Cancer Rate Monitoring and Alert System Report

Problem Statement:

One major job of political and health officials is to monitor and respond to threats to public well-being. Cancer rates are driven by a number of environmental, genetic, and habitual factors, but abnormalities and changes can be a sign of problems within a community. For instance, St. Louis residents near the Bridgeton landfill started to suspect an unusually high incidence of cancer in their community. They fought for years with the EPA to address the issue of nuclear waste in the nearby landfill. A model that could raise an alarm early upon an increase of change in cancer rates would give faster warning and additional evidence that action is needed in these kinds of situations.

Goal:
The aim of the project is to flag communities with anomalous cancer rates. If a city or county has or develops an unexpectedly high rate of cancer, public health officials could begin investigations into environmental factors, working conditions, and the habits of community members to address issues early and at the source. A community with an unexpectedly low rate could also be flagged. On the one hand, this could be indicative of problems with detection and reporting. For instance, a community might be underserviced with regard to healthcare or have alarmingly low rates of insured individuals. This model would give those communities evidence that further assistance is needed. If, on the other hand, insurance and reporting are not in question, outstanding communities could serve as case studies for public health scholars.

Data Wrangling:

The dataset used for this analysis contained information from 3,047 counties across the United States. In addition to cancer rates, the data contained economic and educational information including, but not limited to, median income, percentage of the population covered by public and private insurance, and percentage of the population that completed high school and college. It also contained demographic information such as population and percentages by race.

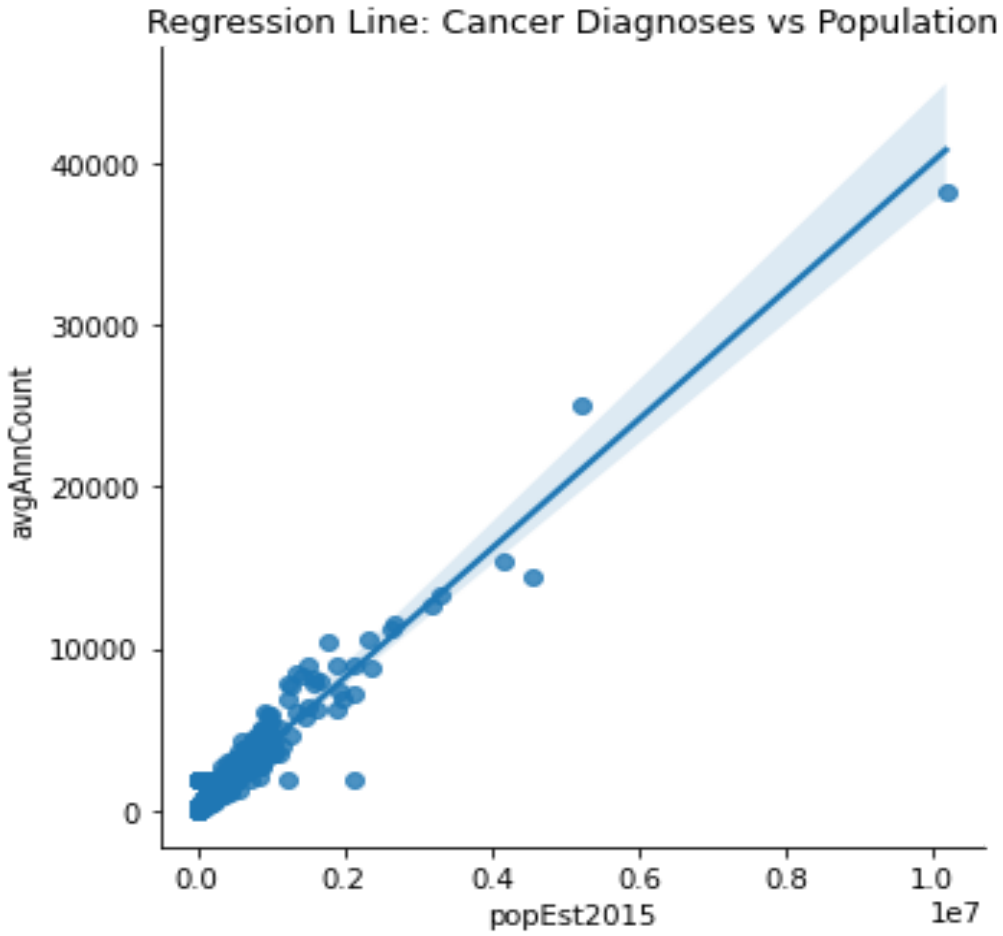
There were a significant number of missing values in the column representing the percentage of citizens 18-24 who had attended some college (but had not graduated). Since there were other similar educational indicators such as percentage who had finished high school and college, I dropped this column from the analysis. Similarly, I dropped the column representing the percentage of people with private coverage alone, as the data provides a number of other relevant health coverage metrics including the percent of the population with private coverage and employer private coverage.

Finally, there were missing values indicating the percentage of the population 16 and older who were employed. These were relatively few in number and the subset of counties missing this value were similar to the larger dataset other regards. Since they were relatively typical counties

otherwise, including other employment statistics, I filled these missing values with mean of the column.

EDA:

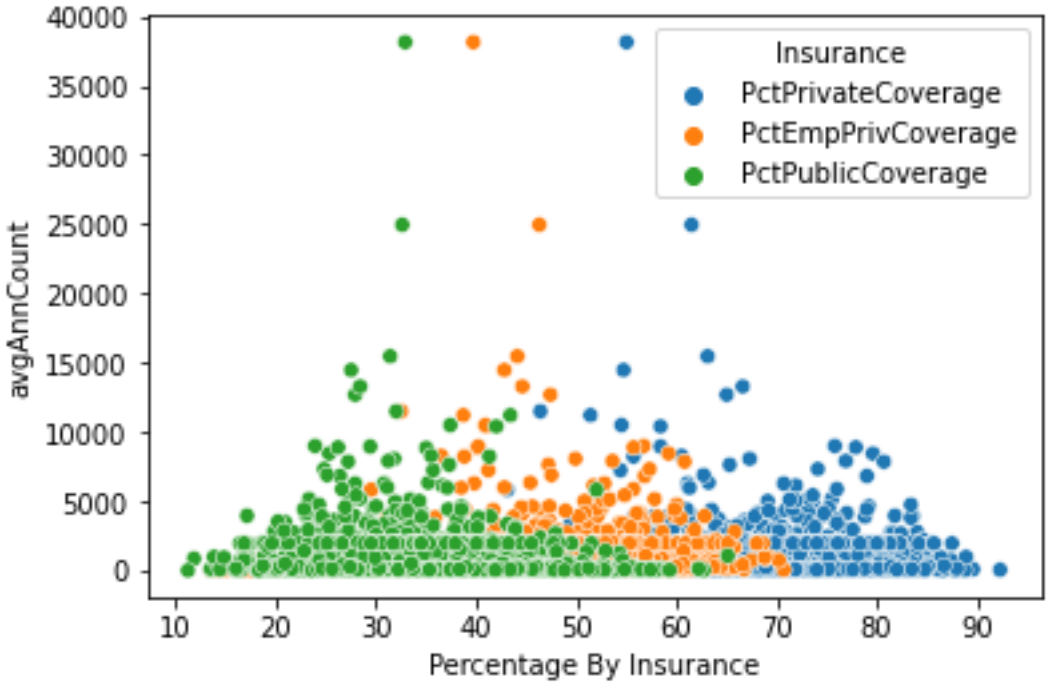
I began exploring the data by trying to get a sense of how the different variables correlated with number of cancer diagnoses in a county. Unsurprisingly, there was a strong linear relationship between the number of diagnoses and the population within a county. The goal of this model is to add more nuance to that figure by taking into account factors such as economic and educational variation among counties. Below is the regression line with a 99% confidence interval on the graph of cancer diagnoses vs. population.



Further exploration revealed some other points that were outside the scope of this particular model but may be of interest to future studies. For example, below is a graph of cancer rates with the percentage of different insurance types designated by different colors. Where the percentage of public coverage is greater than 45% or less than 20%, nearly all the counties have less than 5,000 diagnoses. It could be the case that these counties are almost exclusively small, rural areas with low population counts that account for these small numbers. However, someone

investigating health outcomes of publicly insured individuals might want to take a deeper dive into those numbers.

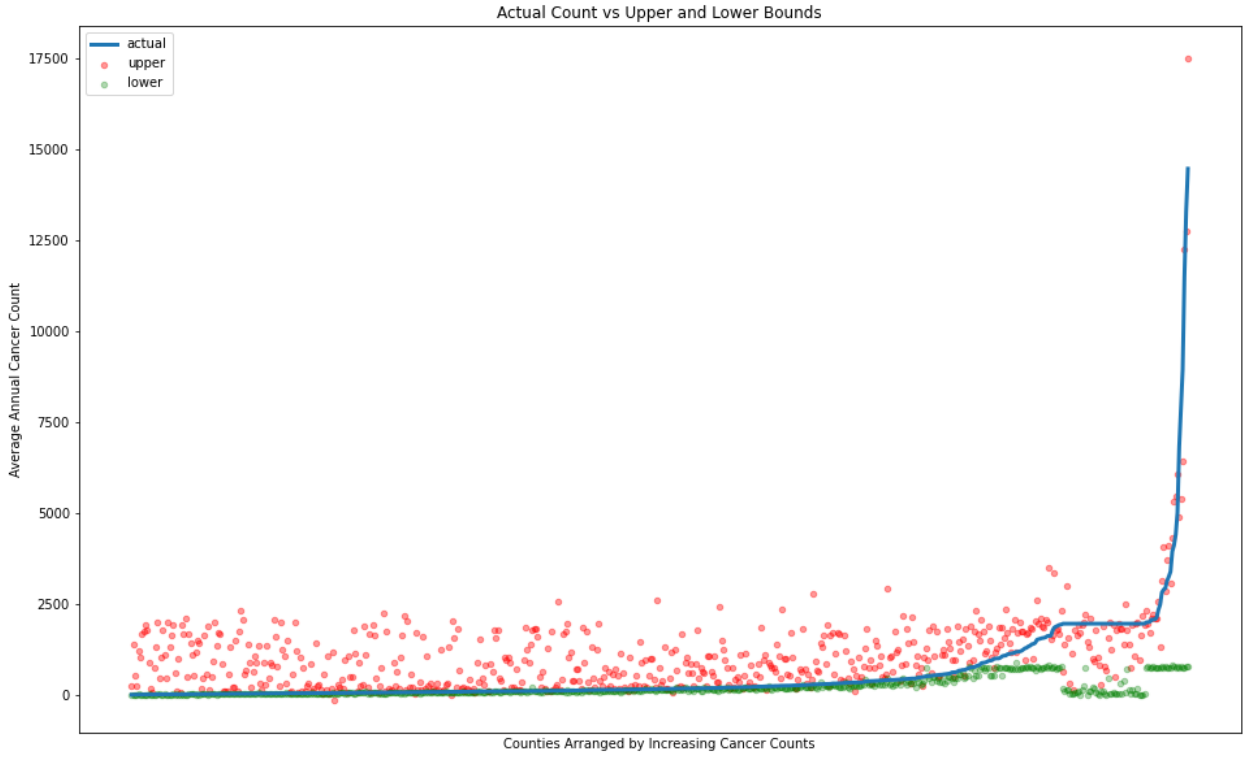
Percent Insurance Types in Relation of Cancer Diagnoses



Modeling:

My initial goal was to not only come up with an estimate for cancer rates in a county, but to be able to flag cases that were much higher or lower than expected. To attempt this, I employed scikit learn's gradient boosting regressor to create three models. The first model was a "mid" model that served as the direct estimator of cancer averages. This model was intended to give a baseline estimate for cancer diagnoses in a given county. The next two models were "upper" and "lower" models. The hope for these models was that they could identify the upper and lower 90% of cancer estimates. In other words, if the model could reliably tell us that the cancer average of a county had pushed into the top 10% given county demographics, that could trigger the alert to public health officials.

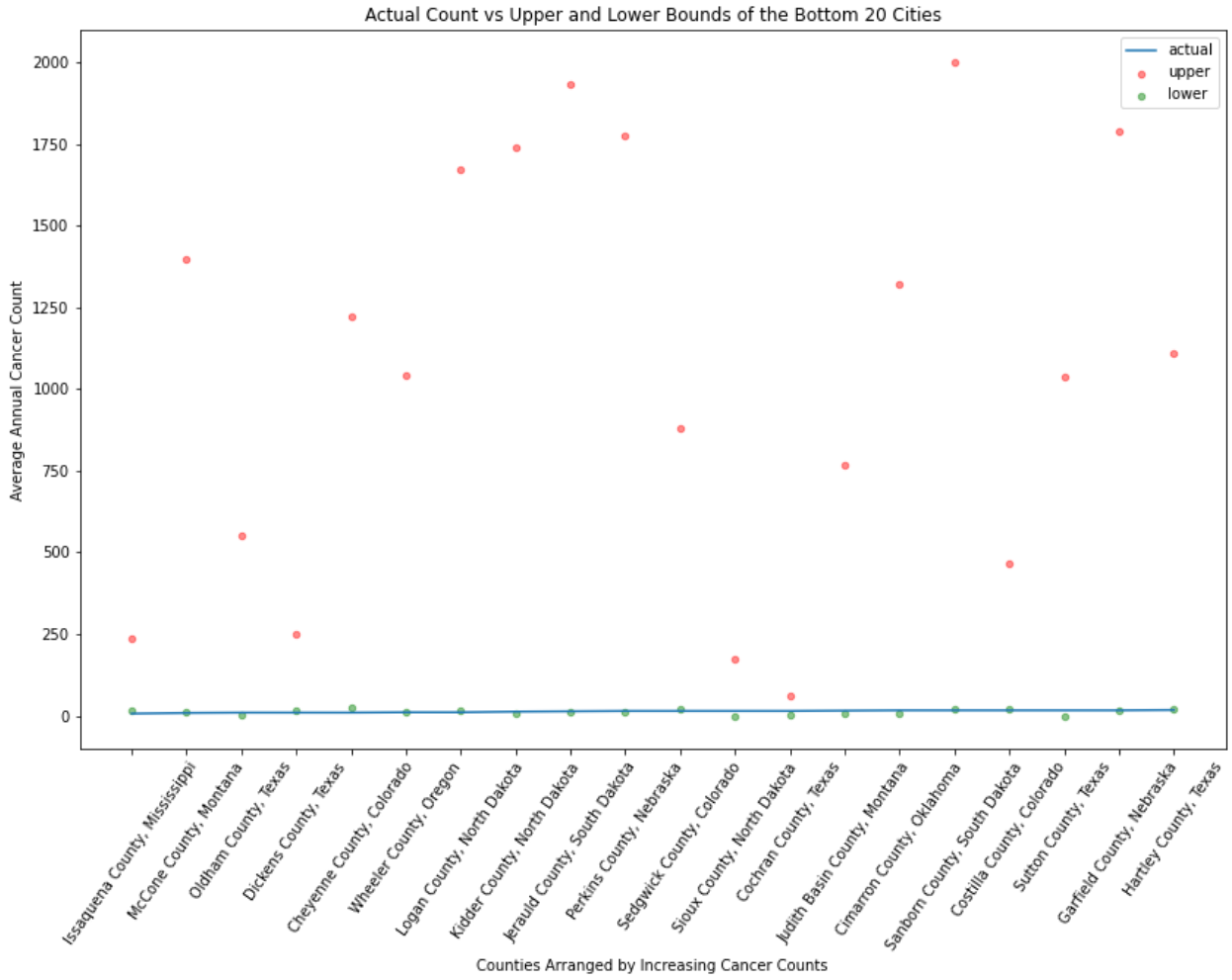
Unfortunately, even after tuning hyper parameters to optimize these models, they still struggled to make accurate predictions, with the upper and lower bounds showing minimal change as the actual cancer diagnoses rose. The first graph below shows all counties' cancer rates in comparison to the upper and lower estimates.



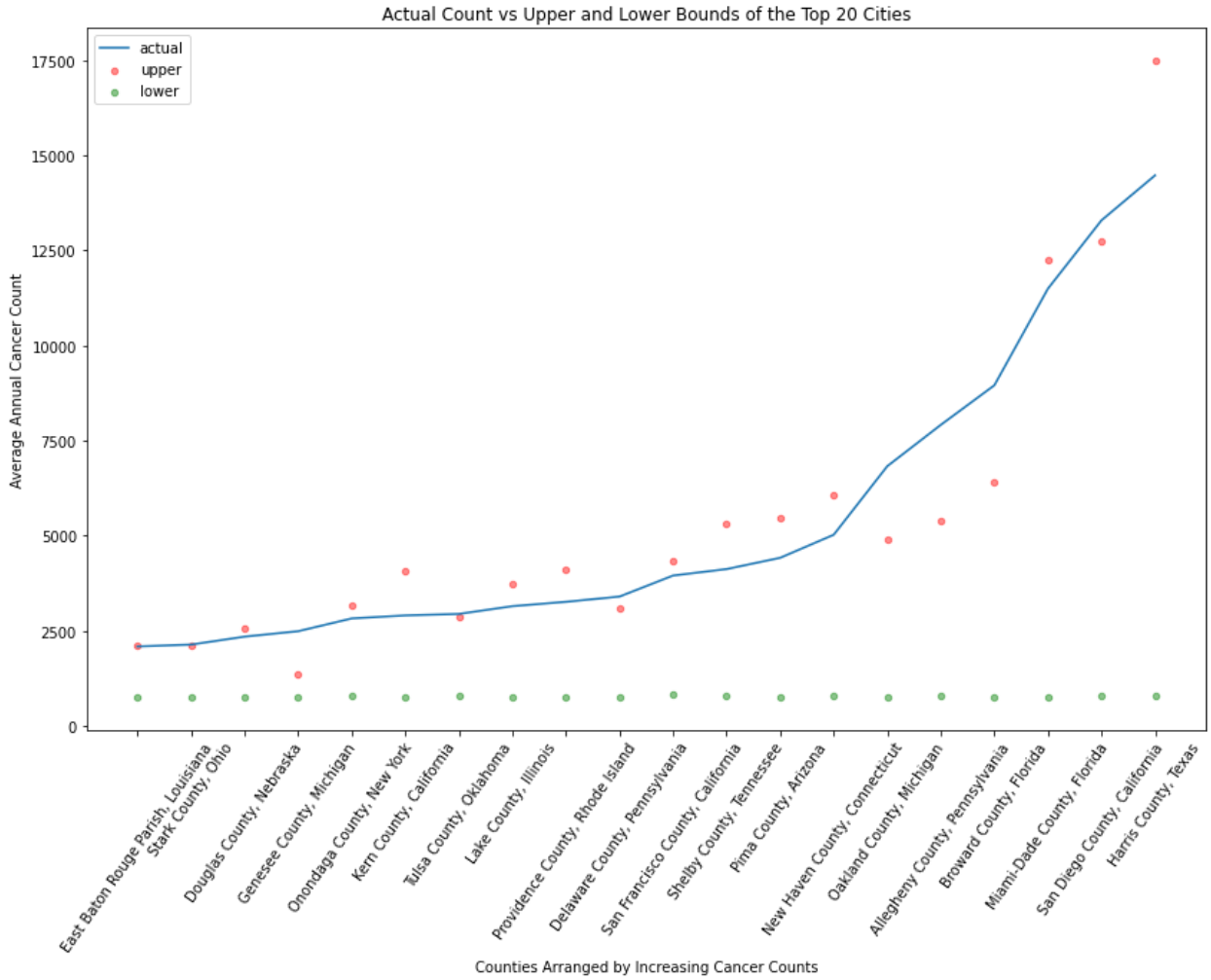
Moving into the second half of the graph the models appear to start converging closer to the actual numbers. This hope vanishes, however, as numbers start to make a more significant increase. Neither the upper nor lower bound track the rise in counties with more diagnoses. While the models do successfully capture 71% of the actual values between the bounds, this is far from the target 90%, particularly given how wide the range is for many of the estimates. For example, many of the counties with actual numbers in the double digits, have upper bounds in the thousands. This would clearly make for an ineffective warning system if cancer diagnoses had to rise 50-100 fold before an alert was raised.

The two graphs below show the bottom and top 20 counties respectively, in relation to upper and lower estimates.

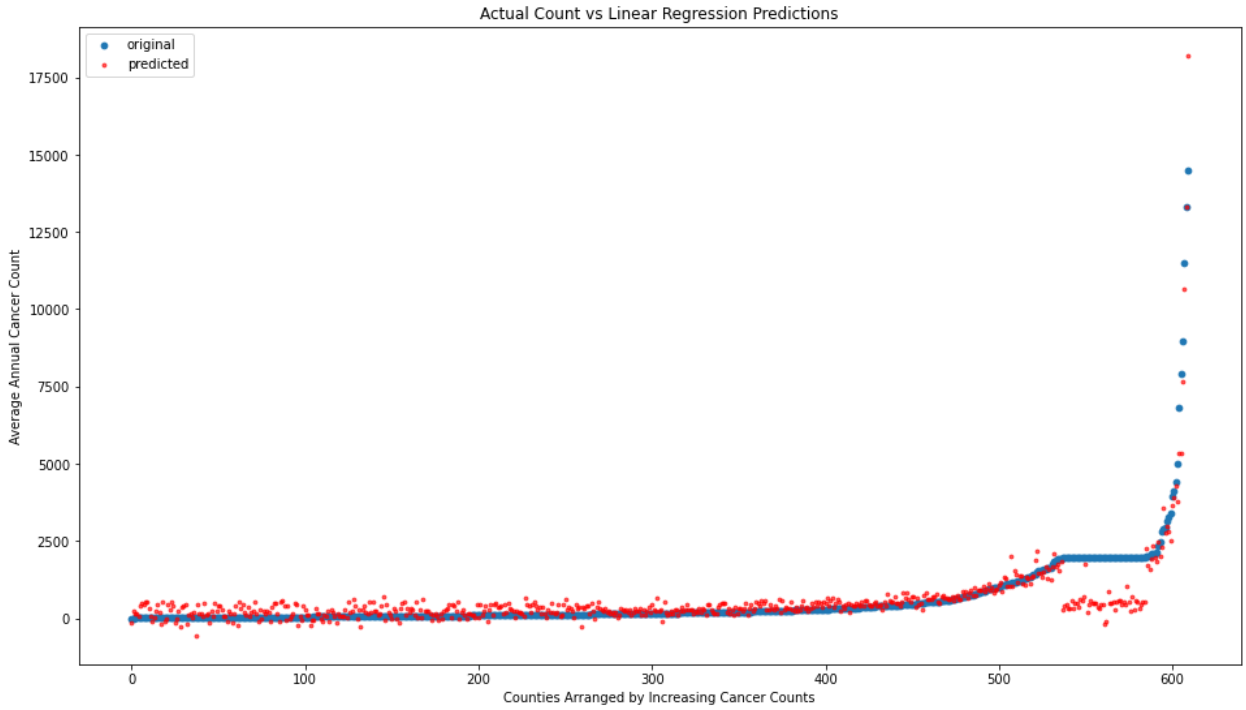
Given that cancer counts in these counties range from seven to eighteen, there it is not surprising to see the lower model tow the line of the actual count. However, only one of the upper model predictions appears within 100 of the actual count. To make matters worse, 12 of the predictions are above 1,000 and half of those are above 1,500. Using these metrics, a county with no change in demographics could experience a 100-fold increase in cancer rates and still not hit the upper bound prediction. An alternative method would clearly be needed to generate a warning.



At the other end with our 20 counties in the test data with the highest cancer average, our bounds look strange as well. While we would expect, on average, two of our upper bound estimates to fall below the actual values, seven do. Furthermore, our lower bound model exhibits no noticeable change as the numbers increase. Given that the goal of the model is to identify unusually high or low numbers, the range of 17,000 given by the upper and lower bound on our final data point is unacceptably wide. This level of uncertainty would not provide meaningful support to a public health team.



I had marginally greater success with scikit learn's LinearRegression model. I first explored how many features to include. According to a kbest analysis, 23 features was optimal. This was a slight reduction in dimensionality, down from 26. Below is a graph of the actual diagnoses in blue and model predictions in red. This model had a mean squared error of 268,127. In other words, the model was making estimates within about 517 diagnoses of the actual values.



While these numbers are not particularly impressive, we can see in the graph above that a majority of our error is concentrated in a couple spots. Much like with our upper and lower bounds, our model had a tough time with the lowest and highest values. This may give a clue as to what counties this model would be suitable for. It may then be necessary to come up with a different approach for other counties. Secondly, our predictions start to rise with the actual

number of diagnoses, but then around the 2,000 mark almost uniformly drop near zero. While it is unclear what is causing this error, it might provide a clue as to further data we need to include to correct our model and give it greater utility.

Conclusion:

My initial hope was a set of three models that could serve as triggers to notify public health officials in cases of unusually high or low cancer levels in a county. Given the error metrics of the upper and lower gradient boosting model, particularly at the low and high ends of our testing data, a different approach seems necessary.

While more work would need to be done, the LinearRegression model did have some initial – albeit limited - success. The model tracked well along approximately 70% of the data. Furthermore, there were clear sections - the lower most, uppermost, and data points close to 2,000 – that generated a majority of the error. Moving forward, this at least offers some areas of investigation and might provide hints if we were to seek more data to tune the model.

If the LinearRegression model was improved, one other approach might be to abandon the three-model approach and instead use the error metric to develop an alert system. For instance, if an average error of 100 diagnoses is expected, one approach might be to scale that error up and down with population and then have a system that sends an alert if the actual number of diagnoses exceeds twice the expected, scaled error.

Further Usage and Investigation:

It might also be worth considering whether these models could be used to elucidate issues in environmental justice. The data provides information on each county's racial makeup broken down by percentage of the population. One exercise could be segment counties that are predominantly communities of color and compare a model trained on that data to a model trained on predominantly white communities. Alternatively, if we gathered data about actual diagnoses by race, we could train a model with no information regarding race and compare its predictions with actual values. If disparities were found in health outcomes based on race, it would not be the first finding of its kind. However, it might add further detail and supporting evidence for activists and public health officials seeking to address inequalities.