**
  - Air Quality data (historical PM2.5) is from the OpenAQ Platform via its API.
  - Meteorological data (weather) is from the Open-Meteo API.
  - Open-Meteo was chosen for its generous free tier (10,000 calls/day) and no API key requirement.
  - The model requires both data types merged into a single time-series dataset.

- **Preprocessing steps during data collection**
  - Raw data was collected at an hourly resolution.
  - It was aggregated to daily averages to match the forecasting objective.
  - The datasets were then merged using the date as the common key.
  - All numerical features were scaled (normalized) for model ingestion.
  - The final dataset was split chronologically (80% train, 20% test) to prevent data leakage.

- **Handling missing values, outliers, etc.**
  - Missing values were addressed using time-series-appropriate imputation techniques.
  - Methods included linear interpolation or forward-fill.
  - This imputation preserves the temporal integrity and autocorrelation of the data.
  - This approach ensures the baseline linear model is not unfairly handicapped by feature scale.

frontier tech leaders

Model

Frontier tech leaders

Model Selection and Training/Testing

- **Details on training the model**
  - Three models were compared: Linear Regression, Random Forest, and XGBoost.
  - XGBoost was selected as the superior algorithm and primary model.
  - A Hybrid Trend-Residual strategy was used to handle time-series trends.
  - This approach allows XGBoost to focus on complex, non-linear weather fluctuations.

- **Hyperparameters used and any cross-validation techniques and details on any hyperparameter tuning**
  - A chronological train-test split (80/20) was used for model validation.
  - This method simulates a real-world forecast and prevents data leakage.
  - Hyperparameter tuning was applied to enhance model performance.
  - Key XGBoost settings included 500 estimators , a learning rate of 0.05 , and a max depth of 6.

- **Comparison of models**
  - XGBoost showed the highest performance on historical test data.
  - XGBoost: Accuracy Score (R Squared) of 0.91 and Average Error (MAE) of 1.02.
  - Random Forest: Accuracy Score (R Squared) of 0.908 and MAE of 1.046.
  - Linear Regression metrics were invalid, so it was excluded from refinement.

- **Visualization**
  - A feature importance plot would show temporal features as most predictive.
  - Top features include the 3-day rolling mean (pm25_roll3) and 1-day lag (pm25_lag1).
  - Meteorological variables (wind speed, humidity) remain secondary predictors.

frontier tech leaders

Result

Frontier tech leaders

Evaluation Results

- XGBoost achieved the best performance among all tested models.
- RMSE: 1.35 µg/m³
- MAE: 1.02 µg/m³
- R^2: 0.91
- The model closely follows actual PM2.5 trends.
- Short-term lag and rolling-average features are the most important predictors.
- The evaluation shows the model is reliable for short-term air quality early warning.

frontier tech leaders

Deployment

Deployment

- The trained XGBoost model was saved using model serialization (.pkl files)
- Feature columns and evaluation metrics were stored for consistent inference
- The model is deployed as a Streamlit web application
- The web app allows users to:
  - View today's PM2.5 level
  - See a 7–14 day PM2.5 forecast
  - Check AQI category and health guidance
  - Monitor model performance and data quality
- This deployment demonstrates a working early-warning system suitable for real-world use

Frontier tech leaders

Conclusion and Futurework

Project Conclusion

- The PM2.5 forecasting system successfully achieved high performance.
- The refined XGBoost model proved superior, achieving an Accuracy Score of 0.91.
- Temporal features (lags and rolling means) were crucial for the model's high accuracy.
- The system is deployed using a decoupled architecture on Streamlit Community Cloud.
- This project established a valuable public health tool for Chiang Rai.

Future Work and Roadmap

- Implement the Automated Retraining Trigger to maintain model accuracy.
- The model monitoring service will detect Model Drift by tracking daily Average Error (MAE) and Precision (RMSE).
- If drift is detected, the system will trigger a new model to be trained on the latest data.
- Explore wrapping the backend in a FastAPI container for future scalability.
- This containerization supports expansion to mobile applications or external consumers.

Thank you!

frontier tech leaders