Data 100
13 December, 2021

AQI Project Final Report

Group Members: David Dangond, Aditya Vallabhani, Jongho Choi, Irfan Syed

Problem Statement

Our group hypothesized that AQI in California has a strong positive correlation with AADT/Traffic, and also varies depending on geographic land use and time of year. We can confirm or reject our hypothesis by checking whether the correlations between these features and AQI are positive or negative, as well as the strength of the relationship. Unlimited access to modern datasets would allow us to develop a thorough model relating AQI with traffic, land use, and time of year, since we would have near-perfect information without any missing data points. This way, our model can reveal information about the correlation between AQI and the features. We plan to train our model based on various sets of features and compare the model's scores to determine if certain features have stronger or weaker correlations with AQI.

We found that certain types of land use and location settings (i.e. military reservations or urban areas) have a positive correlation with AQI. This trend can be accurately extrapolated to the real world, as urban areas would have higher levels of traffic, meaning more pollutants and higher AQI levels.

We also determined that AQI fluctuates depending on the time of year, especially in the period of August-October. This is explained by California's wildfire seasons, which would cause AQI during those months to rise relative to the rest.

Based on our EDA and our model's performance on various feature sets, we confirmed our hypothesis that AADT is positively correlated with AQI. By further analyzing our model's performance on the time of year and site information (Month, Land Use, and Location Setting) features, we determined that a strong correlation exists between site information, time of year, and AQI. This is demonstrated by the improvement in model performance between the feature sets without these site and time features vs. our model performance including those features.

Answer

Based on our results, we landed on the consensus that our hypothesis can be confirmed. The initial model that we developed was based on our hypothesis, and thus used the three basic features: AADT/Traffic, geographic land use, time of the year. We created a model out of this that would predict AQI solely off of these features in order to test the correlation between the features and AQI. This model returned a decent score for our first set of features from the hypothesis: it had around a 60% accuracy of detecting AQI level off of traffic level, geographic land use, and time of year. We had to interpret these results how we saw fit. Our model is set up to predict the AQI level category (6 total) based on the given features. Random predictions would have an accuracy of ⅙ or ~16% at best, since it would be randomly assigning each prediction into one of the 6 AQI Category buckets. Because our model performs at a

significantly higher level than random predictions, we can confirm that the set of features have a strong correlation with AQI.

However, we also came to the conclusion that this set of features we chose in our hypothesis was not the optimal set of features that could predict AQI. So although we could consider our hypothesis confirmed because the strong positive correlation existed, we still believed there were stronger sets of features that are even more positively correlated than solely AADT/Traffic, geographic land use, and time of year. This thought led us to explore more complicated models, which are improvements on the initial model made from our hypothesis.

We can also confirm our hypothesis through Exploratory Data Analysis:

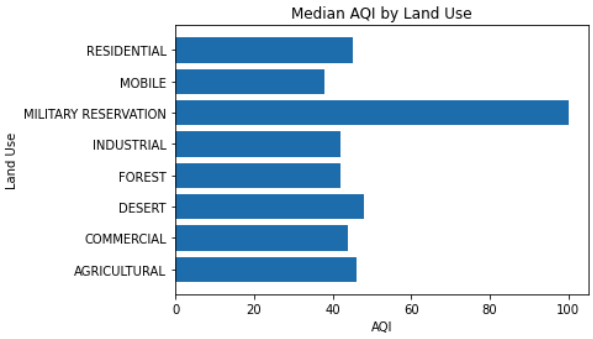


Figure 1

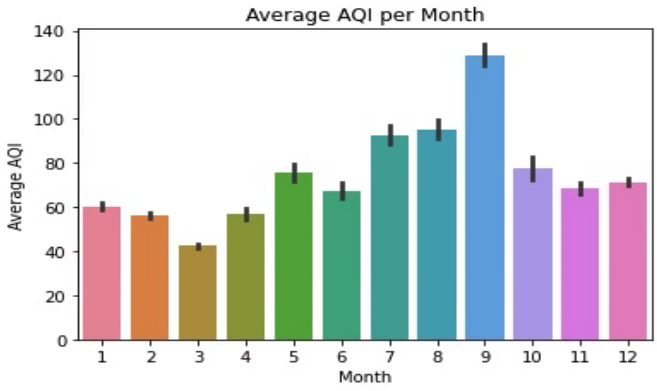


Figure 3

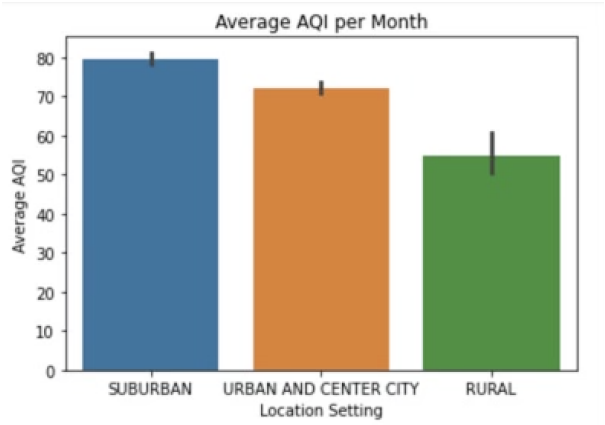


Figure 2

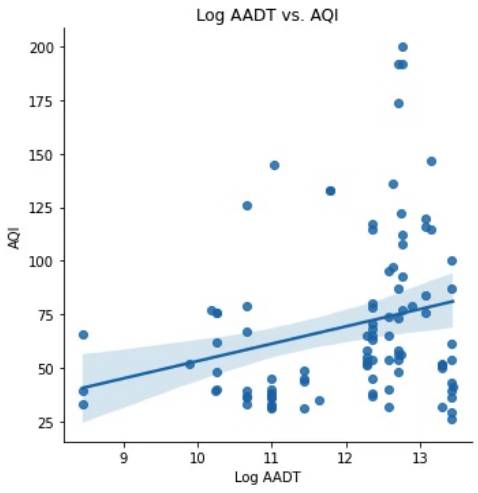


Figure 4

These graphs plot geographic land use (and subsequently the same data grouped by location type), time of the year, and AADT/traffic against AQI to show any relationships between the potential features and AQI levels.

Modeling

For the open-ended modeling portion of this project, we tried various methods of modeling with different combinations of features. We chose the three features for our model based on our original EDA, which showed that these features, as you can see in the above visualizations, had a strong positive correlation with AQI. The three features were AADT/traffic levels, geographic land use (by location setting), and time of year. The model we decided to use originally was a linear regression model. We chose these features because they align with our hypothesis, and so it would be intuitive to use these in order to predict AQI, which would be the predicted output from our linear regression model. We chose linear regression as our model initially because we wanted some type of model to work as a baseline before we explore our other options, and we expected there to be a positive correlation between the chosen features and AQI. This model gave us an accuracy in our validation set of about 60%. This was way too low for us to consider it a satisfactory model, so we decided to then use random forest regression as our next model for the following reasons:

- Since random forests is a decision tree algorithm, it does not make any assumptions about the underlying distribution of the data. It is also much less influenced by outliers than other methods of modeling, such as linear regression, which is better to use when the underlying function is truly linear. With the three features we chose, it is very unlikely that the underlying model would be linear since they are not all directly related to AQI.
    - For example, we have little reason to believe that an increase in the AADT and changing the land use to one that correlates to higher AQI, such as from rural to suburban, at the same time would result in a *linear* increase in AQI, since there are other factors related to AQI that could change depending on the new setting, that could have an opposite effect on the AQI as what we would expect by increasing the AQI. An example of a factor that would have this opposite effect could be the population density of the new location. If this location has a very low population density, then we would expect the same increase in traffic in this new location to have a significantly lower effect on the AQI, since the increase in the amount of traffic would probably not affect the AQI as much for a region that has a lower population density and thus may not already have high levels of traffic.

  The key point here is that having multiple features in our model makes it very unlikely that the underlying distribution of the data is linear, and the aforementioned example shows just one of many indirect factors that could be responsible for this.

- Random forests are also better suited for feature selection. This is because random forests are not heavily impacted by features that are not as useful. In fact, the random forests modeling method does not even include useless features when it comes to splitting on the data. This is a significant improvement from the linear regression model, since it can filter out useless features and hone in more accurately on the underlying distribution. Thus, in the case where one of our features might not be as good of a feature as we believe it to be, random forest will still filter it out or at least not put as much weight into that feature.

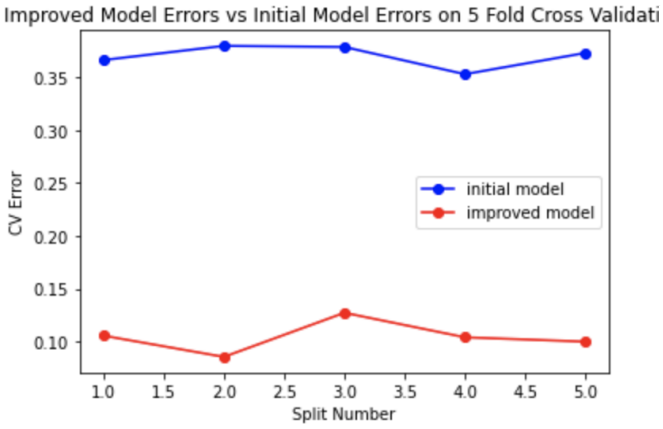
Model Evaluation and Analysis

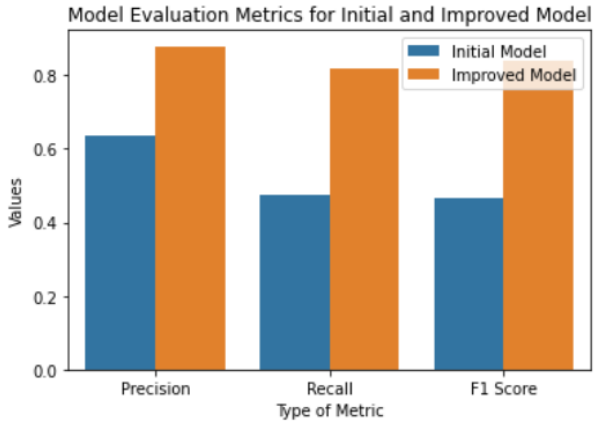
We decided to use cross-validation to test our models. We split our dataset five times and created five training and test sets to validate our model on. After training our model on the five different training sets, we validated it using the model's predictions across the test sets, calculating the total error for each validation split. This is visualized in the graphs below, showing a line plot of the model's performance error per split for our initial and improved model. The mean error for the initial model hovered between the 0.30-0.40 range, while the improved model had a mean error between the 0.07-0.15 range.

Since our model is predicting a final categorical variable (AQI Category), we decided to use precision, recall, and f1 scores to evaluate our model's performance and compare models. Based on our initial feature set for 11a, our precision, recall, and f1 scores were 0.637, 0.475, 0.466, respectively. After improving our model with additional features and changing the type of regression, our precision, recall, and f1 scores improved across the board to be 0.879, 0.818, 0.838, respectively.

This demonstrates that our model improved in both the quality and quantity of predictions, indicated by the increase in precision and recall. The improvement in the f1 score indicates that our model improved in the balance between precision and recall as well.

We chose to examine the cv_error across 5 different folds of our data set to see whether or not our model would consistently predict the correct categories. Since the cv_errors were consistently lower across all five folds, we determined that our model showed significant improvement after better feature selection.





Model Improvement

Improvement #1: As stated in the answer section, the results of the models were represented as accuracies, which showed how accurately the model could predict AQI level. This being the case, our model results were open to interpretation. Although our initial set of features (AADT/traffic, geographic land use, time of year) yielded an accuracy of 60%, which is better than a baseline random model, we were not satisfied. Our problem was that we didn't use the strongest features, and we also didn't use enough features in general. Although the initial features we chose were based on our belief that they would be strong indicators of AQI, there are actually a lot of real world factors that affect AQI, so one or two features was not enough to predict AQI, and we could tell because our accuracy was too low. In order to aim for perfection, we improved our model by adding new and more relevant features. The way we went about this was by conducting even more exploratory data analysis on the rest of the data. By checking all the other external features such as elevation, and site information against AQI, we came up with a list of features that have a stronger correlation with AQI. Just like we did with AADT and the other features, we plotted every feature against AQI to check their relationship. After filtering through the features to choose the ones with the strongest relations, we ended on a final list of features. Overall, it is fair to say that adding these extra features successfully improved our model, as we saw a double digit increase in accuracy. It makes sense that the accuracy increases as we add features such as carbon dioxide levels and temperature since these are factors in real life that impact AQI.

Improvement #2: We decided to one-hot-encode the month feature as an improvement to our model. At first, it was not immediately obvious that keeping month as a numerical variable would throw off our accuracy, but as we examined our model's performance, we realized that month is actually a categorical variable where the useful information comes from which month a particular datapoint is in, rather than the overall sequence of months over the year. By

one-hot-encoding month, our model was able to learn the relationship between a particular month and the AQI levels. By proxy, this allowed our model to learn the effect California wildfire season has on AQI, since AQI spikes drastically during the wildfire season/months. We also one-hot-encoded other categorical variables such as location setting to more accurately tie AQI to particular types of regions. For example, rural regions had relatively lower AQI readings compared to urban environments. By one-hot-encoding this information and feeding it into our model, we saw an improvement in our evaluation metrics (precision, recall, f1 score, cv_error).

Future Work

Our model ended up having a relatively decent accuracy, but there are definitely lots of improvements to be made. While our EDA showed that temperature is positively correlated with AQI, the reality is that this relationship is false. Colder temperatures will prevent hot air from rising, therefore trapping pollutants at ground level. We believe that the flaw in our EDA was due to the fact that AQI levels rise significantly during California wildfire season, when it is generally a hotter time of the year. Because of this conflict, our model was not able to accurately learn the relationship between temperature and AQI. Further exploration can be done by finding ways to decouple these two variables so that our model can learn the actual relationship between temperature and AQI, and thus improve its performance. Moreover, we noticed that adding wind as a feature into our model would lower its accuracy and increase the error. Thus, finding a way to add wind as a feature in order to increase accuracy would be a possible future improvement because we can predict AQI levels better based on wind patterns and air circulation.