CP468: Artificial Intelligence

Dr. Sumeet Kaur Sehra

Shayaan Niazi 211983410

Deergh Gandhi 210629500

Aug 12, 2023

Predictive Modeling of Atmospheric Carbon Monoxide Concentrations

Abstract:

Air quality is an imperative concern in urban environments due to its direct impact on both public health and ecological sustainability. This research endeavors to predict the concentration of Carbon Monoxide (CO) in the atmosphere, a prevalent and harmful pollutant. Utilizing an extensive dataset, which encompasses hourly air quality measurements sourced from sensors positioned in a notably polluted region in Italy, this study implements a Linear Regression model as the predictive tool. The model is trained on a subset of this data and subsequently validated against unseen data to assess its predictive capability. The outcomes of this research not only attest to the viability of machine learning in environmental predictions but also underscore the potential implications for real-time air quality monitoring and urban planning.

Introduction:

Air pollution and its ever-increasing threat have spurred numerous studies and research. Carbon Monoxide (CO), a primary pollutant, has been linked to several health concerns, making its prediction and monitoring crucial. This project seeks to harness the power of machine learning to predict CO concentrations based on various other atmospheric readings. We utilized a dataset with hourly averaged measurements from a range of sensors, offering a comprehensive view of air quality over a significant period.

Project Description:

Understanding and predicting air quality, especially pollutants like Carbon Monoxide (CO), is of paramount importance as urban centers continue to grow and environmental concerns mount. This study was structured to explore this critical area.

Objective: The primary goal was to predict Carbon Monoxide (CO) concentrations in the air. CO, being an odorless, colorless gas, is not only a significant air pollutant but also poses severe health risks in high concentrations.

Dataset Overview: The dataset leveraged in this study is a compilation of hourly averaged air quality measurements. These measurements were sourced from a comprehensive Air Quality Chemical Multi Sensor Device, specially designed to capture a wide range of pollutants and atmospheric conditions.

Location & Time Frame: The sensors providing the data were strategically positioned in a region in Italy known for its elevated pollution levels, making the dataset particularly relevant for a real-world application. The data spans a duration from March 2004 to April

2005, providing a broad temporal scope and capturing seasonal variations, which is essential for a holistic understanding.

Data Attributes: With a total of 9357 instances, the dataset is both expansive and detailed. It encapsulates readings from five metal oxide chemical sensors. These sensors measure various pollutants and atmospheric conditions, providing a multi-faceted view of air quality. The measurements encompass a range of atmospheric variables, including but not limited to, concentrations of CO, Non-Methane Hydrocarbons (NMHC), Benzene, Nitrogen Oxides, and particulate matter. Additionally, the dataset also includes readings related to temperature, relative humidity, and absolute humidity, providing a broader context to the pollutant levels.

Significance: The dataset's richness, both in terms of temporal span and the variety of attributes, offers a unique opportunity. By predicting CO levels, this study not only aims to showcase the power of machine learning in environmental applications but also hopes to provide insights that can be instrumental for on-ground monitoring, policy-making, and public health advisories.

Methodology:

Embarking on this project required a meticulous approach:

- **Data Cleaning:** The initial dataset had several anomalies, including redundant columns and potential outliers. A rigorous cleaning process was implemented, dropping unnecessary columns and handling missing data points. Values that did not align with expected readings were replaced and imputed, ensuring a cleaner dataset.
- **Exploratory Data Analysis:** Before delving into modeling, an in-depth exploration of the data was conducted. Distributions of variables were plotted, and correlation matrices were created to discern relationships and potential multicollinearity among the variables.
- **Model Selection and Building:** Based on the problem's nature and the dataset's characteristics, a Linear Regression model was chosen for its simplicity and effectiveness. The data was split into training (80%) and test (20%) sets, ensuring a balanced approach to model training and validation.
- **Model Evaluation:** Post-training, the model's efficacy was gauged using the RMSE and MAE metrics on both training and test sets.

Results:

- Upon completing the model training and subsequent testing, a comprehensive evaluation of its performance was conducted to ascertain its predictive capabilities. Here are the key findings:
- Model Performance Metrics:
- Root Mean Squared Error (RMSE): This metric provides insight into the model's overall prediction error. The Linear Regression model achieved an RMSE of approximately 0.57. Given the scale of CO concentrations in the dataset, this indicates that, on average, the model's predictions deviate from the actual values by about 0.57 units.
- Mean Absolute Error (MAE): The MAE metric sheds light on the average magnitude of errors between predicted and actual values, without considering their direction. The model reported an MAE of approximately 0.38, suggesting that, on average, the absolute difference between the model's predictions and the true values is 0.38 units.
- Model Consistency: An important aspect of model evaluation is ensuring that it performs consistently across different data subsets. The model showcased a comparable performance on both the training and test datasets. This consistency indicates that the model is generalizing well to new data and is not just memorizing the training data.
- Feature Significance: A deeper look into the model coefficients revealed the relative significance of various features. Variables like NOx(GT) and PT08.S1(CO) exhibited a higher influence on CO predictions. This aligns with the initial correlation analysis, suggesting a strong relationship between these variables and CO concentrations.
- Residual Analysis: Examining the residuals (differences between actual and predicted values) provided additional insights. The distribution of residuals was approximately normal, with most residuals centered around zero. This is a positive sign, indicating that the model's errors are random and not biased.
- In summary, the Linear Regression model showcased a promising ability to predict CO concentrations based on other atmospheric variables. The chosen metrics, RMSE and MAE, affirm the model's robustness and reliability in making reasonably accurate predictions.

Conclusion:

The importance of predicting air quality, particularly pollutants like Carbon Monoxide (CO), cannot be understated in today's rapidly urbanizing world. Our endeavor in this study was to harness the predictive capabilities of machine learning to address this vital concern.

A few key takeaways from our research are:

Model Choice: The Linear Regression model, while being one of the simpler machine learning algorithms, proved to be quite effective for this dataset. Its performance metrics, like RMSE and MAE, provided a satisfactory indication of its accuracy, making it a viable tool for predicting CO levels in real-world scenarios.

Data Insights: The dataset, rich with hourly air quality readings, was instrumental in the study's success. However, it also posed challenges, such as missing values and potential outliers. Proper preprocessing and data handling techniques were crucial in ensuring the data's integrity and relevance.

Feature Relationships: An essential aspect that came to light was the interplay between different atmospheric variables. Certain features, like NOx(GT) and PT08.S1(CO), displayed significant influence on CO predictions. Understanding these relationships can be pivotal for more advanced predictive modeling and for real-world sensor deployments.

Future Prospects: While the results are promising, there's always room for improvement. Incorporating additional data sources, such as meteorological data or urban traffic patterns, could enhance the model's predictive capability. Furthermore, experimenting with more advanced machine learning models or ensemble techniques might lead to even more accurate predictions.

Real-world Implications: Beyond the realm of academic research, the practical applications of this study are vast. Urban planners, environmentalists, and policymakers can utilize such predictive models to make informed decisions, be it in infrastructure planning, pollution control measures, or public health advisories.

In wrapping up, this study underscores the potential of machine learning in environmental science. While challenges remain, the path forward is rife with opportunities for innovation and impact, aiming for a future where urban dwellers can breathe a little easier.

References:

Vito,Saverio. (2016). Air Quality. UCI Machine Learning Repository. https://doi.org/10.24432/C59K5F.