Elevating Predictive Air Quality Models: Integrating Socio-Economic Factors for Precision

Lilian Chen, Humishka Zope

Introduction

Amidst growing environmental concerns and the urgent call for sustainable urban development, the management and prediction of air quality have taken center stage. This paper presents an innovative model crafted to predict the Air Quality Index (AQI) by utilizing three distinct datasets. In the initial dataset, we use geographical coordinates as the foundational element, progressively enhancing predictive capabilities by integrating population and energy consumption patterns in subsequent datasets. The combination of these datasets establishes a comprehensive framework to grasp the interactions between geographical, demographic, and energy-related factors, allowing a more nuanced and precise AQI prediction. The practical applications of our model includes: providing invaluable insights for urban planners, environmental regulators, and public health officials in devising strategies to combat air pollution and by effect, promote more sustainable urban environments.

Literature Review

There have been previous attempts at predicting AQI values using weather forecast models, mainly based on factors such as satellite images, air monitoring data, and models that estimate how pollutants travel in the air and where they originate from. One example is a paper published in by Hindawi Journal of Environment and Public Health titled "Prediction of Air Quality Index Using Machine Learning Techniques: A Comparative Analysis", which uses data from major cities in India of pollutant levels such as ozone, carbon monoxide, nitrogen oxide, etc. This paper implements three models: support vector regression (SVR), random forest regression (RFR), and CatBoost regression (CR) and found that random forest regression models perform the best in predicting AQI.

In contrast to conventional methods that primarily rely on weather forecast models and pollutant-specific data, our model takes on a more distinctive human-centric perspective on how AQI can be affected, and therefore predicted. We do so by using features such as population and residential and energy consumption into our dataset, instead of weather or pollutant data. We've additionally employed a diverse set of algorithms, including Linear Regression, Random Forest, Light GBM, and K-Nearest Neighbors, allowing us to capture different aspects of the complex relationship between input variables and AQI.

Datasets

See implementation of visualizations here:
https://colab.research.google.com/drive/1soB8-dT3e7p62rTEToZupYl9ceH-wgfx?usp=sharing

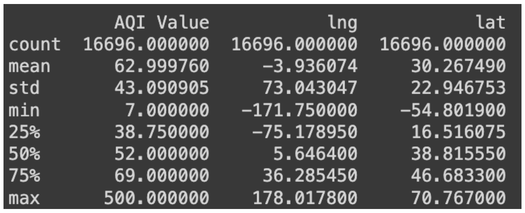
We employed three distinct datasets to enhance the predictive capabilities of our model for forecasting the Air Quality Index (AQI), which we describe below.

First Dataset:

City, Latitude, Longitude

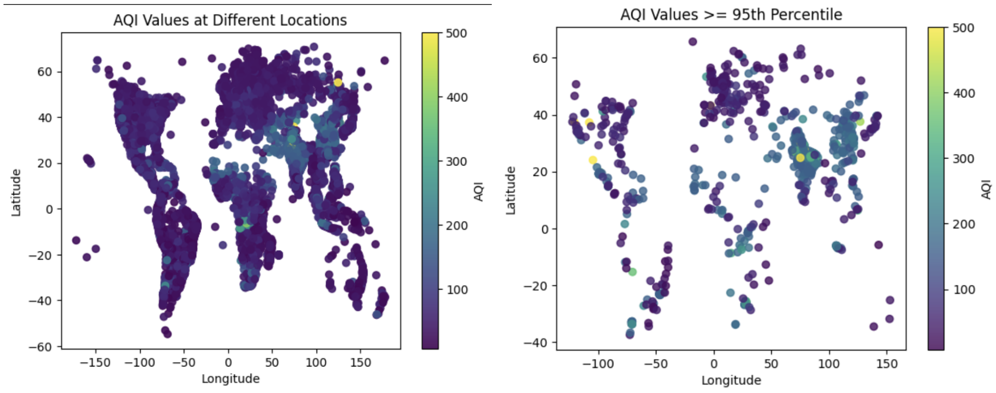
[1] World Cities Air Quality Index by Coordinate: The dataset encompasses up-to-date information, last updated in April 2023, for 16,695 cities worldwide, including their latitude and longitude coordinates, where measurements of AQI indices and various other air pollution indicators were recorded. We preprocess this dataset and merge it with a dataset containing population to yield the second dataset, except we ignore the population column (see next section for more details).

Visualizations:



```
          AQI Value           lng           lat
count  16696.000000  16696.000000  16696.000000
mean      62.999760     -3.936074     30.267490
std       43.090905     73.043047     22.946753
min        7.000000   -171.750000    -54.801900
25%       38.750000    -75.178950     16.516075
50%       52.000000      5.646400     38.815550
75%       69.000000     36.285450     46.683300
max      500.000000    178.017800     70.767000
```

Summary of First Dataset: focusing on the AQI Value aspect of the summary, we see that there is a mean AQI Value of 63 across all datapoints, which we use as our baseline value for the first dataset. Furthermore, the standard deviation of 43 is quite large compared to the mean of 63, indicating that AQI Values are spread out over a considerable range across different locations.



Visualization of AQI Values at different latitude, longitude pairs (left) and Visualization of locations with AQI Values >= 95th+ percentile (right): We can see using this visualization that the dataset is quite robust in that it contains AQI measures from a wide spread of locations across geographical areas. We further see that the vast majority of datasets are <100 as indicated by the color, however in some areas, such as across the coast of the United States and in some parts of South/Southeast Asia for a few examples, there are some points with higher AQI values (as we see blue/green/yellow dots there). We can, at this stage, perhaps hypothesize that these seem to be high populations or known industrial manufacturing areas.

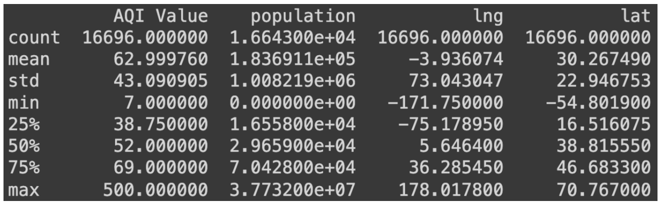
Second Dataset:

City, Latitude, Longitude, Population

[2] World Cities Population by Coordinate: This dataset includes information from March 2023, covering 44,690 cities globally along with their respective estimated populations and latitude and longitude coordinates.

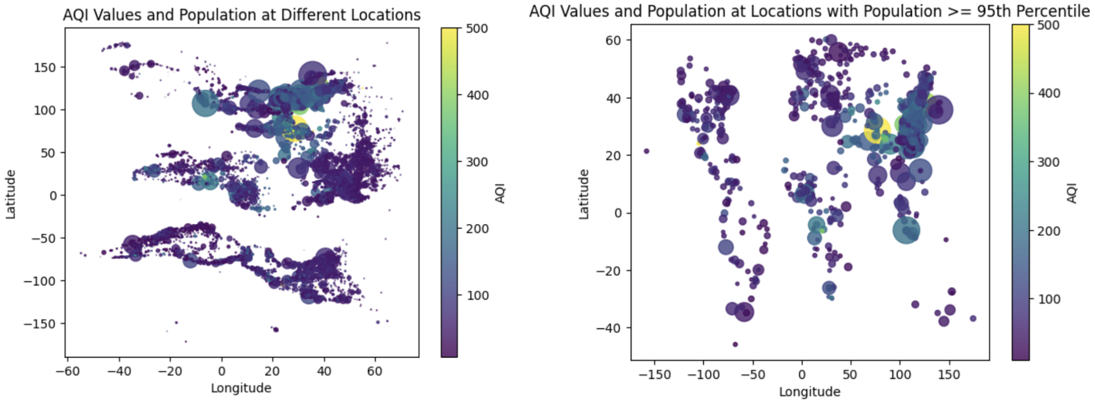
Preprocessing for Final Dataset: To merge the dataset [1] comprising cities and their corresponding AQI values with dataset [2] containing cities and their estimated populations, we merged dataset [2] onto dataset [1] using city name, latitude, and longitude as key identifiers. The inclusion of latitude and longitude as conditions was essential to address the over 2,000 cases where cities shared the same name but were actually distinct locations. Finally, we converted the population, latitude, and longitude sections into floating-point numbers for seamless integration into our models. The final merged dataset contained 16643 distinct cities and their respective coordinates, AQI values, and estimated populations.

Visualizations:



```
          AQI Value    population           lng           lat
count  16696.000000  1.664300e+04  16696.000000  16696.000000
mean      62.999760  1.836911e+05     -3.936074     30.267490
std       43.090905  1.008219e+06     73.043047     22.946753
min        7.000000  0.000000e+00   -171.750000    -54.801900
25%       38.750000  1.655800e+04    -75.178950     16.516075
50%       52.000000  2.965900e+04      5.646400     38.815550
75%       69.000000  7.042800e+04     36.285450     46.683300
max      500.000000  3.773200e+07    178.017800     70.767000
```

Summary of Second Dataset: As we just added population values to the first dataset, we can just analyze the population set. We see that the standard deviation for population is greater than the mean, indicating that we have a very wide range of city populations, which makes sense given that the dataset will contain both major cities (such as New York) and rural towns with populations in the hundreds.



Visualization of AQI Values at different latitude, longitude pairs and population values (left) and Visualization of AQI Values and locations with population >= 95th+ percentile (right): In the left graph, the location of each point is again, determined by its latitude and longitude values, as in the first dataset. The AQI at each point can also be represented as the color, to which we refer to the key on the right to get a sense of its AQI Value. However, we also add population to this visualization in that the size of each point is determined by the population at that location. However, due to the nature in which we have scaled the dataset's population (using the formulation: secondDataset['population'] / secondDataset['population'].max() * 500), city locations with smaller populations do not show up on the map. We can use the right visualization, which only contains points whose populations are >= 95th percentile of the dataset, to see that a good amount of the larger population data points deviate from the dark purple color, indicating that it has a higher AQI (>100). However, there are still a good amount of data points of populous cities with <100 AQIs as well.

<u>Third Dataset:</u>

City, Latitude, Longitude, Population, Residential Electricity Consumption, Residential Natural Gas Consumption, Commercial Electricity Consumption

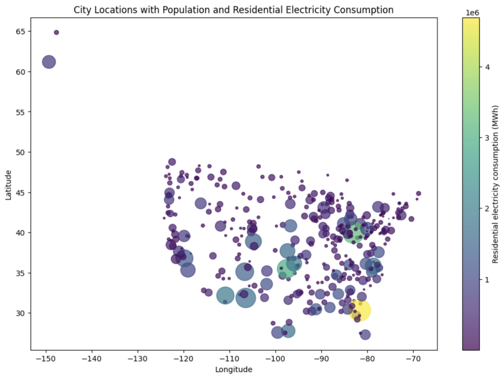
[3] US Cities 2016 Energy Consumption Data: The dataset encompasses information on energy consumption for 23,431 cities in the United States in 2016. For each city, the data includes details such as residential electricity consumption, residential natural gas consumption, and commercial electricity consumption.

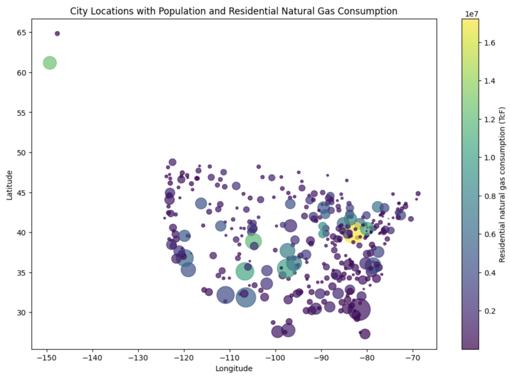
[4] US Cities 2016 AQI Data: The dataset comprises the Air Quality Index (AQI) for 523 cities throughout the United States, with a total of 170,424 rows. This dataset provides AQI values for each of these cities, spanning various dates throughout the year 2016.

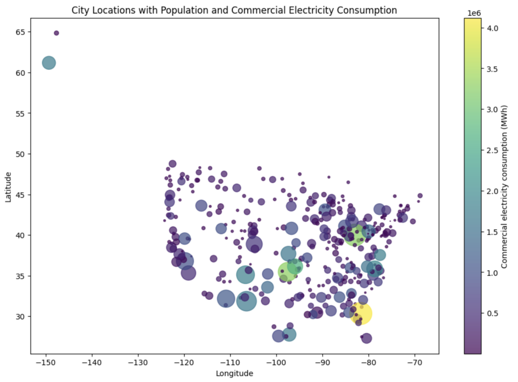
Preprocessing for Final Dataset: We opted to utilize dataset [4] for its AQI indices in 2016, aligning with the energy consumption data from dataset [3] collected in the same year. Ensuring similar timeframes is crucial because it enables a meaningful and interpretable comparison between the two datasets. Given that dataset [4] presents AQI values at 2-3 day intervals throughout the year for each city, we took the approach of averaging these values for each city and consolidating them into a new dataset with 523 rows representing each city's average AQI. This transformation streamlined the dataset [4] from 170,424 rows to a more manageable 523 rows.

To facilitate comparison and merging, we standardized the formatting in both datasets. For instance, dataset [4] expressed cities as "Anchorage, AK," while dataset [3] separated them into two columns, "AK" and "Anchorage City," with the added word 'city' after each town name. We addressed this by splitting dataset [4]'s city column into separate city name and state ID columns. Simultaneously, we reformatted dataset [3]'s city column to exclude the extra "city" tag. Finally, after merging dataset [3] onto dataset [4] based on state abbreviation and city name, we converted AQI values, latitude, longitude, and population into floating-point numbers for ease of use in our modeling endeavors. The final merged dataset contained 354 distinct cities and their respective coordinates, AQI values, estimated populations, and energy consumption metrics.

Visualizations:



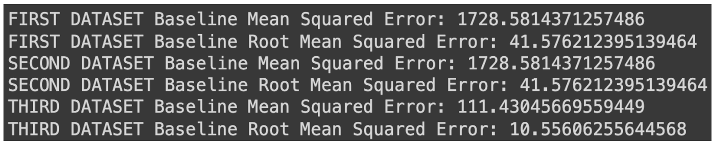




Visualization of (Residential Electricity, Residential Natural Gas, Commercial Electricity) Energy Consumptions at different latitude, longitude pairs and population values: The location of each point is determined by its latitude and longitude values, while the energy consumption at each point can also be represented as the color, to which we refer to the key on the right to get a sense of its value. The size of each point is determined by the population at that location. We are not able to get a full understanding of how consumption is correlated to population based on the chart, however we hope to see that, based on intuition, a higher population means more energy consumption, which could perhaps contribute to the AQI.

Baseline

See implementation of baseline here: 221 Final Project Models.ipynb



```
FIRST DATASET Baseline Mean Squared Error: 1728.5814371257486
FIRST DATASET Baseline Root Mean Squared Error: 41.576212395139464
SECOND DATASET Baseline Mean Squared Error: 1728.5814371257486
SECOND DATASET Baseline Root Mean Squared Error: 41.576212395139464
THIRD DATASET Baseline Mean Squared Error: 111.43045669559449
THIRD DATASET Baseline Root Mean Squared Error: 10.5560625564568
```

First Dataset: In the baseline analysis, we randomly divided the dataset into an 80% training set and a 20% test set. We computed the average Air Quality Index (AQI) across the entire dataset and, for each test point, determined the error in relation to the calculated average AQI.

Second Dataset: The exact procedures as the first dataset were followed, with the only distinction being the usage of the second dataset. (notice that first and second dataset's baselines are the same–this is because we merged the datasets for AQI, latitude, longitude (dataset 1) and population (preprocessing step of dataset 2), except ignoring the population feature on the first dataset.

Third Dataset: The exact procedures as the first dataset were followed, with the only distinction being the usage of the third dataset.

Model & Evaluation Metrics

Models:

For our model, we decided to try 4 different models per dataset (so, 12 models total) to determine what was the best approach for each dataset. For each dataset, we implemented Linear Regression, Random Forest, K Nearest Neighbors, and Light GBM. The chosen models span a range of capabilities. Linear Regression provides a baseline linear model, Random Forest captures non-linear relationships and interactions, KNN leverages the proximity of data points, and Light GBM optimizes for gradient boosting efficiency. Using multiple models allows a more nuanced understanding of the dataset, as we hope to choose the most suitable algorithm for predicting AQI.

Evaluation Metric:

To evaluate the predictive models for Air Quality Index (AQI) across the three datasets, we used Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) to assess the performance. MSE measures the average squared differences between the predicted and actual AQI values, providing a reliable evaluation of the model's accuracy. Meanwhile, RMSE offers a more interpretable metric by presenting the average magnitude of prediction errors.

Implementation of Linear Regression and Random Forest here:
https://colab.research.google.com/drive/1b53gaYbHZFVrYXMZtZWjcYDW7Nke9GDu?usp=sharing

Linear Regression:

Model:

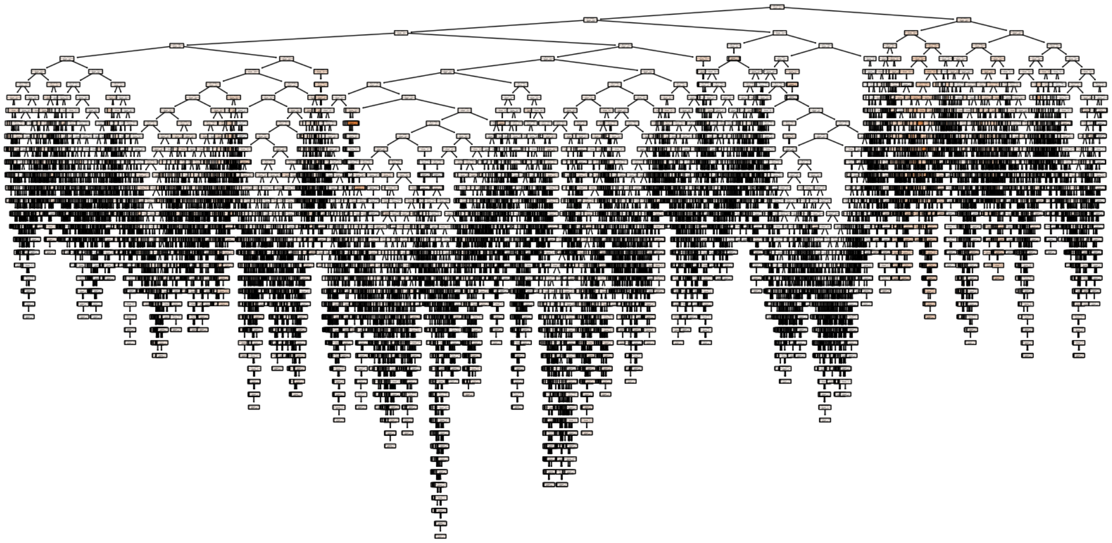
This model assumes linearity of our data, and fits a coefficient vector of w = (w_1, w_2, … w_k) + y_intercept, with the goal of reducing RMSE. In our particular problem, the model will attempt to find a linear relationship between AQI and each feature in our dataset. Our coefficients for dataset #1 were [-0.17273084, 0.08856948] which correspond to a feature vector of [latitude, longitude], and our intercept was 68.55. For dataset #2, our coefficients were [-1.64195055e-01, 7.68747439e-02, 6.96871346e-06] with an intercept of 66.95, which corresponds to a feature vector of [latitude, longitude, population]. For dataset #3, our coefficients were [-4.28627962e-01, -3.52214548e-02, 1.89592875e-05, 6.45030230e-07] with an intercept of [49.70695271], corresponding to a feature vector of [latitude, longitude, population, residential natural gas consumption (TcF)].

Random Forest:

Model:

Random Forest fits a set number (in our case, 100) of decision trees on various subsets of the dataset, and then outputs the average prediction of those decision trees as the final AQI prediction. The model utilizes bootstrapping to build subsets to create each tree and averaging at the end to prevent overfitting. For our model, we fit a set number of 100 decision trees in each random forest and had no maximum on the depth of each tree. Random forest allows the model more flexibility in how it makes predictions based on aspects of the data than other models like Linear Regression.

A visualization of just one of the decision trees for dataset #1 is shown below.



Implementation of KNN and Light GBM here:
https://colab.research.google.com/drive/1soB8-dT3e7p62rTEToZupYl9ceH-wgfx?usp=sharing

KNN:

Model:

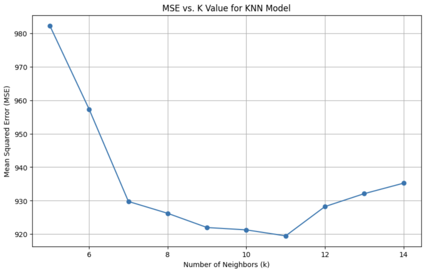
The fundamental principle behind KNN (K Nearest Neighbors) in regression tasks is to predict the value of a new data point by considering the average of its k-nearest neighbors in its feature space–in other words, the algorithm determines a data point's value by inspecting the characteristics of its closest data points.

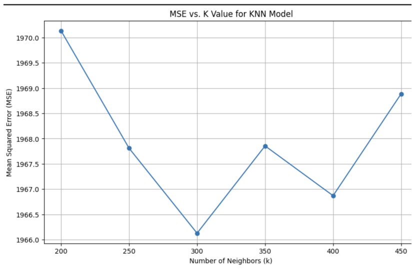
In our context of predicting Air Quality Index based on different datasets, KNN can be applied effectively. For instance, when using the AQI dataset based on latitude and longitude alone, KNN would identify neighboring geographical locations with similar coordinates and derive predictions for a given location based on the AQI values of its nearest neighbors. When incorporating population information in the second dataset, KNN would consider both geographical proximity and population similarity to make predictions. In the third dataset, which includes energy consumption metrics along with population and geographical data, KNN would take into account the multidimensional feature space to find the most similar instances when predicting AQI. KNN's strength lies in its simplicity and adaptability to diverse datasets, making it a suitable choice for predicting Air Quality Index in different scenarios.

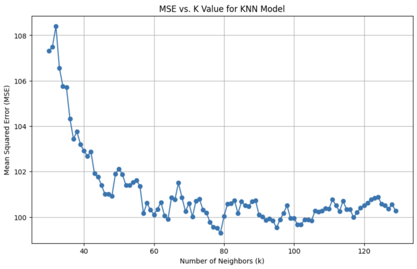
Visualization:

Because we are using multiple features such that the feature space cannot be visualized on a 2D-plane (as there are more than 2 features we are using for the second and third dataset), we do not have a visualization of KNN and its decision boundaries.

The choice of k-value is important because the parameter "k" represents the number of nearest neighbors considered when making predictions for a new data point (a too-small or too-large k-value can lead to poor generalization). Thus, we use cross-validation in order to find the optimal value of "K" in order to minimize the MSE for our KNN model for each of the three datasets. We show the results of cross-validation on KNN models for the three datasets below in order.







Light GBM:

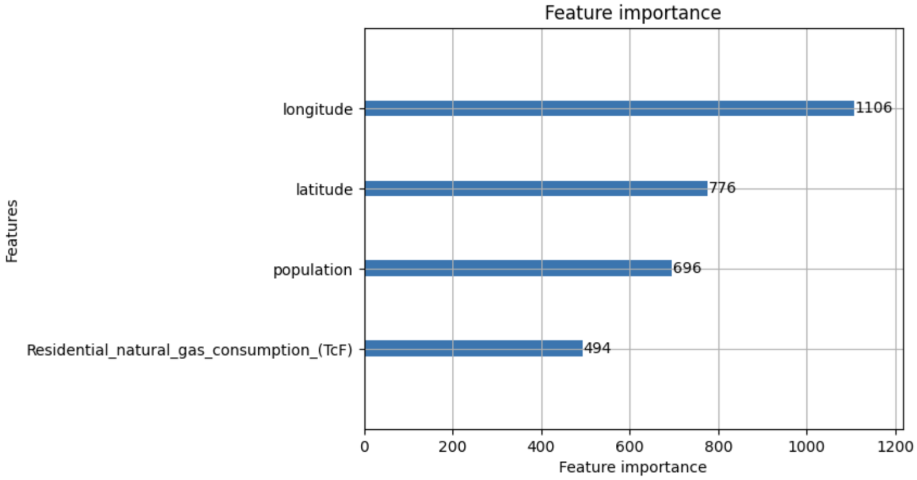
Model:

Light GBM (Light Gradient Boosting Machine), unlike other tree-based methods, grows trees leaf-wise, allowing it to prioritize higher-loss nodes and deeper trees.

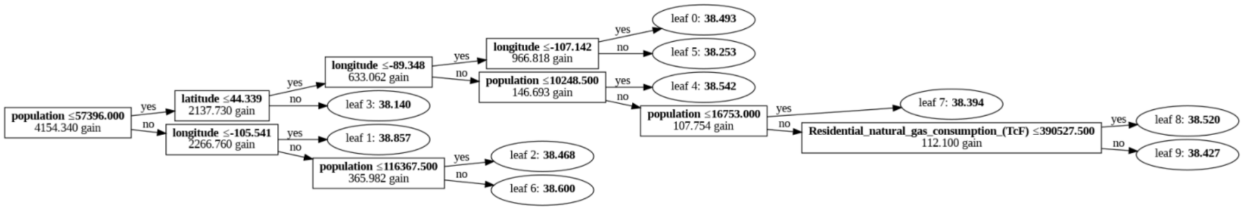
In our context of predicting Air Quality Index (AQI), the diverse nature of the input features in our datasets, such as latitude, longitude, population, and energy consumption metrics, combined with this model's ability to handle large datasets with intricate relationships between variables makes it suitable for effectively capturing the non-linear interactions and dependencies within the data, providing even more accurate predictions for AQI. We see this is true as it outperforms non-tree based models such as linear regression and KNN for all datasets.

Visualizations:

We can visualize the contribution of each feature into the model's prediction by using a feature graph, as shown below. The x-axis represents the importance score of each feature (based on metrics like improvement the feature makes to prediction accuracy and number of times the feature is used in tree splits). While energy consumption has the lowest importance score, it does not necessarily mean it is irrelevant to the model (as we have tested, including the energy consumption feature lowers MSE as compared to without it), but rather that it perhaps helps distinguish nuances in the pattern to make even more precise predictions.



Take the resulting tree split from running Light GMB on Dataset 3, for example–we see that initially, population is the top split, followed by latitude, longitude features, which indicates that these features have a rather significant predictive power in distinguishing varying patterns in the data. However, later we see that energy consumption plays a more significant role in later splits, meaning that perhaps this feature is used for fine-tuning or making finer-grained predictions (on nuances) in the data.



Finally, as proof that energy consumption metrics are at least somewhat related to predicting AQI, we see that the MSE of dataset 3 without energy consumption metrics is higher than if we had added "Residential Natural Gas Consumption" as a feature (listed below).

Dataset 3 model performance with versus without energy consumption metrics:
Mean Squared Error without energy consumption metrics: 95.329
Root Mean Squared Error without energy consumption metrics: 9.763
Mean Squared Error with energy consumption metrics: 85.998
Mean Squared Error without energy consumption metrics: 9.764

Results & Analysis

The Root Mean Squared Error for each model and dataset is shown below (lowest scores are bolded to indicate best performance):

	Dataset 1	Dataset 2	Dataset 3
Baseline:	41.57	41.57	10.55
Linear Regression:	42.26	41.39	10.02
Random Forest:	**30.01**	**28.89**	9.56
K Nearest Neighbors:	30.32	44.33	9.97
Light GBM:	30.18	28.90	**9.27**

Best Performing Models:

For dataset #1 (where input is latitude/longitude and output is AQI), the best performing model was Random Forest with a root Mean Squared Error of 30.01.

For dataset #2 (where input is longitude, latitude, and population and output is AQI), the best model was Random Forest, with a RMSE of 28.89.

For dataset #3 (where input is latitude, longitude, population, Residential natural gas consumption and output is AQI), the best model was Light GBM, with a RMSE of 9.27. We also found that only using the input feature of Residential natural gas consumption for Light GBM lowered the MSE the most (rather than including electricity consumption features as well).

We found that for datasets #1 and #2, Random Forest performs the best, and for dataset #3 Light GBM performed the best (though there were only slight differences in performance between both models in all three datasets). In general, we see that for all three datasets, tree-based models perform the best in reducing the root mean squared error of our data. We believe this is because of a few reasons. First off, tree based models allow us to better model the complexities of our dataset. Both linear regression and K-means clustering make more assumptions about our initial dataset, and we found that linear regression performed the worst for all three datasets, followed by K-means clustering (both of which in general only performed slightly better than the baseline). In addition, tree based models are less susceptible to outliers and overfitting, making them useful models for larger datasets (as dataset #1 and #2 have over 16600 data points).

Feature importance and analysis:

We analyzed our Random Forest models to understand which features were given the most importance in predicting AQI. To measure feature importance, we used Gini feature importance as our metric, which measures the sum over the number of splits that include the feature, weighted by the amount of samples that the feature splits. For dataset #1, we received feature importance values of [0.45974257, 0.54025743] (where higher values means more importance) for features [latitude, longitude], showing that the model placed slightly more importance on longitude than latitude. For dataset #2, we received values of [0.3728111 , 0.45352022, 0.17366869] for features [latitude, longitude, population] and for dataset #3, we received values of [0.33992252, 0.34137689, 0.2006403 , 0.11806028] for features [latitude, longitude, population, residential electricity consumption]. Our feature importance values for dataset #2 and #3 show that latitude and longitude are still the most important features being used to predict AQI, with population and then energy data being of lesser importance.

In addition, despite the fact that the RMSE goes down when moving from each dataset to the next, when compared to the baseline, we see that our Random Forest model improved upon our baseline a roughly equal amount in dataset #1 and #2, and much less in dataset #3. While it is difficult to compare dataset #3 to #1 and #2 because it only includes US cities, this implies that including both population and energy data did not have as significant of an improvement as we hypothesized. For dataset #3, since our baseline (predicting the average AQI across the entire dataset of US cities) performed fairly well, this implies that perhaps predicting AQI in the United States is not a good case study for our problem statement because there's not much variation in between AQI values in difference cities, when compared to other countries and globally.

In conclusion, we found that tree-based models perform best, and that including population and energy metrics did not have a significant impact on our model's ability to predict AQI. We attribute this to lack of comprehensive data on energy consumption around the world (see future work section) as well as lack of variation in AQI values in the US cities dataset (dataset #3).

Future Work

As one of the main issues we ran into was finding good datasets to appropriately run our predictive models on (especially ones containing comprehensive data on energy consumption for different cities around the world–hence why our third dataset is limited to the United States), the next steps we hope to take with our model is the inclusion of more datasets and features relating to air quality mitigation measures. For example, other features that we would have liked to test out, should there have been appropriate data pertaining to it, include cars per person, number of factories, policies to mitigate air quality in each city (quantitatively measured by, for example, measurement of public transportation usage or amount spent on these policies).

Code

For Dataset (csv files) and Google Colab Notebook (ipynb), see here: https://github.com/lilianchs/CS221FinalProject

---

References

Literature Review sources:

N. Srinivasa Gupta, Yashvi Mohta, Khyati Heda, Raahil Armaan, B. Valarmathi, G. Arulkumaran, "Prediction of Air Quality Index Using Machine Learning Techniques: A Comparative Analysis", *Journal of Environmental and Public Health*, vol. 2023, Article ID 4916267, 26 pages, 2023. https://doi.org/10.1155/2023/4916267

Dataset sources:

[1] Ramachandran, Aditya. "World Air Quality Index by City and Coordinates." *Kaggle*, 7 May 2023, www.kaggle.com/datasets/adityaramachandran27/world-air-quality-index-by-city-and-coordinates.

[2] "World Cities Database." *Simplemaps*, United States Census , simplemaps.com/data/world -cities. Accessed 6 Dec. 2023.

[3] Day, Megan. "Open Energy Data Initiative (OEDI)." *City and County Energy Profiles*, 20 Dec. 2019, data.openei.org/submissions/149.

[4] "Daily AQI." *EPA*, Environmental Protection Agency, aqs.epa.gov/aqsweb/airdata/download_files.html#AQI. Accessed 6 Dec. 2023.