Using a Temporal Fusion Transformer in AQI Forecasting

Naveen Sukumar

Shuwei Li

Pin-Lun Hsu

December 13, 2021

The notebook for this writeup is runnable on Colab. Please click the button below to open it.

Introduction

AQI (Air Quality Index) is an important factor to measure the quality of the air and assess if the air is safe for humans to breathe.We use pollutant concentration mainly, as well as other features, to predict AQI. However, we discovered through analysis and research that many areas lack sufficient equipment to detect pollutant levels. To help tackle this important real-world issue, we wanted to build a forecasting model that can predict future AQI without relying on any pollutant concentration levels or previous AQI values.

Primary Exploratory Data Analysis

a. Air pollution is closely related to health ailments and premature death in the world. Across all the states in America, the state of California has the worst annual median AQI across all US states. We analyzed the AQI data of the year 2020 from the US EPA dataset to understand air pollution in California. Several counties in California have missing dates in the 2020 AQI data, but there is no similarity among different counties. Del Norte and Trinity have the most missing days. These missing dates occur in blocks of days throughout the year which indicates that this may have happened due to equipment error and malfunctioning measurement instruments.

b. When comparing monthly AQI across counties in California, the monthly average is relatively low in the first half of the year 2020, but the average AQI in October, September, and October are relatively high, especially in September. A reason for this might be because AQI has some association with temperature which is relatively high during these months. Wildfires often begin during these months as well which might also explain the high average AQI in these months. To analyze the differences and similarities in climate patterns across different counties, we made a line plot where we plotted the Median AQI values over 12 months (Figure 1). In this line plot, we see that the increases and decreases of Median AQI values follow a very similar pattern. This hints that time is very correlated to AQI and that the months could be a very beneficial feature to have in a model. Also, there could potentially be clusters of counties with similar AQI values and patterns.

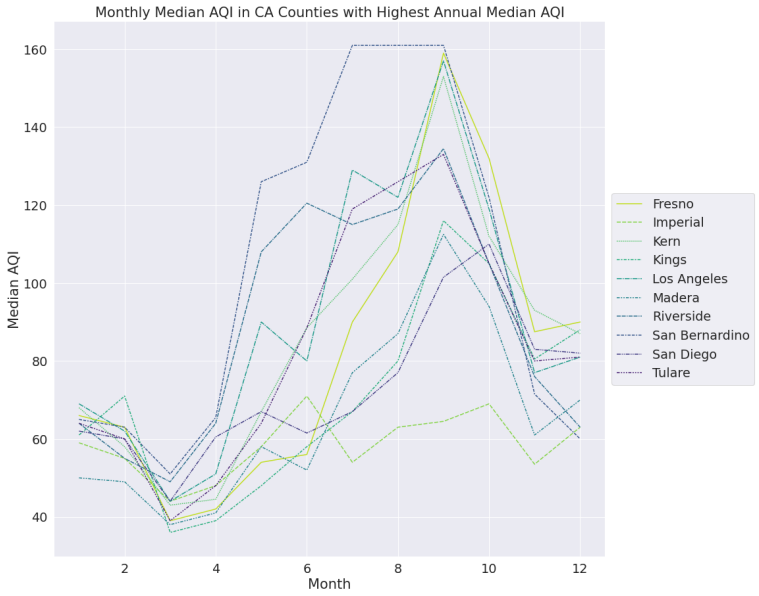

Figure 1: Monthly Median AQI in CA Counties with Highest Annual Median AQI

c. To understand the geographical distribution of air pollution in the month with the most air pollution, September, we generated a heat map to visualize the geographical distribution of AQI across California in Figure 2. We found that the median AQI in September is not spread uniformly across California. Darker regions are the areas with relatively high yearly mean AQI. These regions correspond to areas with a relatively dense population or relatively dry climate. This suggests that heavy traffic or wildfires in these areas might be associated with these high average AQI values.

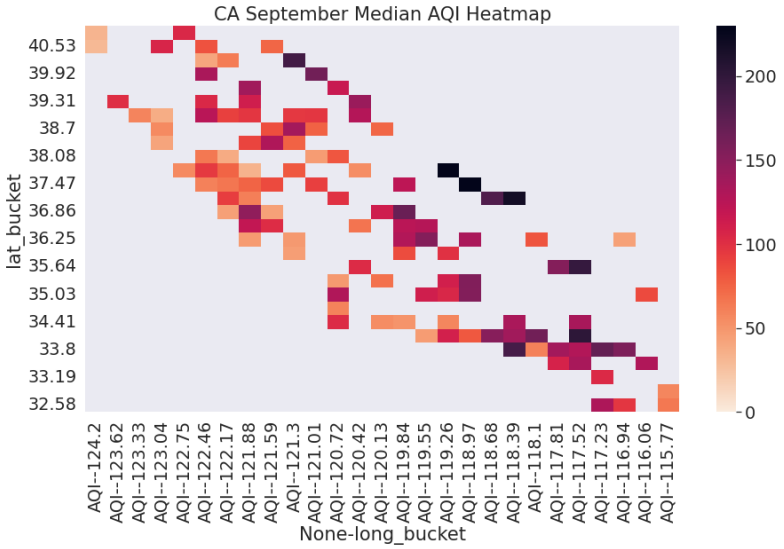

Figure 2: CA September Median AQI Heatmap

d. AQI is based on the five "criteria" pollutants: ground-level ozone, particulate matter, carbon monoxide, sulfur dioxide, and nitrogen dioxide. To determine the correlation between the concentration of these compounds and the AQI, we pair-plot the correlation between different air pollutants and AQI in FIgure 3. The result clearly shows that some pollutants possess a close

relationship with the AQI. This can help us identify pollutants that we can use as features to best improve the accuracy of our model. The pair plots also revealed that our model should not necessarily be using every pollutant as a feature. CO and NO2 have an extremely high correlation coefficient for example. This could indicate that our model should be using one of these two compounds as a feature since having them both could cause our model to have redundant features.

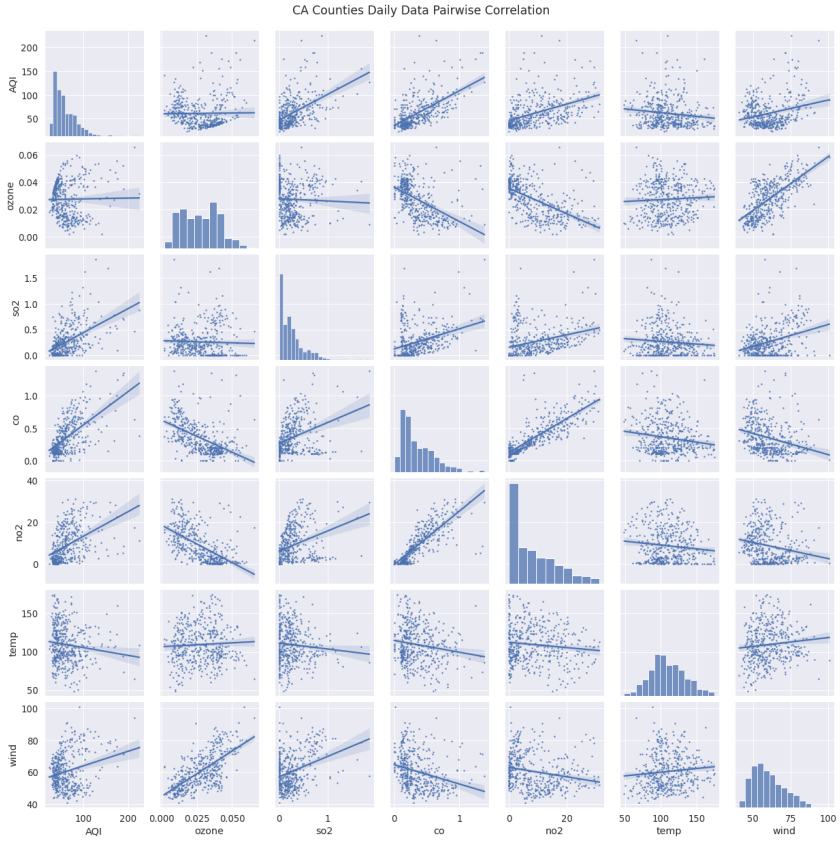

Figure 3: CA Counties Daily Data Pairwise Correlation

e. We also performed principal component analysis on each CA county's monthly mean AQI. The first two principal components in PCA explained 84% of the variance, indicating we can project the data into a two-dimensional space to identify the similarity of AQI changing patterns in CA counties. After the projection (Figure 4), we can identify a cluster of counties sharing a similar AQI changing pattern, including Mendocino, Colusa, Solano, Yolo, Sonoma, Marin, San Mateo, San Francisco, Napa, Santa, San Luis Obispo, etc. We can also identify counties with very different AQI changing patterns such as Mono and San Bernardino. This suggests we might build different models for different groups of counties in CA.

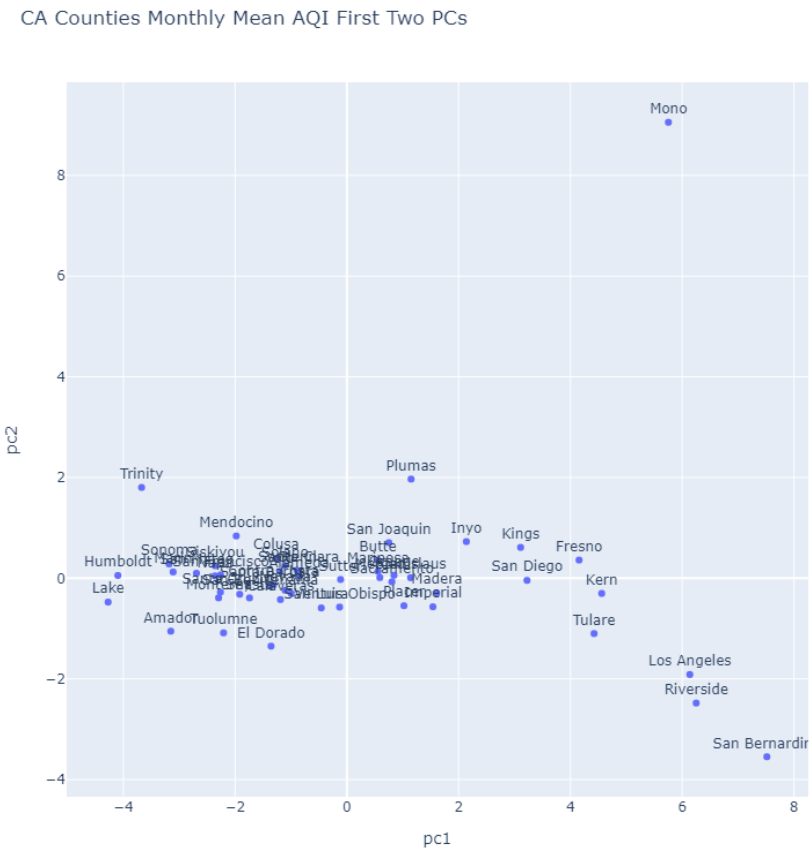

Figure 4: CA Counties Monthly Mean AQI First Two PCs

f. We also wanted to analyze the differences and similarities in climate patterns across different counties. For each county, we plotted the mean temperatures against the mean wind speeds of each month in Figure 5. We also set the hue of every data point in the scatter plots to represent the AQI levels. In most counties we can see there seems to be a positive association between mean temperatures and the mean wind speeds.

Figure 5: CA Counties Monthly Pct Pollution Days by Mean Temperature and Wind Speed

Open-ended Questions

a. Is there a correlation between AQI levels and COVID-19 cases such that we can use COVID-19 cases as a feature in our model to improve accuracy? Will the shelter in place order have an effect on AQI because traffic will be reduced?

b. Will AQI differ in different climate types? Will areas with Mediterranean climate type have relatively lower AQI compared with regions with other climate types?

c. How does AQI differ among different months for regions in each climate type?

As we explored the relationship between our internal datasets and external datasets, we conducted **Further Exploratory Data Analysis** to address some of our open-ended questions in the next section.

Data Collection and Cleaning

1. External Datasets

Gas consumption: From California Energy Commission, the dataset contains the gas consumption of a county from 2016~2020. We assume that gas consumption can indirectly impact air quality. Each column contains gas consumption for a county in a year.

Wildfire: From California Natural Resource Agency, the dataset consists of wildfire records in California. The columns useful for us are `ALARM_DATE` and `FIRE_NUM`. By aggregating the fires that occurred on the same date together, we can find the total number of fires on every date. We decided to do this because the frequency of wildfires can have a large effect on the pollution concentration levels in the atmosphere.

Climate type: From climates to travel, they categorize counties in CA into different types: the Mediterranean, continental, semi-arid, etc. We hypothesize that different climate types should generally have different air quality.

PM2.5: tiny particles in the air that reduce visibility and cause the air to appear hazy when levels are elevated

Barometric Pressure: the pressure within the atmosphere of Earth.

Dew Point: related to the humidity of the air.

For the data preparation, we aim to build a data frame with labels, internal features, and external features.

Labels	1. AQI (air quality index) 2. AQC (air quality category)
Internal features (existent in original dataset)	1. Ozone 2. SO2 3. CO 4. NO2

	5. Temperature 6. Wind 7. Date 8. State and County
External features (crawled from external sites)	1. PM2.5 (EPA) 2. Barometric Pressure (EPA) 3. Dew Point (EPA) 4. Gas (CEC) 5. Climate Type (Climates to Travel) 6. Wildfire (CNRA)

Table 1: Variables for Supervised Learning

The label includes AQI value and Category, which can be extracted from *daily_county_aqi*. This category is the main target that we want to predict in guided modeling.

The internal features refer to the ones already existing in the part-1 dataset, including Ozone, SO2, CO, NO2, Temperature, Wind, Date, state, and county.

Moreover, we considered that one year of data might not be enough to train a good model, so we expanded our data set to include data from 2016-2019 from the EPA (Environmental Protection Agency) website. Besides the internal features and labels, we also collected new features (external features) from the website, including PM 2.5, Barometric Pressure, and Dew Point.

We also collected data on gas consumption from the California energy commission, wildfires from California Natural Resource Agency, and climate type from Climates to travel.

2. Data Cleaning

In the first step, we have a lot of tables that need to be merged. They can be represented as

1. daily_county_aqi_2020, daily_ozone_2020, daily_no2_2020 ...
2. daily_county_aqi_2019, daily_ozone_2019, daily_no2_2019 ...
3. …
4. ...
5. daily_county_aqi_2016, daily_ozone_2016, daily_no2_2016 ...

We use left_join to keep merging features with AQI by each year. After this operation, the data of each year will be aggregated together. It will become like the following:

1. daily_all_data_2020
2. daily_all_data_2019
3. …
4. ...
5. daily_all_data_2016

In the second step, each row contains a merged table from a year, and we concatenated them together. In the end, the table will include the data from 2016~2020 with all features.

3. Missing and Extreme Values

After the data cleaning stage, there are still several cells with the value of NaN. We filled them with the mean value of the feature. Later, we normalize the feature value by using *Normalizer* from scikit-learn. Moreover, we removed records containing AQI values over 1,000 since these extremely large values were outliers that could have biased the results of our model. Also, the improved model removes some outlier counties because they have very different AQI patterns unlike others (from PCA).

4. Transformation

There is no clear relationship between AQI and other features. This indicates that under the assumption of not using air pollutants, the relationship between AQI and other features is complex and we might use a relatively complex model for AQI prediction.

Further Exploratory Data Analysis

1. California AQI Trend 2016 - 2020

We aim to find the pattern of AQI by plotting the AQI in 5 years. Although we have already removed outliers in AQI values. It is still hard for us to find its pattern, which indicates that it might be a challenging task to do forecasting on AQI value.

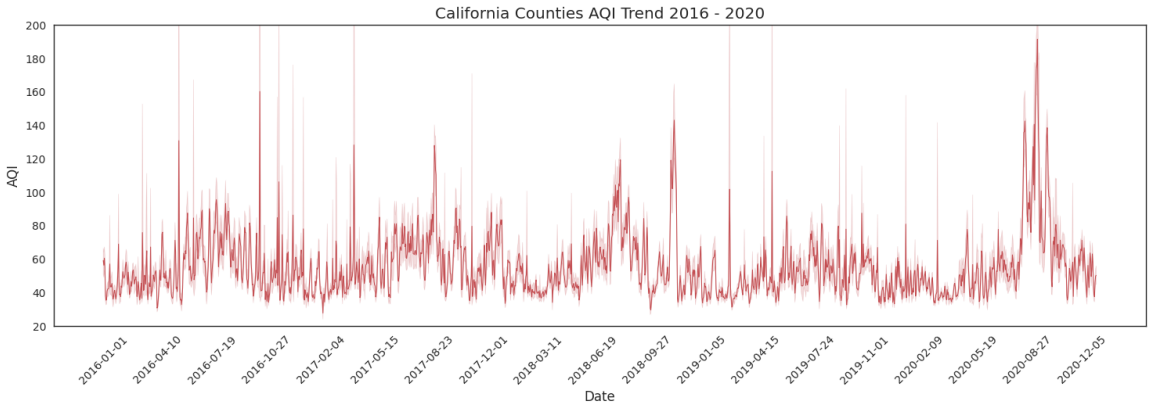

Figure 6: California Counties AQI Trend 2016 - 2020 with 95% CI

2. California AQI vs. COVID-19

We assume that COVID-19 should have something to do with the AQI value. The reason behind this is that the shelter in place order might affect the traffic and some productions causing the

emission of air pollutants. Thus, we plotted the AQI value before and after the shelter in place order, March 11, the 80th day of the year 2020, in California and compared it with the same period of the last year, 2019. However, we cannot find a significant impact by COVID-19 either before and after the order is in effect in 2020, or between the similar periods of 2019 and 2020.

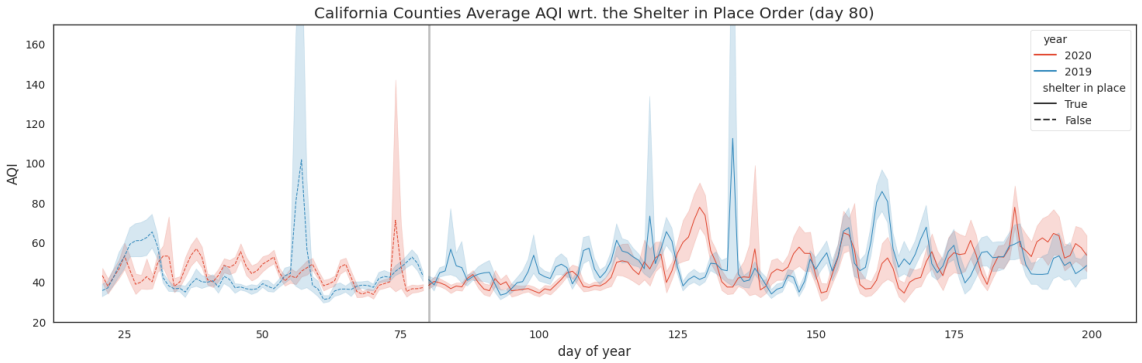

Figure 7: California Counties Average AQI wrt. the shelter in place order with 95% CI (day 80)

3. Los Angeles AQI vs. COVID-19

Because the AQI over the entire California counties has more variation, larger cities with high population density may be more impacted by COVID-19 as there is more traffic. We generated the following plot to show the AQI trend of Los Angeles before and after the shelter in place order and compare it with the similar period of the last year, 2019.

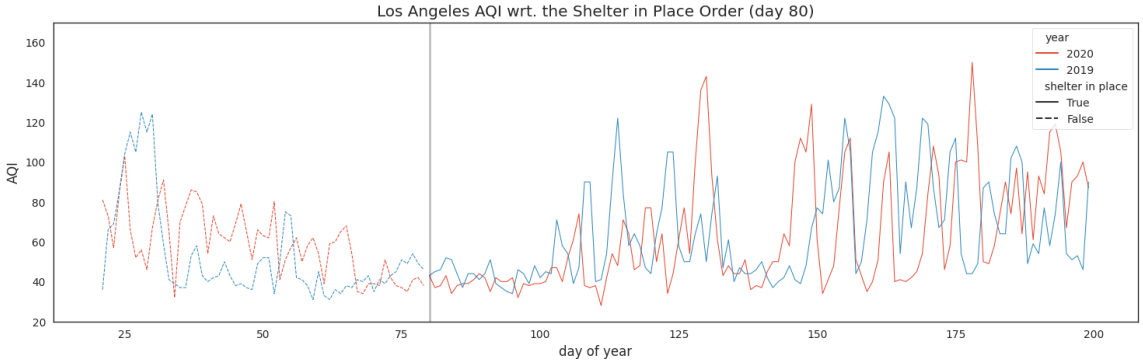

Figure 8: Los Angeles AQI wrt. The Shelter in Place Order (day 80)

4. Weekly Median AQI by Climate

In the following figure, it shows the AQI for different climate types. We can notice that for continental climate, the AQI will change most drastically. For the other three types, they do not fluctuate so much. It explains that climate type might have something to do with AQI.

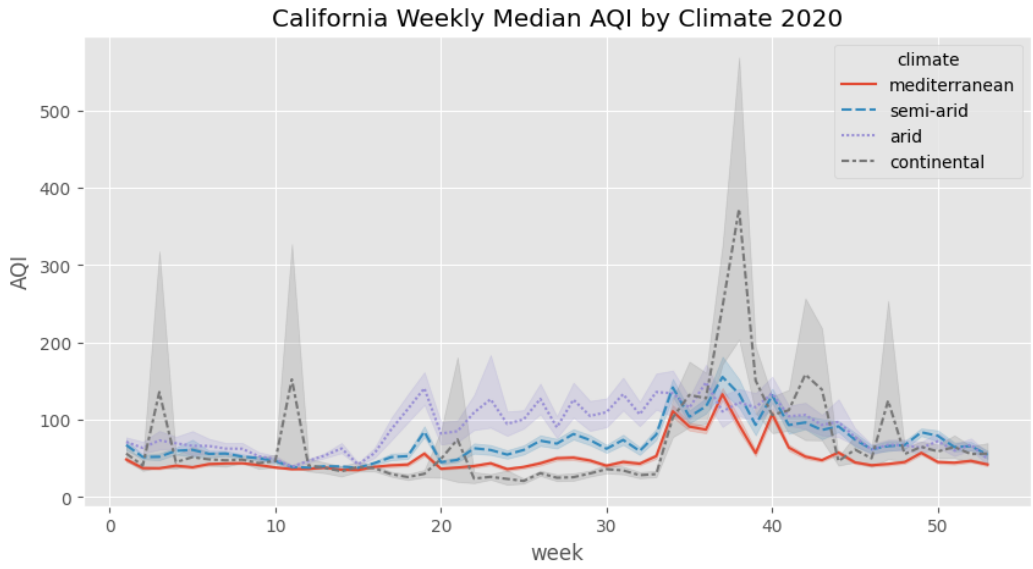

Figure 9: California Weekly Median AQI by Climate Types with 95% CI

5. County Monthly AQI

These four counties stand for different climate types. We plot the monthly AQI distribution by state, and we notice that the month is highly related to AQI value. For example, in summer, the AQI tends to be much higher than in winter.

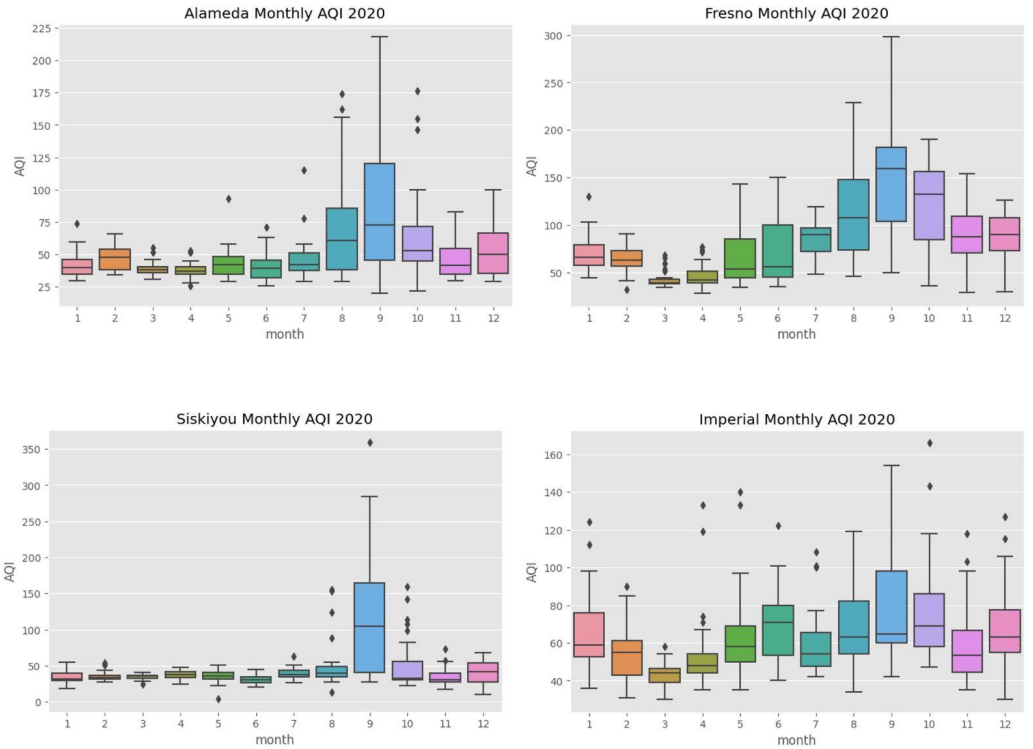

Figure 10: California Counties with Four Different CLimate Types Monthly AQI

Hypothesis

Our initial hypothesis was that AQI can be predicted more precisely if we also take weather patterns, such as temperature and wind speeds, into consideration to train a relatively accurate model. Although we are moving forward with the premise of this hypothesis, we have decided to create a new hypothesis that explores how much of an impact features from external data sets can have on the accuracy of our model. Also, we wanted to make our tasks more challenging. During our EDA we found that some regions did not have recorded AQI values. We wanted to answer the question of whether we can train a model that can accurately forecast AQI values for these regions as well given no prior AQI data. Therefore we revised our hypothesis to:

> "With the integration of creative features, gas consumption and climate types, from external datasets and **without** relying on any previous AQI or air pollutant concentrations levels, future AQI can be forecasted with at most 15% symmetric mean absolute percentage error (SMAPE). Gas consumption and the frequency/periodicity of wildfires will be the two most important features that contribute towards the accuracy of our model via a 'feature importance score' computed from our final model."

We will accept this hypothesis if the SMAPE of our model is at most 15% and gas consumption and the frequency/periodicity of wildfires are chosen as the two most important features from the 'feature importance score.' We have all needed data to verify this hypothesis as explained in a section above.

Feature and Model Selection

In this section, we investigate which features and models we should apply in the forecasting model. We trained a random forest regressor and inspected which features contributed the most. Based on this information, we selected the top 5 features as our input in the forecasting model.

Regarding the model selection, we tried a wide variety of model types and we found that MLP (MultiLayer Perceptron) and Random Forest have the best performance. Therefore, we realized that we can use a neural network model and ensemble technique to construct our forecasting model. Based on these observations, we thought that a transformer is a good fit for our case.

1. Feature Selection

We set a constraint for ourselves: we will not use AQI or directly related AQI features (co, ozone, pm2.5 …etc) in our forecast model because some regions have no AQI due to the equipment trouble, but we still pursue strong accuracy in the model.

To select which features we should choose, we build a random forest regressor to predict the AQI value, and then we extract the feature importance from the model. The results shown in the chart below logically make sense. The first most important feature is wind. This logically makes sense because higher wind speeds disperse and spread air pollutants throughout the atmosphere, which

could lead to lower AQI levels in the atmosphere. The second most important feature that contributed towards predicting AQI is the number of wildfires occurring over a period of time. This also logically makes sense because wildfires release large amounts of carbon dioxide into the atmosphere which can cause AQI levels to drastically increase. The third most important feature that contributed towards predicting AQI is relative humidity and dewpoint. This is because higher humidity levels in the atmosphere can increase the levels of noxious chemicals in the atmosphere, effectively increasing AQI levels. A notable feature worth mentioning is gas consumption. The relatively big effect that gas consumption has on our model can be because since gas consumption is linked to the usage of vehicles, more gas consumption can lead to the use of more transportation, which causes AQI levels to rise since gas-powered vehicles emit pollution in the atmosphere. As we can observe in the figure, ‘wind’, ‘num wildfires’, dew point’, ‘barometric pressure’, and ‘longitude’ have the highest feature importance, so we would take these 5 features into our forecast model.

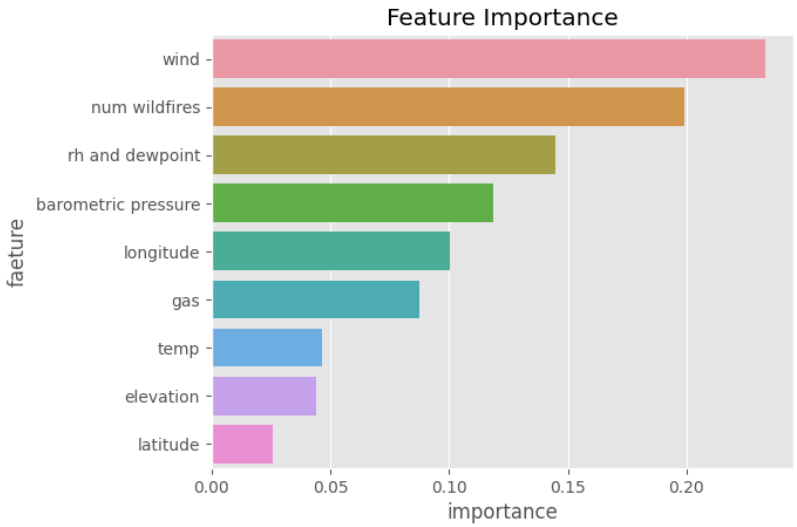

Figure 11: Feature importance by random forest regressor

To check the possibility of dimensionality reduction, we performed PCA for the most important 6 features that we selected. This shows that it is possible to decrease the feature dimension from 6 to 4 while remaining around 95% of the variance. But the computing costs of PC projection and dimensionality reduction are also high and therefore might not be worth it.

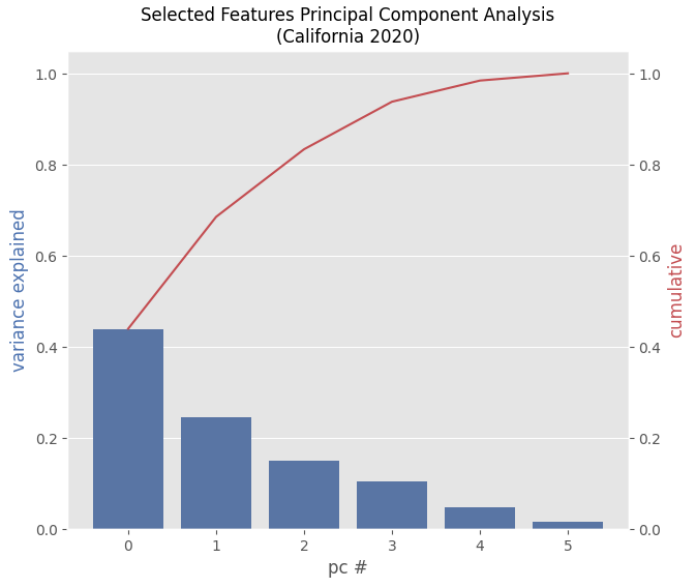

Figure 12: Selected Features Principal Component Analysis

2. Model selection

After selecting those features, we observe which models can achieve the best performance. Air quality categories (AQC) is an “ordinal qualitative variable” as defined in the table, and AQI value is a “continuous quantitative variable”. We propose two kinds of models to predict AQI. The first (classifier) will directly output the AQC and the second (regressor) will directly output the AQI value which would then be mapped to the category. To elaborate, we will use the following AQI-AQC conversion table to map a regression result to an AQI category. Because we believe overestimating and underestimating AQC are equally bad, we used real mapping.

AQC	AQI
Good	0 to 50
Moderate	51 to 100
Unhealthy for Sensitive Groups	101 to 150
Unhealthy	151 to 200
Very Unhealthy	201 to 300
Hazardous	301 and higher

Table 2: AQC-AQI Mapping

We listed 8 kinds of models including 4 classifiers and 4 regressors.

Model #1	LogisticRegressionClassifier
Model #2	MLPClassifier
Model #3	RFClassifier
Model #4	KNClassifier
Model #5	LinearRegression
Model #6	MLPRegressor
Model #7	RandomForestRegressor
Model #8	KNeighborsRegressor

Table 3: Models to be Compared

Model	Training Binary Accuracy	Testing Binary Accuracy
LogisticRegression	0.8508	0.8231
MLPClassifier	0.9934	0.827
RandomForestClassifier	0.7128	0.6972
KNeighborsClassifier	0.9913	0.8094
LinearRegression	0.7527	0.7424
MLPRegressor	0.8882	0.7864
RandomForestRegressor	0.9282	0.834
KNeighborsRegressor	0.9995	0.8019

Table 4: Model Accuracy

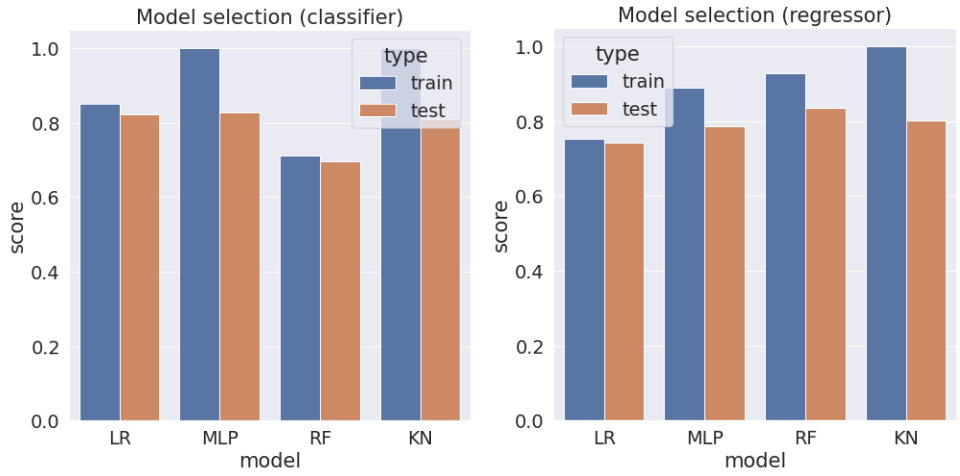

Figure 13: Model Accuracy

We found the MLP Classifier and the RandomForest Regressor have the highest test score. Thus a model combining the advantages would yield the best result. Because MLP achieves good scores, we can consider building our forecasting model based on a neural network. We can also utilize ensembles in our forecasting from Random Forest Regression.

The transformer is a well-known model for time-series prediction so we decided to use this model. Random forest allows us to train different models of the same type and intercorporate their results to get a better prediction. To utilize this idea, we decided to use an ensemble together with the transformer.

Metric

Our focus is on AQI. We will be using symmetric mean absolute percentage error (SMAPE) as the metric for evaluating the performance of our model.

$$\text{SMAPE} = \frac{1}{n}\sum_{t=1}^{n}\frac{|P_t - A_t|}{(|A_t| + |P_t|)/2}$$

We decided to use this metric because it is a better metric for measuring the inaccuracy of our model, as well as the significance of this inaccuracy, which metrics such as mean squared error are not good at conveying. For example, if the real AQI is 200 and the predicted one is 210, the mean squared error is 100. If the real AQI is 0 and the predicted one is 10, the mean squared error is also 100. However, it is apparent that in the latter one, the significance of this difference is larger. The formula for symmetric mean absolute percentage error is shown above.

Modeling

Our baseline model consists of two features, the county feature (which is the geometric information of a record of a country formed by concatenating the state code with the county code) and the time index feature (which is a discrete numerical way of identifying dates). We decided to do this because both of these are vital fundamental features that can help the model use basic information about counties and the dates that were significant to make important predictions.

For our model, we have decided to use a temporal fusion transformer. It relies on deep neural networks for effective forecasting. The temporal fusion transformer will utilize input data consisting of the features from 7 days to predict the AQI levels, which is the output of the model, for an average of two days in the future. More detailed information about the inputs and output models is in the Transformer Architecture section. The temporal fusion transformer allows us to extract information, which if interpreted correctly, can show us which factors contribute towards influencing the predictions. Since our hypothesis revolves around exploring which features contribute the most towards building an accurate model and whether the use of external data sets is needed given the added complexity, space, and runtime that they add to the model, we felt that the temporal fusion transformer was the best model to use.

1. Constructing a Time Series Dataset for Supervised Training

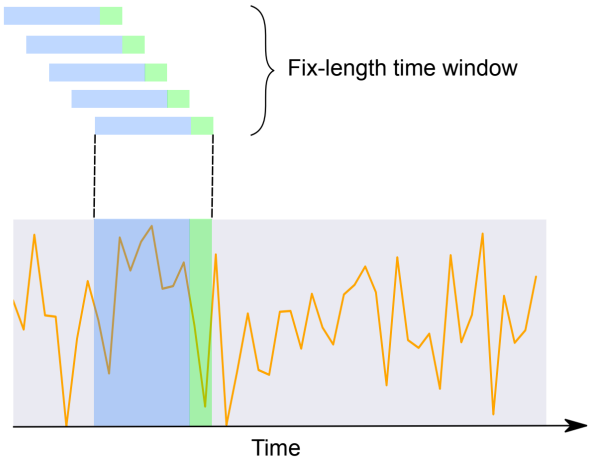

Image credit: https://arxiv.org/pdf/2001.08317.pdf

To prepare the training and validation data, we will split one country's entire data of the five years into small fix-length time windows, each containing a five-day history and two-day prediction targets. We then use our model to fit the data. The last 2 days of our entire training data are used for validation. Because using real data in reality to test the model performance, we do not contain a test set in our modeling process.

2. Temporal Fusion Transformer Architecture

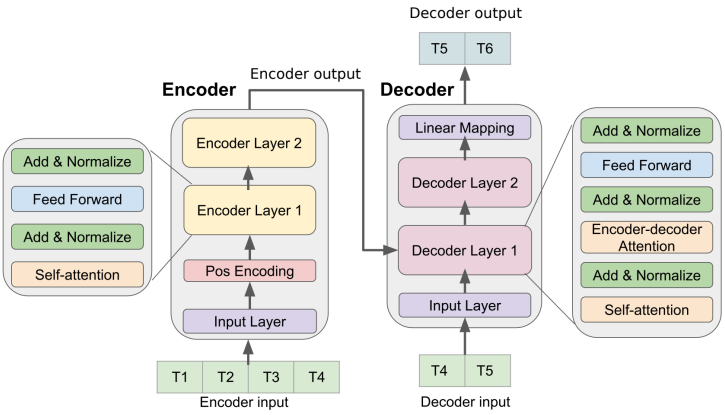

Image credit: https://arxiv.org/pdf/2001.08317.pdf

The parameters used in the temporal fusion transformer can be described as encoder variables, decoder variables, and static variables. Static variables are variables whose values do not change over a while. Encoder variables are variables that will serve as inputs and features to our model. Decoder variables are variables that our model can use to predict AQI. In our baseline model, county, which refers to the unique identification numbers of counties, and time index, which refers to the dates of all our records in our datasets, will be our static variables while in our final model, county, time index, and elevation will be our static variables. In our baseline model, time index will be our encoder variable while in the final model, wind speeds, gas consumption, climate patterns, barometric pressure, PM 2.5, and the frequency of wildfires over some time will be our encoder variables. In our baseline model, the decoder variables are the time index while in our final model, the decoder variables are the time index, and gas consumption.

In our experiment, following the idea of ensembling, we constructed 5 transformers, each of which are trained on 30 epochs, and used CA counties with full AQI data on all dates to test our model on and compute the SMAPE error value. We will then average the prediction of the five models and call the averaged result “final model” prediction.

Model Improvements

1. Problem: Some counties are predicted especially badly from the residual plot of the baseline model. The first steps involved how we could improve and optimize the training data that we already are using in the baseline model. Solution: In Part 1, we conducted a Principal Component Analysis and noticed a couple of outliers in our data. Specifically, we noticed that the counties LA, Riverside, Mono, and San Bernardino were outliers in our data (Figure 4). These outliers can cause the predictions of our model to be skewed. We removed any records containing these values to improve the accuracy of our model.

2. Problem: The SMAPE of our baseline model is not significantly lower than 15% hence we should consider how we can provide more features for the transformer. The second step involved how we can utilize new features from external data sets to further improve the accuracy. Solution: Since the baseline model only contains two features, state_county and time_index, which are features that logically wouldn't seem to have a correlation to AQI as high as features that are about environmental data, we wanted to add new features to our model to try to improve accuracy. Our improved model will consist of ten features, two of which are from the baseline model, and the others will be climate patterns, temperature, wind speeds, gas consumption, longitude, latitude, elevation, and the number of wildfires happening over a period of time. By using these features, some of which are from the AQI data set, and some of which are from external data sets, our improved model can help us improve the accuracy of the model to a somewhat significant level by finding and interpreting stronger patterns.

3. Problem: The introduction of new features increases the complex of the transformer, and some hyperparameters should be adjusted accordingly. The third step we took was adjusting hyperparameters in the transformer. Increasing the amount of features in our model can cause the trends and relationships in our training data to be very complex. This can cause the model to have a tough time fitting the training data. Solution: Thus, we decided to increase the number of hidden layers in our neural network model from 16 to 32. Increasing this amount will allow our model to better fit the training data and understand more complex relationships. We also found that a learning rate of 0.05 and a dropout rate of 0.3 produces the lowest training loss, which can be achieved by increasing the number of hidden layers from 16 to 32.

   The results of those improvements are reflected in the improved model in the next section.

Evaluation and Analysis

1. Baseline Model

To evaluate the model performance, we use the previous five days' AQI to predict the later two days' AQI. To visualize the prediction, we can see the following figure: the gray dot with solid line indicates the observed AQI value for the previous five days, the red dot with dashed line indicates the predicted AQI for the last day, and the gray dot with dashed line indicates the actual AQI for the last day.

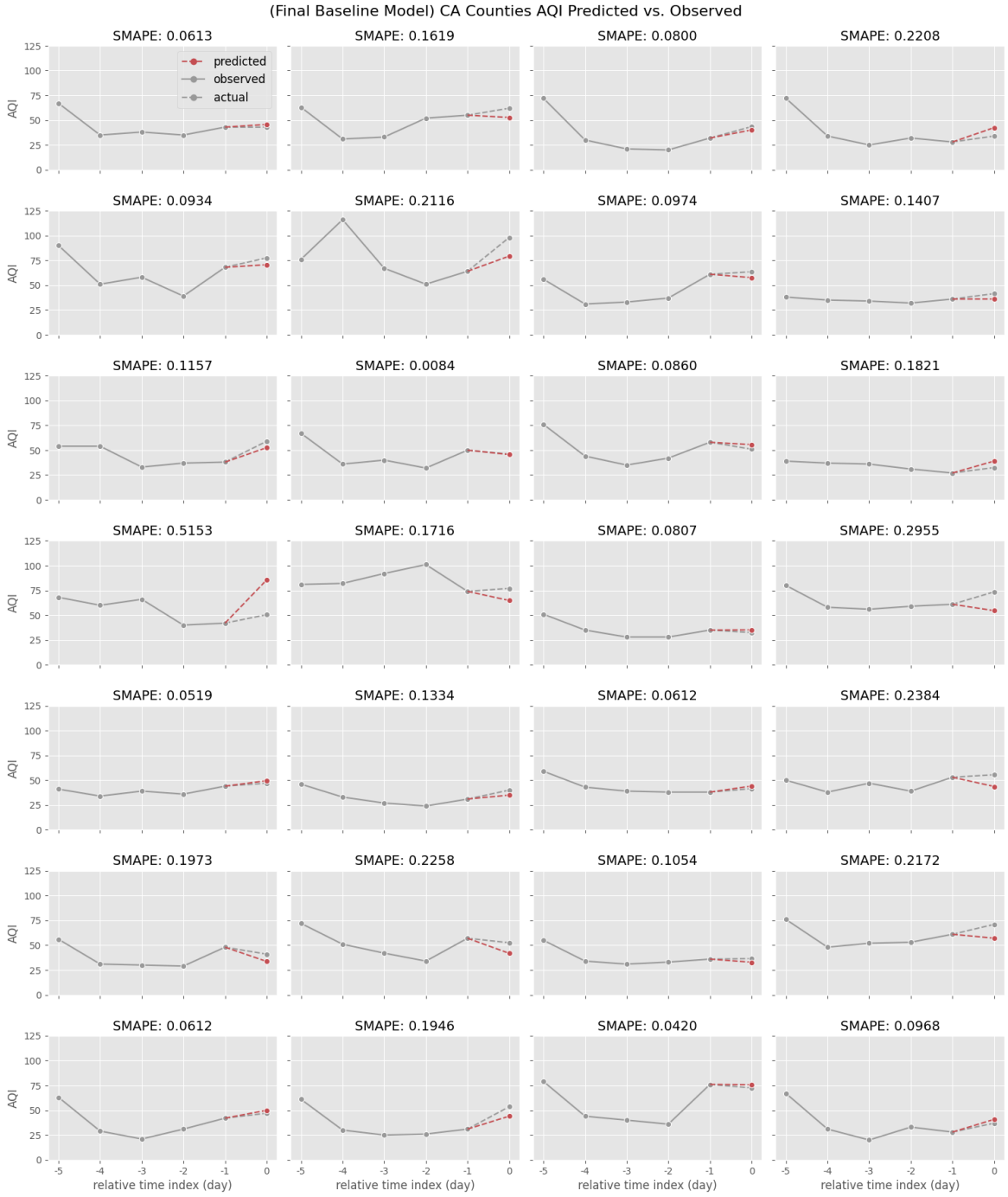

Figure 14: (Baseline) CA Counties AQI Predicted vs. Observed

For every plot above, we calculated the SMAPE, which is shown above every plot. Looking at the plots, we can see that the red dot is very close to the gray dot on the last day which shows that the predicted AQI values are relatively close to the actual AQI values. To understand the distribution of the predicted and actual value for each period, we constructed a scatter plot and calculated the mean SMAPE error.

If the value is predicted perfectly, it will lie on the line `y=x`. We can observe that around half or more of the points lie somewhat close to `y=x`. Some outliers are difficult to predict.

The second plot is useful in helping to interpret the first plot. The second plot shows the residuals and the calculated residual squared sum, which is 3162.97. This value is an indicator of the variance and how closely our model fits the data. Since the RSS is quite large, it seems that this is not the case.

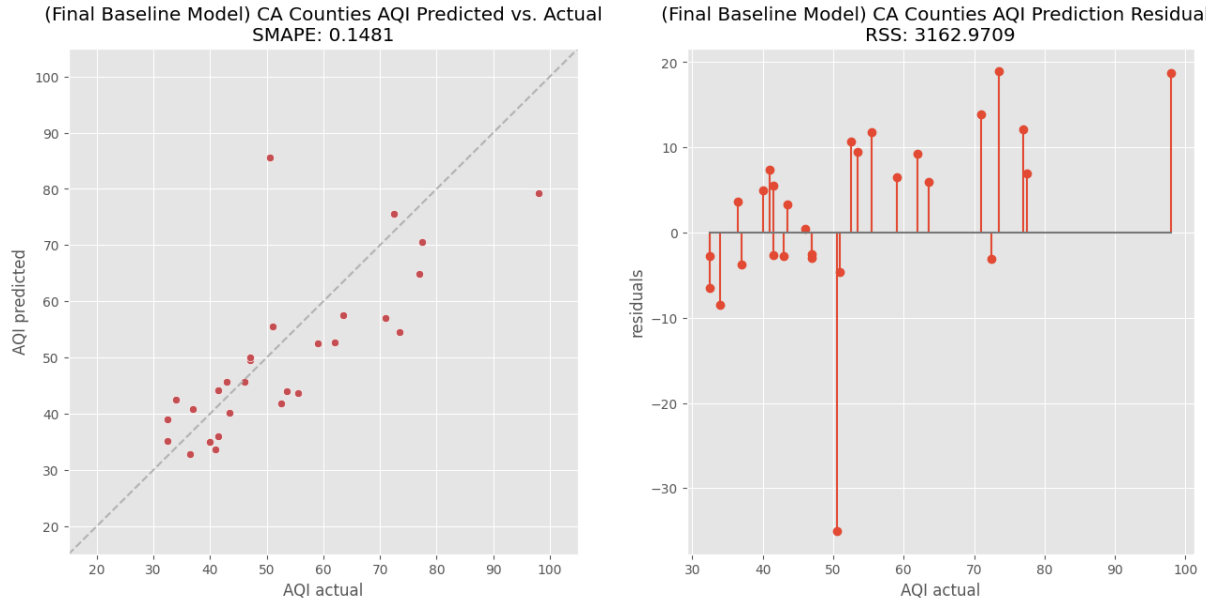

Figure 15: (Baseline) CA Counties AQI Predicted vs. Actual & Residuals

2. Improved Model

After we applied the improvement technique mentioned in the previous section, our model improved significantly, and we successfully reached our hypothesis with under 15% SMAPE error. (as the following figure)

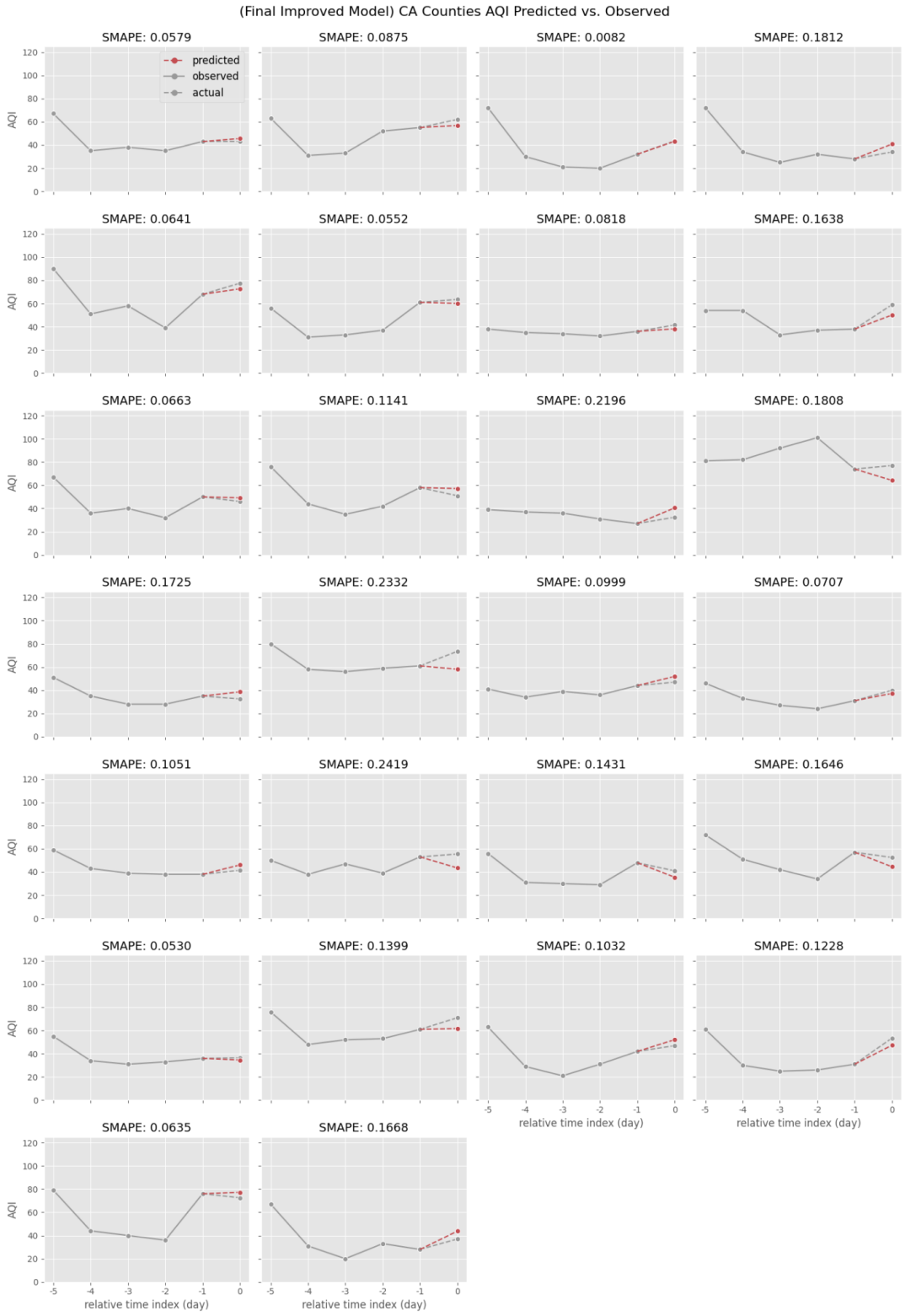

Figure 16: (Improved) CA Counties AQI Predicted vs. Observed

In the plots above, we can see the same pattern that we saw in the same plots for the baseline model, except that the predicted AQI values look a lot closer to the actual AQI values. This shows that there has been an accuracy improvement.

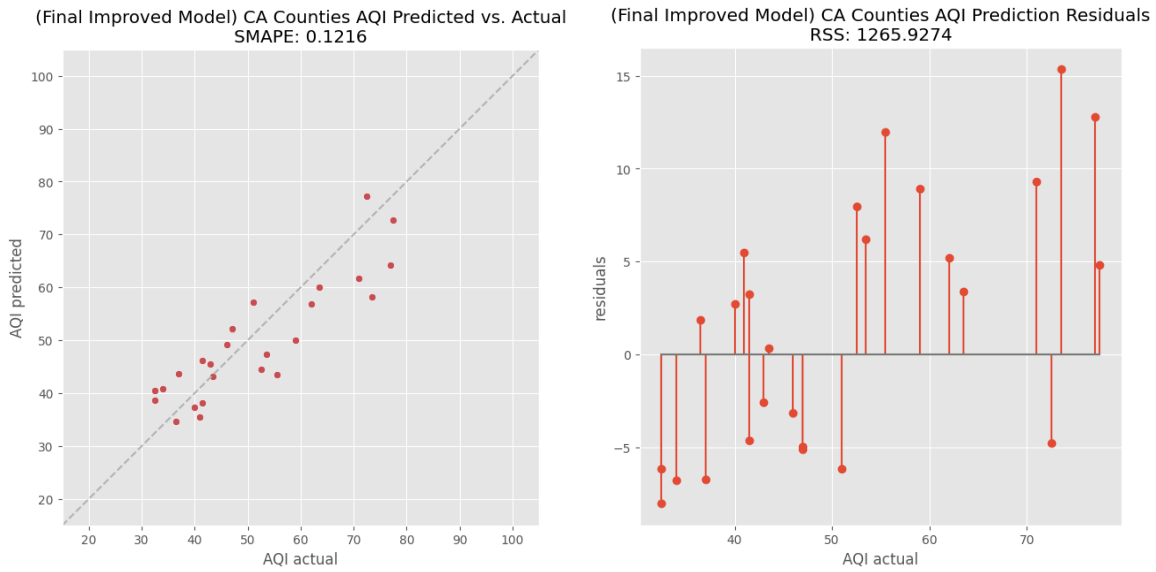

Figure 17: (Improved) CA Counties AQI Predicted vs. Actual & Residuals

In the first plot, we see that the points seem to be a lot closer to the y = x line. This shows that there has been a decrease in the gap between the predicted and actual AQI values. Moreover, the residual plot shows that there has been a decrease in RSS by around 2000 which further proves that the points in the first plot are much closer to the y = x line than the points in the first plot for the baseline model.

3. Improved Model Transformer Variable Importance

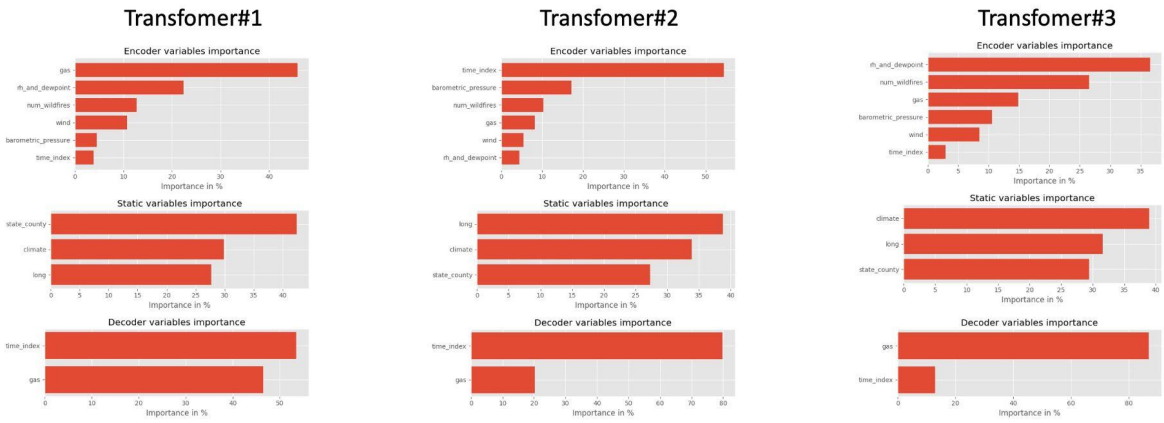

Figure 18: The Same Transformer Setting produced Different Variable Importances

While analyzing which model has the best performance, we also analyzed which features contributed the most towards predicting future AQI levels. However, we extract the feature importance from different components in transformers and find that the result is inconsistent for each time.

Discussion and Interpretation

1. Verifying the Hypothesis and answering Research Questions

We were able to verify that **the first part of our hypothesis is confirmed**, which was "With the integration of creative features such as gas consumption and climate types from external datasets and without relying on any previous AQI or air pollutant concentrations levels, future AQI can be forecasted with at most 15% symmetric mean absolute percentage error." This is because we were able to create an improved model with a SMAPE of less than 15% using only indirectly related features, some or many of which were from external data sets. However, we verified that the second part of our hypothesis is incorrect, which was "Gas consumption and the frequency/periodicity of wildfires will be the two most important features that contribute towards the accuracy of our model." As mentioned above, **the importance of features in our transformer model is inconsistent.** This can be ascribed to the fact that in a neural network, features will have more weight than others in certain scenarios and will have less weight than others in other scenarios. As a result, **the hypothesis that gas consumption and frequency of wildfires are the most important features in our model is rejected.**

A research question that we posed and answered during the experiment was whether the reduction in SMAPE from our baseline model to our improved model warrants the use of using more than twice the amount of features. Our answer to this question is that since the improved model only had a reduction in SMAPE of only around 3%, the use of 8 extra external features cannot be justified since it drastically increases model complexity, runtime, and space complexity for only a minor reduction in SMAPE. Our final model shows that the 2 features that our baseline model utilizes have relatively high accuracy on their own.

2. Outliers

Through our Principal Component Analysis in Part 1, we found that there were four counties in our training data that were outliers. These counties were Los Angeles, Riverside, Mono, and San Bernardino. The AQI level in these four counties had a substantial difference from the AQI level in the rest of the counties in the data set. Seeing how these counties are all located on the south side of California, one plausible explanation for this could be because Southern California is highly prone to wildfires. Higher frequencies of substantially more devastating wildfires could have disproportionately affected these counties and caused them to have much higher AQI levels than other counties.

3. Reasons Behind the Limited Improvements

Another important observation we made earlier was the very small reduction of 3% in SMAPE from the baseline model to the final model. A possible explanation for this could be the fact that the Temporal Fusion Transformer is powerful enough to make accurate predictions using only two features from our datasets. Therefore, the extra 8 features added in our final model did not make much of an impact even though they would seem to make more of a contribution to AQI levels.

Also, referring to Figure 6, we can tell that AQI has little periodicity. It may be another cause that AQI is hard to predict in the manner of time series.

Future Work

1. In further experiments, we would like to analyze the relationship between AQI levels and the number of coronavirus-related deaths and cases based on geographical location. Although we have done some analysis on further EDA we were unable to find a significant relationship between AQI and COVID. We still believe there is some underlying relationship between them. As many of our daily lives have been affected by COVID, working with COVID data would allow us to analyze a problem that is relevant and significant to our society today. Seeing how the number of COVID cases varies by region, identifying if a relationship exists between COVID and AQI can help figure out important trends such as if AQI can cause more people to be prone to COVID. Other types of scenarios that we would love to explore that are related to AQI are car accidents, the agricultural yield of crops, behaviors of animals like their migration patterns, airplane delays, etc.
2. Moreover, we would love to explore other forecasting models utilizing frameworks such as recurrent neural networks, that could potentially further improve the accuracy and quality of our model. Recurrent neural networks are designed to work with complex sequential predictions problems. Since AQI data is made up of the measurements of air quality over a period of time, and since we wanted to build a forecasting model that can predict the AQI levels for future dates, recurrent neural networks can help us achieve our goal.

References

Beitner, J. (2020, Sep 19). “Introducing PyTorch Forecasting.” *Towards Data Science*. https://towardsdatascience.com/introducing-pytorch-forecasting-64de99b9ef46

California Energy Commission. “Gas Consumption by County.” Retrieved Dec 12, 2020, from http://ecdms.energy.ca.gov/gasbycounty.asp

California Natural Resource Agency. “California wildfire perimeters.” Retrieved Dec 12, 2020, from https://data.cnra.ca.gov/dataset/2020s2

California State Geoportal. “California Fire Perimeters.” Retrieved Dec 12, 2020, from https://gis.data.ca.gov/datasets/e3802d2abf8741a187e73a9db49d68fe_0

Climates to Travel. “Climate - California (United States).” Retrieved Dec 12, 2020, from https://www.climatestotravel.com/climate/united-states/california

Environmental Protection Agency. “Pre-Generated Data Files.” Retrieved Dec 12, 2020, from https://aqs.epa.gov/aqsweb/airdata/download_files.html

Wu, N., Green, B., Ben, X., and O’Banion, S. (2020, Jan 23). “Deep Transformer Models for Time Series Forecasting: The Influenza Prevalence Case.” *International Conference on Machine Learning*. https://arxiv.org/pdf/2001.08317.pdf

Appendix

1. AQI Analysis Notebook

   This AQI jupyter notebook contains the implementation code regarding guided modeling and open-ended modeling. Some of them cannot be properly run on datahub, so if you want to reproduce the result, you can refer to this code.

   https://github.com/ByronHsu/ds-100-final/blob/master/AQI.ipynb

2. Dataframe Generator Notebook

   This dataframe generator jupyter notebook contains the implementation details of how we crawled the data from 2016 to 2020 and how we merged the original dataset with the external one.

   https://github.com/ByronHsu/ds-100-final/blob/master/ds_final_data_gen.ipynb