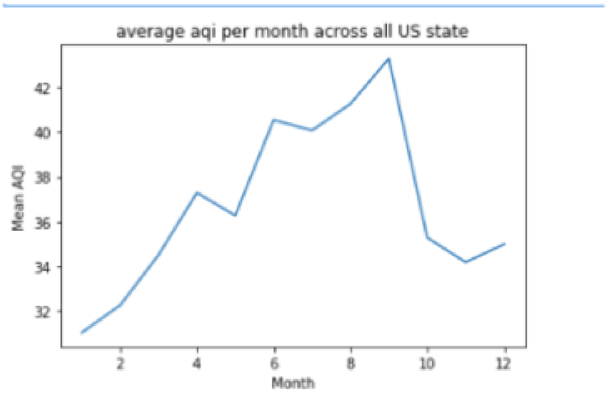
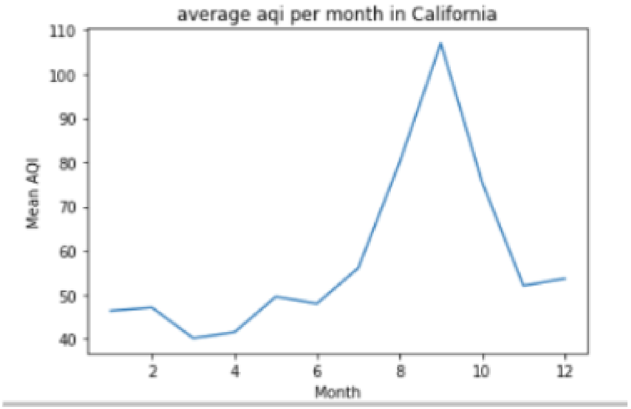
Data 100 Final Project Report: AQI

Group Members: Dingyue Zhang, Larry Gan, Bruce Xu

*Exploratory Data Analysis (Design Doc Clobber):

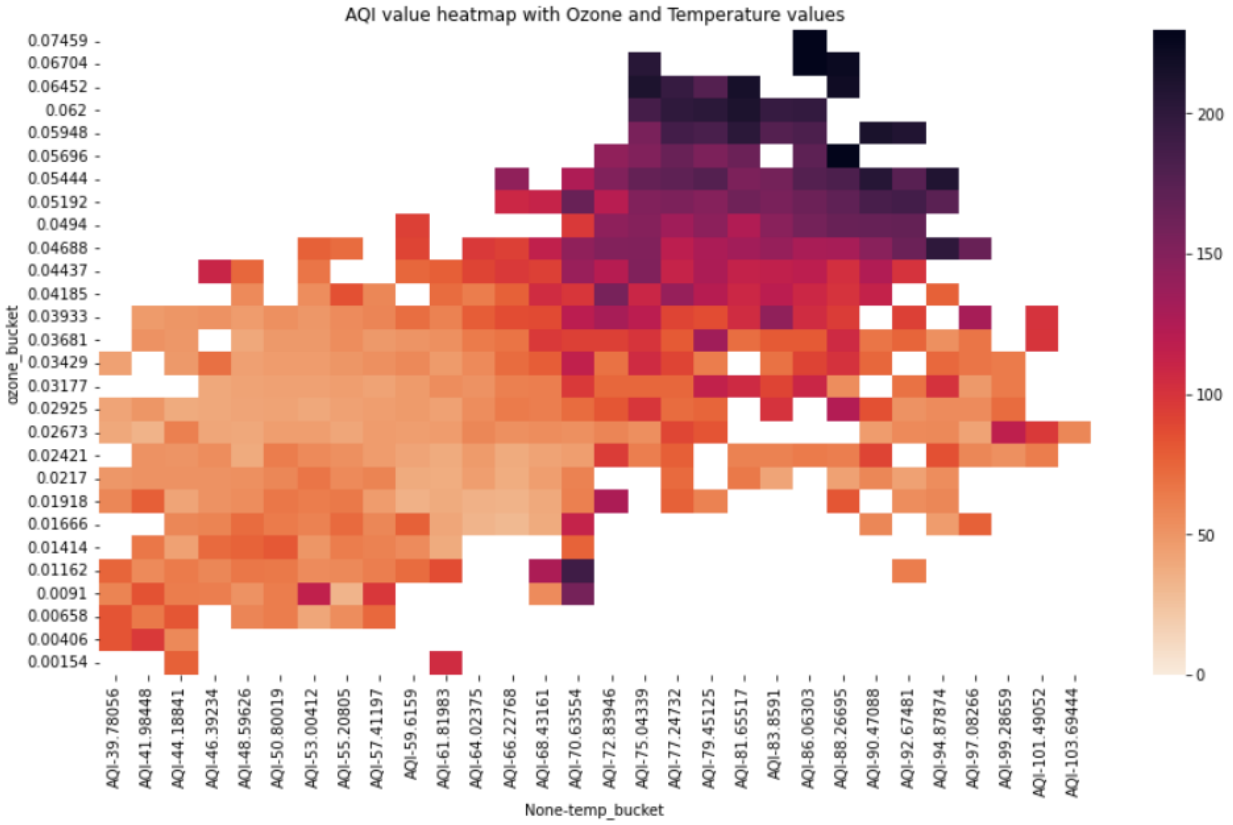
We made a line plot that shows the changes in the average AQI on different months for both California and across all US states. We found it interesting that there seems to be a seasonal pattern to the mean AQI per month where the mean AQI nationally increases from January to August, reaches its peak in September, and decreases from September to December. Meanwhile, California, AQI still peaks in September, but the increase is more abrupt.





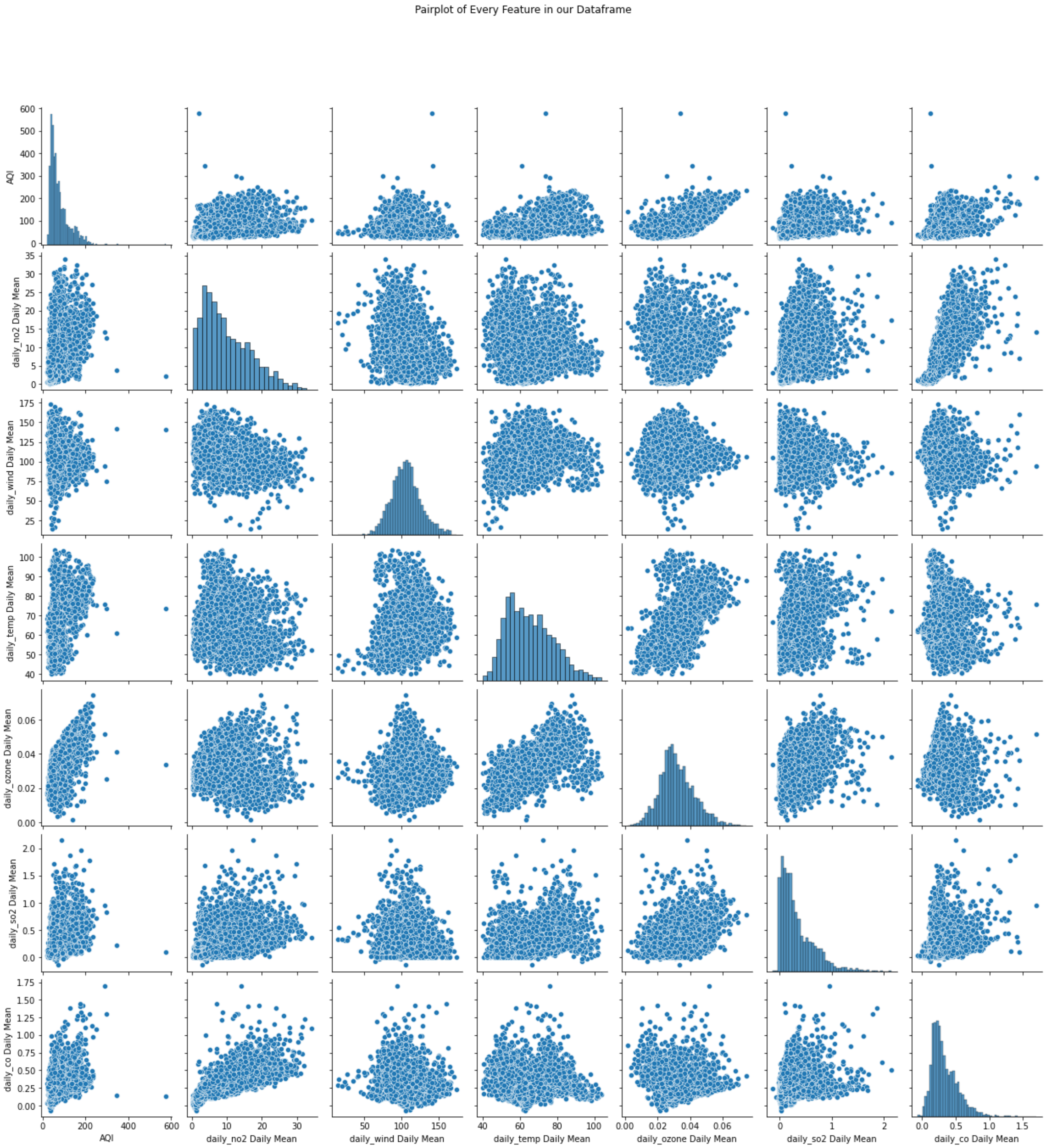
In addition to understanding the change in AQI for the different states and their location, we also want to discover whether there is any relationship between AQI and the temperature, wind, particulate matter, and chemical concentrations. Therefore, we created a data frame with the AQI data and the arithmetic mean for each of the possible features merged using the defining site and the dates.

We also created a heatmap for ozone and temperature values to see their influence on AQI values. We find that both of them tend to have a positive relationship to AQI values. On the one hand, at the bottom left corner, both ozone and temperature values are small, and the AQI values are small. On the other hand, the top right corner marks a cluster of dark pixels that suggest high AQI values when ozone and temperature values are high.



*(*Design Doc Clobber Plot Title)*

After creating the data frame, we made a pair plot that would show all the relationships between each of the variables against each other. We see a positive linear relationship between daily means of co and no2 and daily means of ozone and temperature. We also see a negative linear relationship between daily means of ozone and co and ozone and no2. We also observed that there exist some weak to moderate positive correlations between AQI and daily means of so2, co, no2, or temperature; a weak negative linear correlation between AQI and wind.



(Design Doc Clobber Plot Title)

*Additional Open-ended Questions:

- Will the data for 2020 not be representative because of the great wildfire this year?
- Does moisture content have an effect on AQI?
- Why do states closer to the ocean have worse AQI than states inland?
- Are social factors responsible for worsening AQI?

Problem:

****Hypothesis*** **(Design Doc Clobber)**:
Null Hypothesis: AQI doesn't have any **seasonal pattern**, and the location, social, and time features such as population density, smoker density, and traffic AADT are **not as important as** the meteorological parameters such as wind and temperature in predicting AQI; adding those features would **not result in a significant increase in prediction accuracy**.

****Alternative Hypothesis(Design Doc Clobber):*** AQI values follows a **seasonal pattern** and has a strong correlation with location and social features such as population density, smoker density, and traffic AADT. The location, social, and time features are **as important as** all the meteorological parameters such as wind and temperature we have (importance difference smaller than 10 percent) and the additional features can **improve the accuracy** of predicting AQI categories (by at least 15%).

*How to Test (Design Doc Clobber):

We will test the seasonal pattern of AQI values by measuring the importance score of the date loc feature (date loc is a feature we create by concatenating the county with the month the measurement was made in). We will also import additional data that could affect AQI such as population density, smokers density, and traffic AADT. Importance score is a value between 0 and 100 for each of the features used in the model; the sum of importance score for every feature is 100). If it has an importance score larger than 10, we say we reject this portion of our null hypothesis. To compare the importance of meteorological features and location features, we will compare the proportion of the importance scores the features in those two categories take up. If we see the percent difference of the two sums is less than 10 percent, we reject this portion of the null hypothesis. To test if the improvement of predicting accuracy of AQI after adding location features is at least 15%, we will do an accuracy test comparison for the model with and without the additional features.

Modeling:

Model Selection and Reasons:

We chose to use the random forest model to predict AQI because although the random forest model has lower interpretability, it generally provides higher accuracy with cross-validation. It does not make any assumptions about the distribution of the data, and it is less influenced by outliers than other algorithms. In addition, random forest naturally works well with categorical variables - we don't have to predict the AQI scores then translate that into categories manually.

Input, Output and Reasons:

Our goal is to predict AQI categories (based on the standard AQI categorization according to EPA).

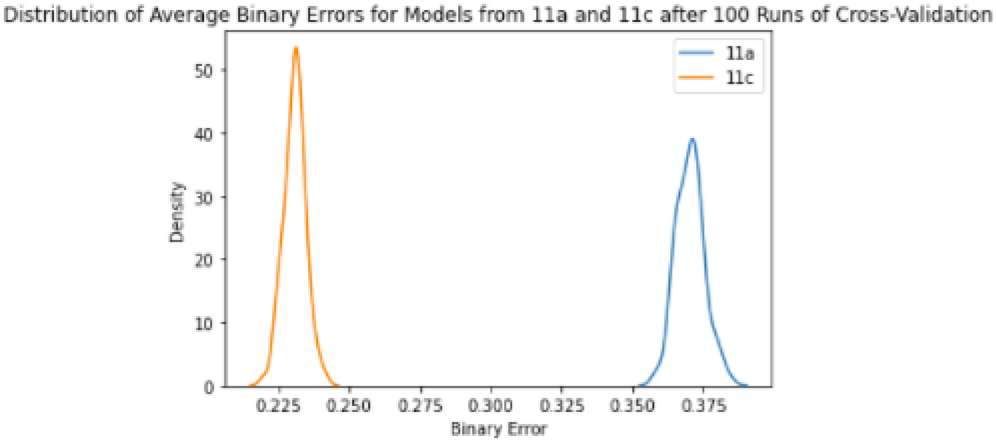
For the baseline model, the 3 features we chose are the daily wind mean, daily temperature mean and daily ozone mean. We chose these 3 features because of our discovery in EDA, especially through the pairplot. We find those three features have a more clear relationship with AQI. Therefore, we believe these features will be a sufficient baseline model with some predictive power because ozone itself is a harmful air pollutant; air temperature and wind affects the movement of air, and thus air pollution. Additionally, we want to see how well the model predicts without the addition of seasonal, location and geographical features.

To improve our model, we added dateloc (a feature we create by concatenating the county with the month the measurement was made in), population density, smokers density, and traffic AADT. From our previous data exploratory analysis, we believe AQI follows a seasonal pattern, which led to the idea of using dateloc. We also think population density will increase the predictive power because denser area tends to have more daily activities that could contribute to worse AQI. By the same token, the area with higher smoker density could have worse AQI. Finally, we think higher traffic volume will make air quality worse because car usages create pollution.

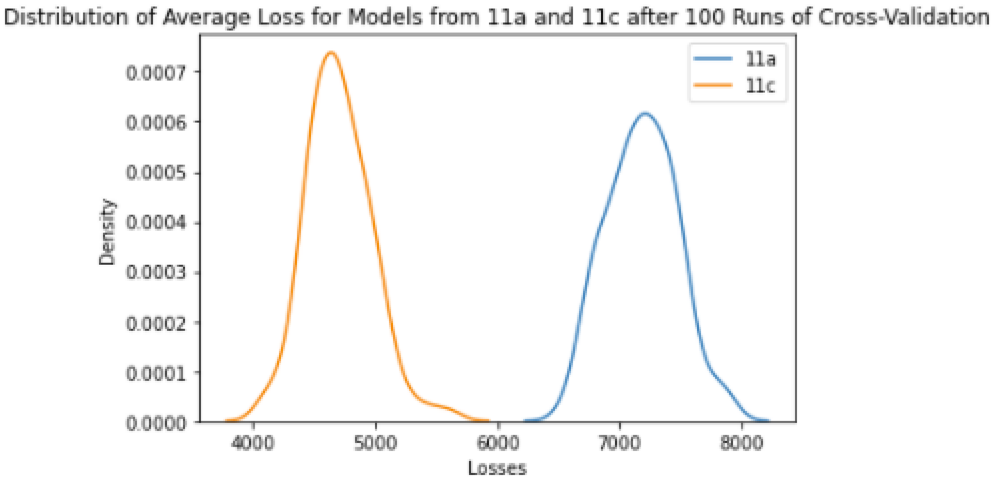
Model Evaluation and Analysis:

We created our own loss function in order to evaluate our model. Since we want to penalize big mistakes more than small mistakes, we designed a custom loss function where predictions that fall farther from the actual category get penalized more. We calculate how far a prediction is by assigning numerical values to each category of AQI (good: 0, moderate: 1, unhealthy sensitive groups: 2, unhealthy: 3, very unhealthy: 4, hazardous: 5, these pairings are analogous to their index in a list sorted by severity of AQI, ascending). We calculate the loss as equal to the 4 raised to the absolute difference between the index of the predicted category and the index of the actual category. This way, larger errors will increase loss exponentially, thereby punishing big mistakes more than small mistakes. We chose a base of 4 for our exponential arbitrarily. Any positive number could theoretically work because the exponential function penalizes large mistakes more than small ones. We also calculate the binary_error, which is the fraction of inputs on our validation set that our model classifies incorrectly.

Below are the graphs that show the distribution of the binary errors and the average loss using the baseline model and the improved model. Neither distribution overlaps the other. The baseline model has an average binary error of ~0.37 and an average loss of ~7185. The improved model has an average binary error of ~0.23 and an average loss of ~4705. There is a 37.8% decrease in the binary error rate and a 34.5% decrease in average loss. Looking at the percentage decrease and the distribution, it is clear that the improved model has a significant improvement on both the binary error as well as the average loss (the average loss is computed using our own distance loss function).

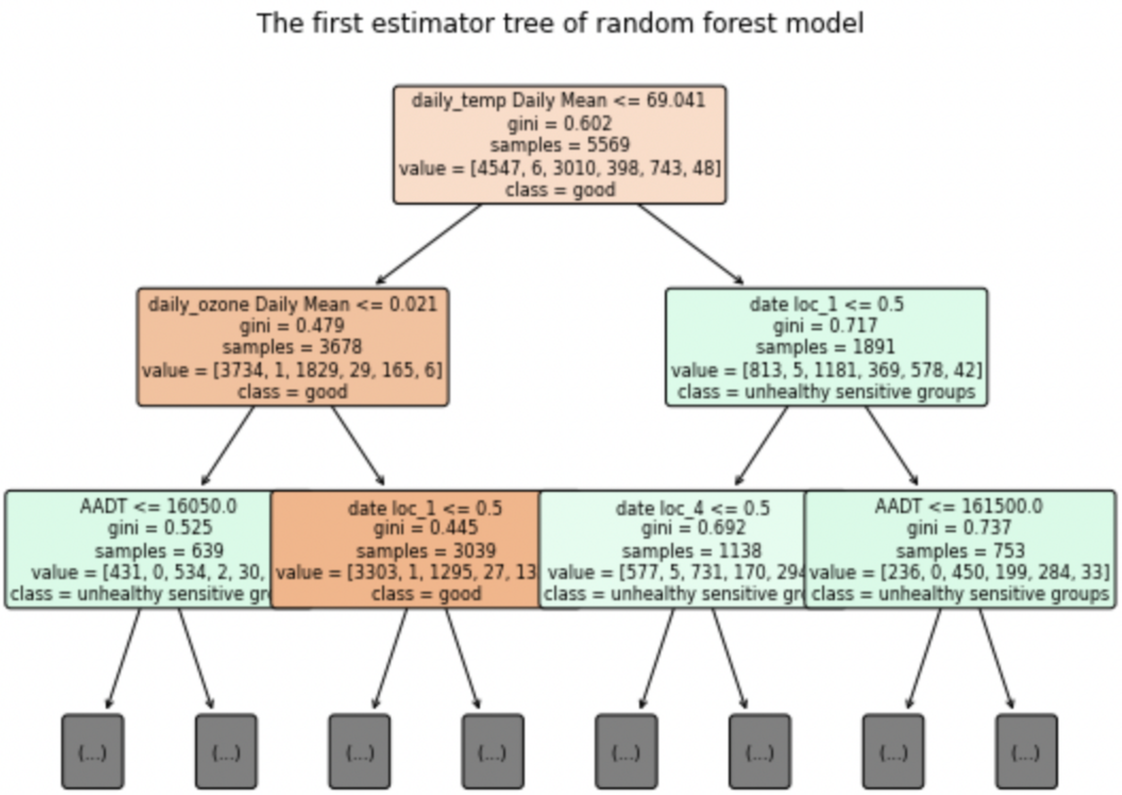


Average Binary Error for our baseline model (which corresponds to the model in 11a) is 0.370349592885698
Average Binary Error for improved model (which corresponds to the model in 11c) is 0.23080656212776368

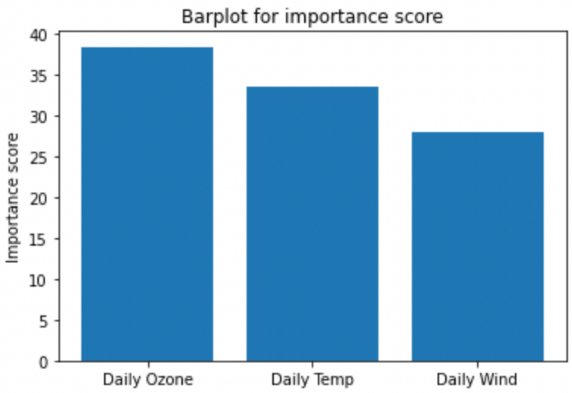


Average Loss for baseline model (which corresponds to the model in 11a) is 7185.404
Average Loss for improved model (which corresponds to the model in 11c) is 4705.484

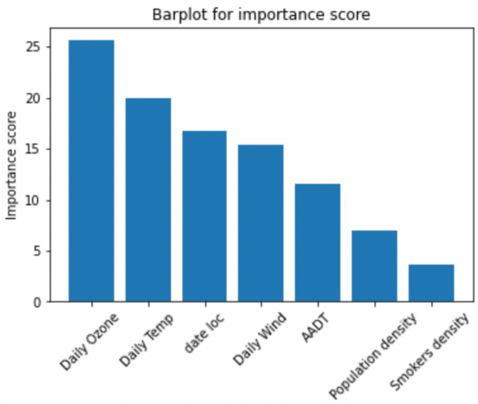
This is a plot of the first estimator tree of our improved random forest model. Although this is not a final decision tree, it still provides some insights into the significance of the features and the important split points. We can see that temperature, ozone, and dateloc are important as they are on the top of the tree. We can also read the split point and the Gini index printed on each node.


The first estimator tree of random forest model
daily_temp Daily Mean <= 69.041
gini = 0.602
samples = 5569
value = [4547, 6, 3010, 398, 743, 48]
class = good
daily_ozone Daily Mean <= 0.021
gini = 0.479
samples = 3678
value = [3734, 1, 1829, 29, 165, 6]
class = good
date loc_1 <= 0.5
gini = 0.717
samples = 1891
value = [813, 5, 1181, 369, 578, 42]
class = unhealthy sensitive groups
AADT <= 16050.0
gini = 0.525
samples = 639
value = [431, 0, 534, 2, 30,
class = unhealthy sensitive gr
date loc_1 <= 0.5
gini = 0.445
samples = 3039
value = [3303, 1, 1295, 27, 13
class = good
date loc_4 <= 0.5
gini = 0.692
samples = 1138
value = [577, 5, 731, 170, 29
class = unhealthy sensitive gr
AADT <= 161500.0
gini = 0.737
samples = 753
value = [236, 0, 450, 199, 284, 33]
class = unhealthy sensitive groups
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)

To further understand the random forest model in a more complete picture, we made a bar graph of the importance scores, computed using Gini impurity/information gain, of all the features. We can see that the importance of ozone > temperature > dateloc > wind > AADT > population density > smoker density. Comparatively, smoker density and population density are not as important as the others, and they both have importance scores below 10. We can also see that the sum of importance scores of the location and social features is 39.3. The sum of importance scores of the meteorological parameters is 60.8. The difference between the two sums is 35%.



Importance Score of Baseline model



Importance Score of Improved model

Result of Experiment:

From our model and analysis above, we can see that dateloc, the feature design to test the seasonality of AQI, has an importance score of 16.8 > 10. And our improved model with newly added location and social features has a 37.8% decrease in the binary error rate and a 34.5% decrease in average loss. However, we cannot conclude that the location, social and time features are as important as the meteorological parameters because there is a 35% difference in their importance scores, which is much larger than our 10% line. In conclusion, we cannot reject our null hypothesis.

Model Improvement:

Our baseline model used three features (ozone, wind, temperature) in order to make predictions. Given our relatively high average binary error rate of ~0.37 with cross-validation and the fact that it did not include every feature that could affect AQI, we believed that our baseline model was underfitting the data.

Improvement 1 (problem and solution):

Since our original model does not have a feature that synthesizes all the seasonal and location features, we add a date loc feature to account for that, which is a feature that represented location and time created by concatenating the county with the month the measurement was made in.

We chose this feature because intuitively, if we believe that AQI follows a seasonal pattern, we should use some kind of a time feature to capture the seasonality.

Improvement 2 (problem and solution):
To fix the problem that our original model does not have any location-related features, we added 4 additional features of average daily traffic, population density per square mile of land area, and density of smokers. Intuitively, we think that AQI is also related to social factors that change depending on the area, which led to our final selection.

Result:
By adding in these features, we were able to decrease our average binary error rate to ~0.23 with cross-validation, as well as lower our average loss from our custom loss function from ~7185 to ~4705. Because both our average binary error and average loss went down, adding additional features did improve our model.

Future Works:

Currently, we only have one model to predict AQI for the entirety of California. However, from our previous EDA, we see that states closer to the ocean generally have worse AQI than states inland. Maybe to further improve our model, we can build multiple models and choose the model depending on whether the county is near the water. We believe that using different models can increase our accuracy. We can even perform clustering then predict to further increase accuracy.

Our current model only records the 5 fold cross-validation error; however, we can use cross-validation to select the best max number of features to use to further increase accuracy and decrease chances of overfitting.

We may also consider some social factors such as median income. We can also explore whether AQI has an effect on people’s health, including mental health, by incorporating the percentage of mental illness for each county.

Some possible extensions can be to build models for the other states since our model is only for California, and we believe it is not reasonable to train the model only using California data to be applied to the other states.