Data 100 Final Project (AQI)
Dec 13, 2021
Group Members: Tyler Freund, Sai Achalla, Sushant Vema, Sheer Karny

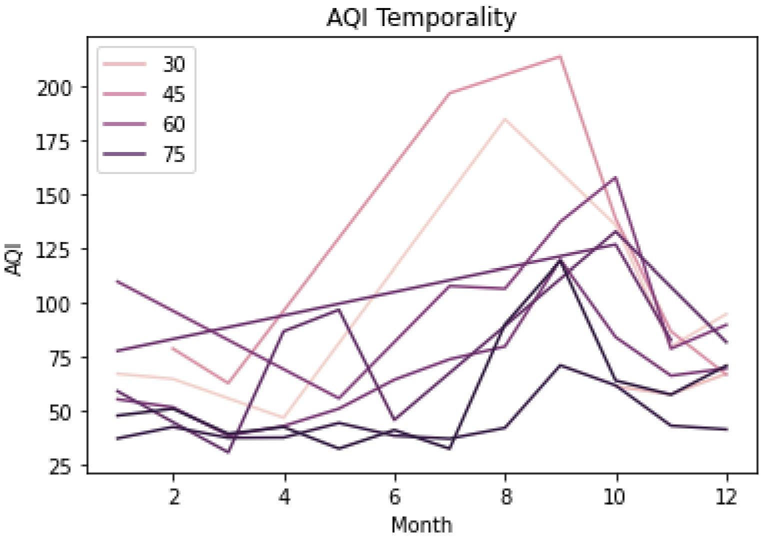
Wildfire Activity and AQI

1 Introduction

Between 2020 and 2021, California has experienced over 17,000 fires and nearly 7 million acres have burned (https://www.fire.ca.gov/stats-events/). As climate change accelerates, the risk of wildfires worsens. Along with structural damages and loss of life, wildfires create hazardous air quality events, which cause serious health concerns for vulnerable groups. This report explores the relationship between AQI, wildfire, and traffic data in order to understand the impact that fires have on air quality.

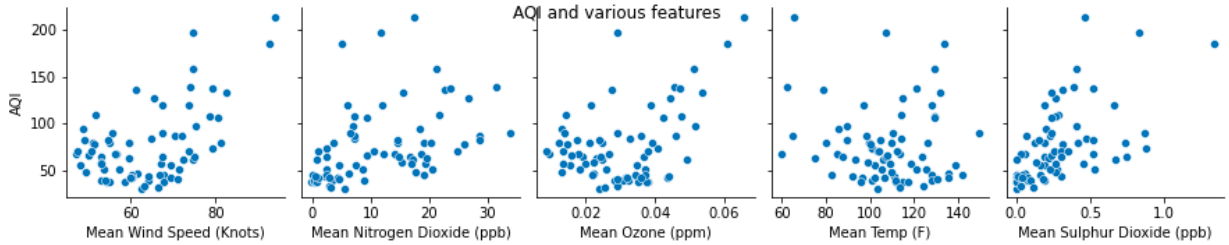
2 Open-Ended EDA

Visualization 1: AQI levels by month



Above is a line plot using the seaborn package, it displays the temporality of AQI levels (y-axis) over the months (x-axis). The largest spikes occur during peak wildfire season between the months of August and October. We were able to leverage this knowledge about the temporality by incorporating the date as a feature in our modeling portion of the project.

Visualization 2: AQI correlation with selected features



Our second visualization is a pairplot that shows the correlation between multiple features in our dataset and AQI. Features of particular interest were wind speed, nitrogen dioxide levels, ozone, temperature, and sulphur dioxide levels. Again, this insight aided us in better choosing features for our models by determining which features share the strongest correlation with AQI levels.

EDA performed:

1. The first part of EDA we conducted was a new temporal split of the data. We explored whether weekdays or weekends are associated with higher levels of AQI. Our findings proved that (as expected) weekdays displayed higher levels of AQI than weekends.
2. Another mode of exploration was to repeat EDA analysis on the new data in part two of our project. While we expected this would yield significantly different results than our EDA in part one; we in fact found similar trends as found in part 1.

Open-Ended Questions for Further Open-Ended EDA:

The EDA indicated which features may yield the best results in predictive modeling, there is more room for future EDA and investigation. Some additional questions include:

- Which particulates are most harmful to air quality? In other words, which particulates are associated with the worst AQI levels?
- In contrast, what are some features that may be associated with optimal AQI levels? For instance, proximity to forests, frequency of solar powered homes, and low level of traffic. By finding features that are correlated with better air quality perhaps we can improve AQI through subsidizing these activities/natural solutions.
- Rather than correlation, what is the causation of poor AQI?
- How is agricultural and food production associated with AQI? Merging farming/agricultural datasets could provide useful insights to this question.

3 Problem

Hypothesis and its Feasibility:

We hypothesize that wildfires are strongly positively correlated with higher AQI levels. More precisely, we believe that using features from our merged wildfire dataset such as acres burned, number of fires within a county, and duration of a fire; we can predict AQI categories with over 70% accuracy in our model. If the open modeling section yields at least 70% accuracy using the wildfire features, we accept our hypothesis. If not, we will reject the hypothesis. Not only does this dataset exist, but we were able to merge this external dataset and reach a conclusion with respect to our stated hypothesis; thus, our hypothesis is indeed feasible. In this case, our target/response variable (Y) is AQI level and our design matrix/features (X) are the multiple features from our merged wildfire dataset.

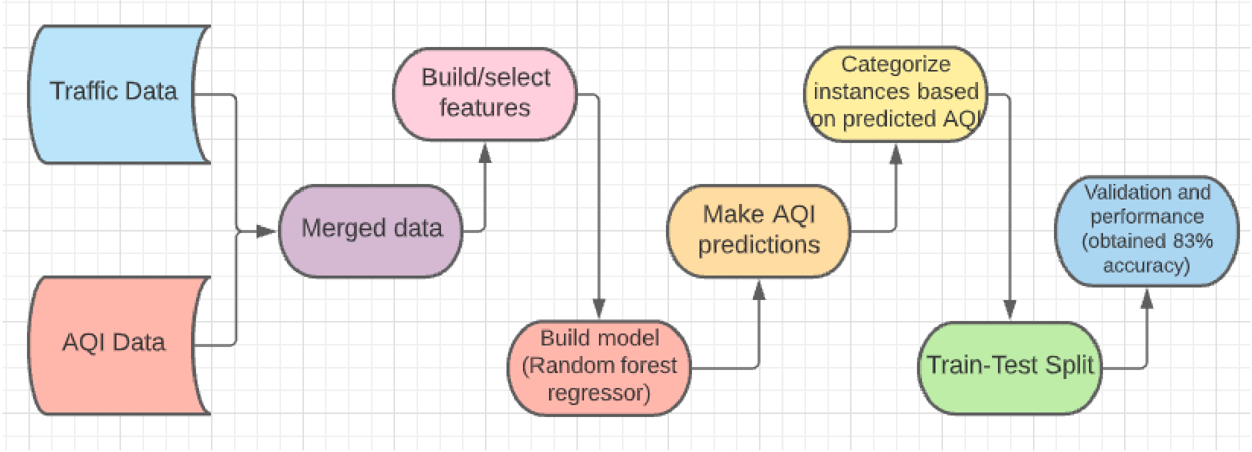
4 Answer and Interpretation

Results:

Because our random forest model in question 11 scored an accuracy below 70% we reject our hypothesis that wildfires can be used to predict AQI. However, our wildfire data in conjunction with our epa data is indicative of average monthly aqi at a given site at a given. In other words, the correlation observed in our linear model proved that there indeed is a positive correlation between AQI and the selected wildfire features.

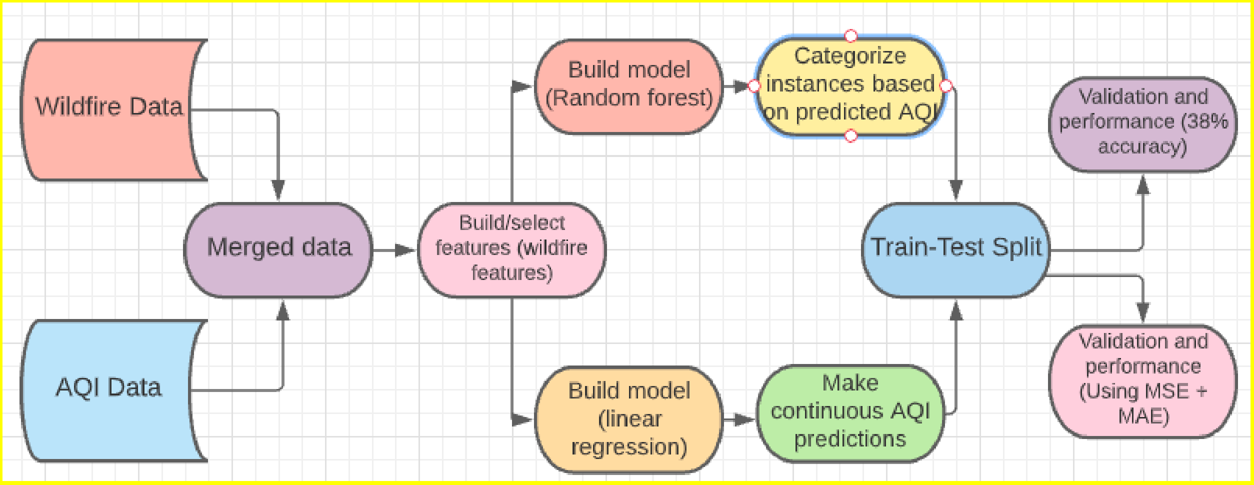
5 Modeling

Guided Modeling:


Traffic Data
AQI Data
Merged data
Build/select features
Build model (Random forest regressor)
Make AQI predictions
Categorize instances based on predicted AQI
Train-Test Split
Validation and performance (obtained 83% accuracy)

Our guided modeling first merged the traffic dataset (AADT) with our AQI data. We then executed some feature engineering and used log AADT, Temperature, NO2, Setting, and Elevation as the features for our design matrix (X/inputs). We then employed a random forest regressor model to predict AQI (Y/output). Our task; however, was to categorize our AQI predictions as good, moderate, unhealthy for sensitive groups, unhealthy, very unhealthy, or hazardous. Thus, after predicting AQI we sorted our output into these categories and checked the performance of our model using the validation/testing dataset to calculate model error.

Open-Ended Modeling:


Wildfire Data
AQI Data
Merged data
Build/select features (wildfire features)
Build model (Random forest)
Categorize instances based on predicted AQI
Build model (linear regression)
Make continuous AQI predictions
Train-Test Split
Validation and performance (38% accuracy)
Validation and performance (Using MSE + MAE)

The open ended modeling portion of the project involved introducing an external dataset containing many wildfire features which was merged upon county code of the AQI dataset. The features used in this model are median_dailyacres (median of daily acres burned in all fires for a particular county thus far); mean_FireDurationDays (average length in days of all the fires for a particular county thus far); and count_ExtremeOrActiveBehavior (the number of fires amongst all the fires for a particular county thus far which have been recorded as having extreme or very active behavior) - a binary feature. The features were selected to test the hypothesis that more severe fires would lead to substantially worse air quality in surrounding regions. The output/response variable for our model was AQI which could be categorized into different levels. With such an outstanding result from the guided modeling using a random forest regressor, we decided to again use the same model type for this section. The reason we chose to regress is because our hypothesis stated that we believed wildfire features were significantly (positively) correlated with AQI levels.

6 Modeling Evaluation and Analysis

Guided Modeling Evaluation:

The error metrics used were the binary classification error (whether the correct class was predicted) and the CV error (K Fold cross validation). The features we selected for this model are log AADT, Temperature, NO2, Setting, and Elevation. Our NO2 variable proved to be the strongest feature of our model due to it's high correlation with AQI. This is to be expected as AQI is calculated using particulate matter like N02. Initially, a linear model was used to predict

AQI but the accuracy was low even after trying different features and tuning the hyperparameters of the model. Our model yielded a binary classification score of 83% which is much higher than the baseline of 50%. That is, the model was a strong predictor of AQI.

Open-Ended Modeling Evaluation:

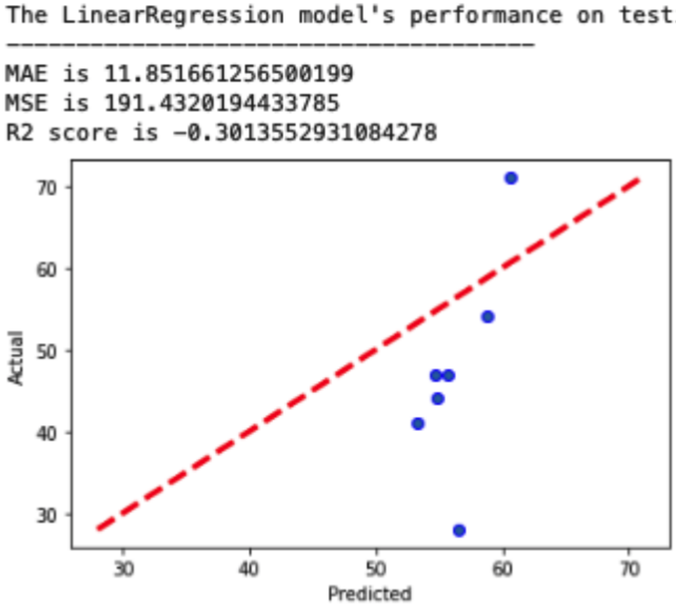
Part 1

The first model that we used was a RandomForestClassifier model. It's target variable was the annual county aqi. We calculated the accuracy of the model and displayed it below.

The second model that we used was a linear regression model that had a target variable as the annual county aqi. The metrics that were used to evaluate the performance of the model were Mean Absolute Error (MAE), Mean Square Error (MSE) and the R2 score.

The third model that we used was the RandomForestRegressor model that had the same target variable and metrics used to evaluate it as the second model that we used.

Baseline Model: The baseline model only includes the wildfire dataset and is absent of granular data like month-by-month temporal data and specific site locations within a county. The reason we used this as the baseline model was so that we could get a basic understanding of the general trends between the county and the AQI. The model did not perform well because of the lack of a large number of data points which was largely limited by the way the data was presented - granularity: each row represented an individual fire in a particular state.
The metrics and the plots for the above mentioned baseline model are shown below:
The plot below is for the Linear regression model that we implemented. The MSE and MAE are definitely low due to the fact that the range of the points is very low as well. However, it is not clear that a linear trend is correct for the actual and predicted values.



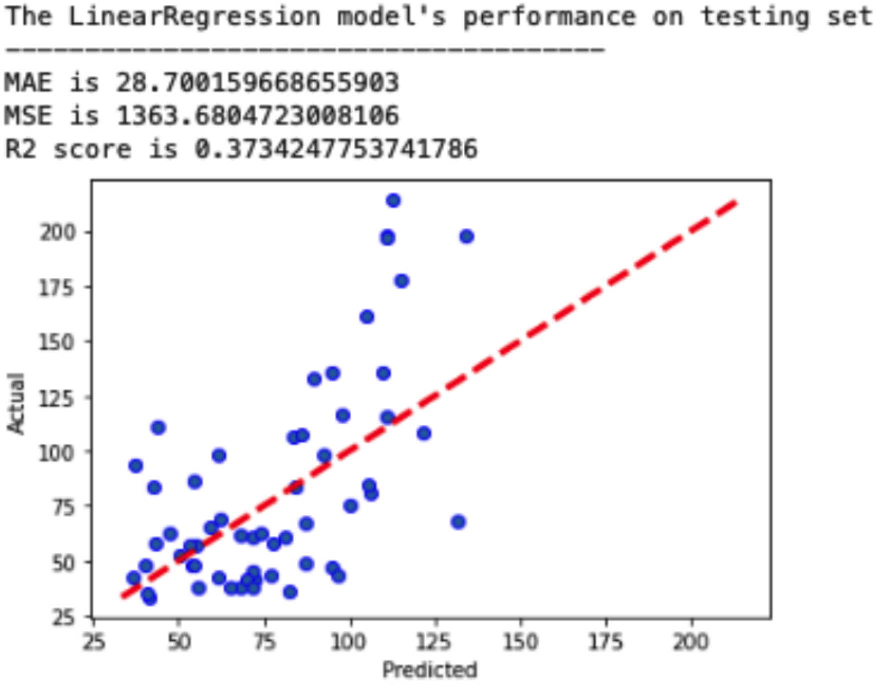
```
The LinearRegression model's performance on testing set
--------------------------------------
MAE is 11.851661256500199
MSE is 191.4320194433785
R2 score is -0.3013552931084278
```

Part 2

Improved Model: Taking into consideration the problems that we faced using the baseline model and the fact that it was not performing well with the less granular data set, we decided to use a data set with more granularity. We achieved this by merging the dataset that we gained from part 10 of the ipython notebook with the fire data set that we used in the baseline model. The way that the granularity was increased was because the dataset that we merged on was grouped by county code, site number and month. This means that for every single site in a given county, every month had a mean AQI value associated with it. This resulted in an increase in the number of data points. This also improved the model since the number of features that the AQI values depended on increased from 1 to 3. We used the same 3 models that we stated above

The metrics and the plots for the above mentioned baseline model are shown below:

The plot below is for the Linear regression model that we implemented. As can be seen, the number of points increasing definitely improved the regression line that was fit to the model. The MSE and MAE definitely increased due to the fact that the range of the points increased. However, it is much more clear now than before that a linear trend is correct for the actual and predicted values.



```
The LinearRegression model's performance on testing set
------------------------------------
MAE is 28.700159668655903
MSE is 1363.6804723008106
R2 score is 0.3734247753741786
```

7 *Model Improvement*

Model Problems:

1. Originally we implemented a linear regression model to predict AQI from our merged dataset and chosen features. Even after attempting many variations of chosen features we still lacked a reasonable accuracy score.
2. Another problem that we ran into during our modeling was underfitting. We were concerned that if we added too many features our model would have an excellent training score but output substantially lower test scores.
3. Finally, we realized that even though our model had improved after changing the model type and adding features we still had room for improvement. Specifically, we noticed that our daily traffic feature had extremely large values and may stifle our model's performance.

Model Solutions:

1. To solve our first problem we changed our model to a random forest regressor instead of a linear regression model. Our reason for doing so is that this model is better suited for non-linear data and our features lacked the clear linear relationship needed for a successful linear model. After simply rebuilding a new model (the random forest regressor) we observed approximately a 23% improvement in our AQI classification accuracy.

2. While overfitting certainly poses a threat to your model's non-training accuracy, adding features can also yield far better results when it has a larger design matrix to "learn" from. In the case of our model, adding more features actually improved both training and testing accuracy.
3. To amend this issue we conducted some simple feature engineering on our daily traffic variable by taking a logarithmic transformation. This proved to be more easily digestible for our model and again boosted both the training and test set accuracy.

8 Future Work

Further research:

Our exploration provided valuable insights regarding the relationship between wildfires and AQI. But wildfires and traffic rates are only a small piece of the puzzle in terms of what impacts our air quality. There are a plethora of other factors/features to explore with respect to air quality. One such direction we could delve into is how agriculture/food production impacts air quality in its region. Current revenue in the food industry is approaching 9 trillion dollars (https://www.statista.com/outlook/cmo/food/worldwide) and is expected to grow about 4.5% in the coming years. The wide scale production of food comes at a cost; this cost comes in the form of the animals' waste and the food they need to eat, the packaging of food, and the transportation needed for delivery. The amalgam of these threats from the food industry likely harm the quality of our air. This particular direction for further research is interesting because the harmful effects of the food industry on the environment are widely discussed but its direct relationship to air quality lacks the necessary data analysis.