Jack Dwyer

Springboard Data Science Intensive

**Predicting Energy and Silencing Counterfactuals**

**DATA DICTIONARY:**

**Dataframe 1:**

**3 columns of interest:**

* Year (1960-2014)
* State (Alabama – Wyoming)
* Generation Quantity (In MMBTU)

**Dataframe 2:**

**5 columns of interest:**

* Year (1990-2015)
* State (Alabama – Wyoming)
* Type of Consumer
  + Commercial Power
  + Electric Power
  + Industrial Power
  + Electric Utilities
  + Independents
* Energy Source/Input Type (most prominent listed below)
  + Coal
  + Natural Gas
  + Wind
  + Hydroelectric Conventional
  + Nuclear
  + Solar Thermal and Photovoltaic
* Generation (Megawatt Hours)

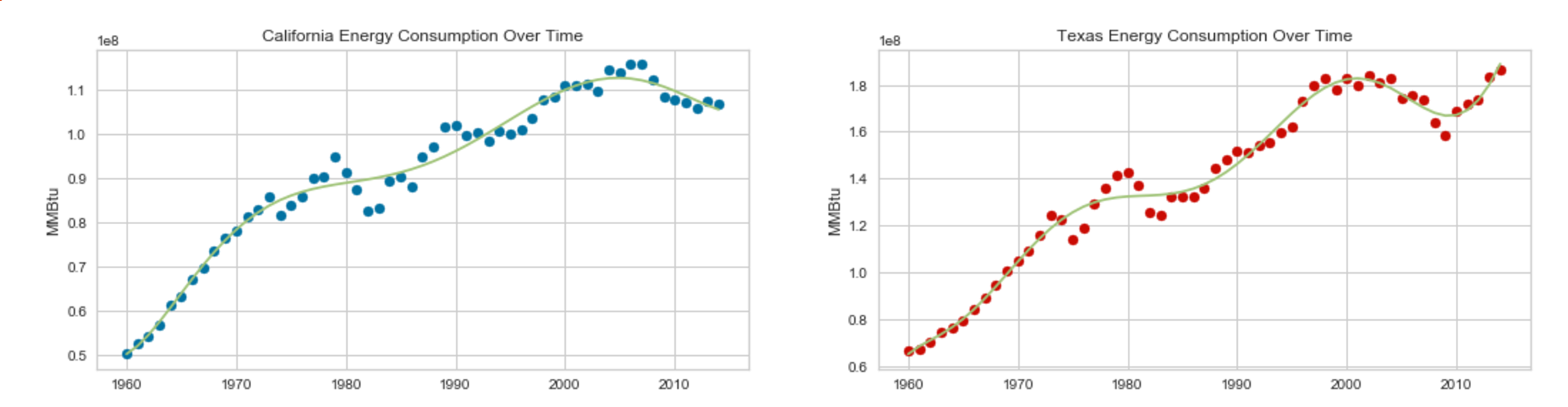
Energy is the hot topic of 2017, especially with the new conservative administration. There are a lot of counterfactuals being thrown around, and having the energy education I do, and the ability to make use of data, I decided to show just what is happening with energy in this country, and to show that being conservative or liberal in identity will not actually have too profound an effect on how a state goes about generating its electricity.

Starting out, I had two data sets I wanted to work with. The first would provide a more macroscopic view of what energy is doing in this country. By state, it outlines generation by state, grouped by year. Using an unstack, I turned this into a year by state dataframe, making it easy to perform analysis on.

The other dataframe was initially a bit overwhelming. It has every state, grouped by year, with their energy inputs by quantity, and the type of consumer using these inputs. I spent more time than I probably should have performing various “groupbys” and “sorts” on this to make it into something I felt I could reasonably analyze.

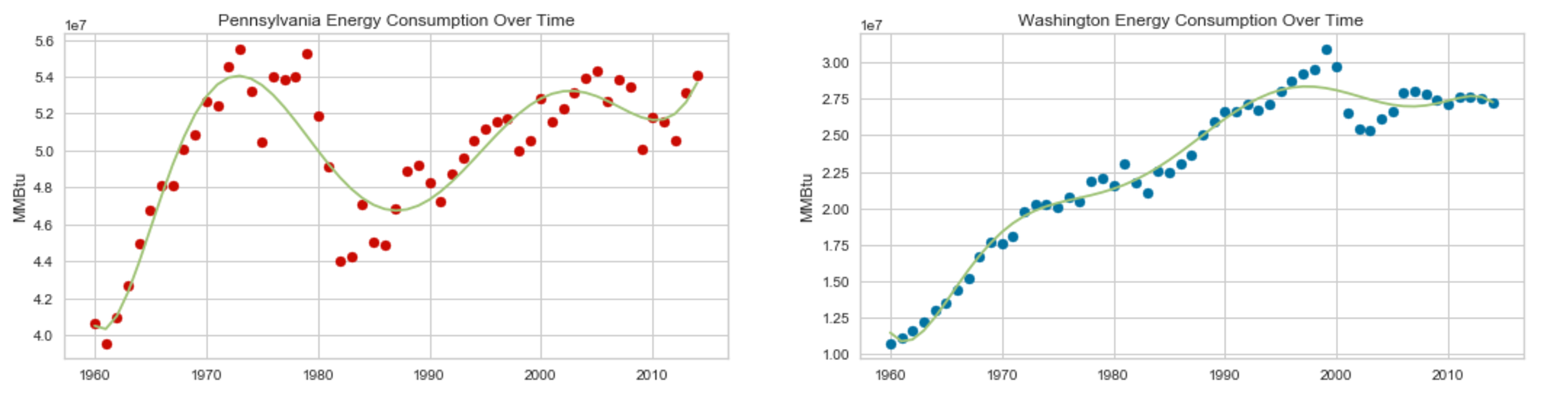
Each frame was nearly completely populated. Initially there were a few zero values. I contemplated forward filling or interpolating them, but they had so little an impact I ultimately dropped the rows because there was enough accompanying data to still provide what I needed for analysis.

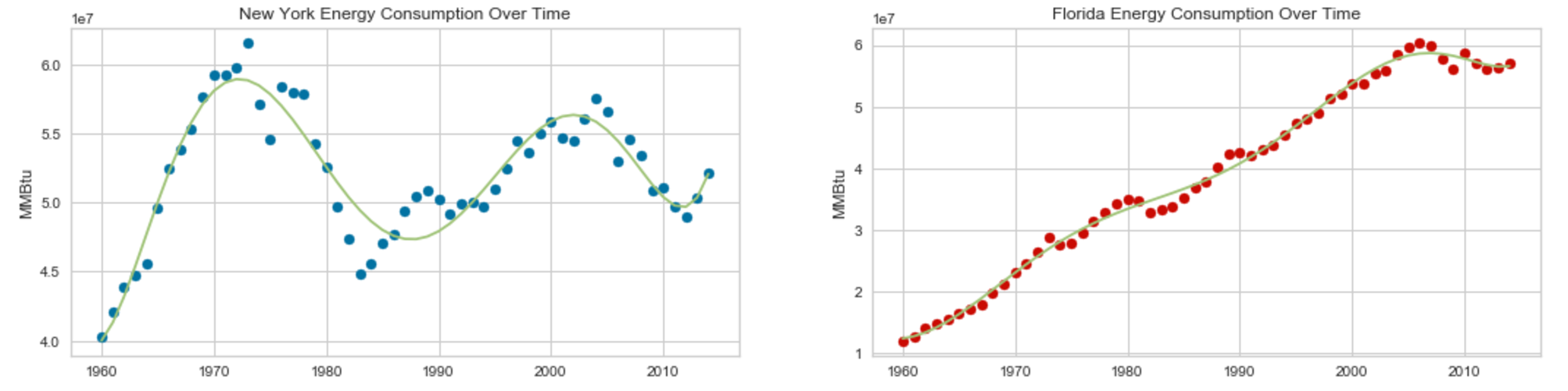
For the first data frame, once set up in the year by state manner using the unstack, I was able to sort by generation to find which states were consuming the most energy in various time periods (I did 1960, the first year on record, and 2015, the last year on record). From this, I chose states I thought would really represent not only the lion’s share of generation, but also the tumultuous political climate. Texas and California were obvious choices to plot together, and they initially painted a very predictable picture. California had been needing to generate less electricity, while Texas more; could it be that conservatives really were more wasteful?



Drawing on my background, and knowing how to think critically, I suspected there was a little more going on than just political climate, so I performed polynomial linear regressions on the top states, as well as the US. I chose polynomial linear regression because using a linear model cast too wide a stroke over the data; everything would have been a positively sloped line, which does not reflect what some of the nuances in the data show. The results were probably surprising to most, but really aligned with what I thought; it is not only about political party. I used various degrees to regress on, and with lower order polynomials, Texas and California looked nearly identical. Of course, once I raised the order of the function, some of the nuances were captured, and Texas has a negative (positive sloped) energy outlook, while California a negatively sloped, positive outlook.

Not satisfied there, I looked at a few other states. New York, a liberal state is very much oscillating around a center. A pure linear regression of the data yielded a nearly flat line, and a higher order regression actually has them projected to grow; neither of these align with what California projected. Washington is the same story, but it actually seems to be slightly growing over time with it only recently plateauing. Regardless, maybe generation is not all that political after all. The conservative states were equally as inconsistent, with Pennsylvania oscillation around a mean and Florida climbing over time, but ultimately plateauing as well.





Okay, so maybe politics aren’t such a great indicator of how much a state is going to generate, but could they be a factor in deciding how they generate? This is where dataframe 2 was heavily utilized.

The first thing I did was a quick analysis to determine if who is consuming energy has changed enough over the decades that a forest could find it. I broke up a subset of dataframe that contained information about what industries are consuming the generated energy. A quick graph showed us that the mix had been changing over the years. Using a Random Forest, I ran a grid search cross validation and found my n\_estimators should be three and I should not set a max\_depth. I chose the forest because I know it performs bootstrapping and I knew this would benefit the outcome (this is my reasoning for every implementation of it). This forest predicted our decade based on the breakdown of who was using power in a given year, and yielded a ~88% accuracy.

What really interested me though was whether or not states are using coal. The reason I am picking on coal is because it is the only dirty fuel still pervasive in today’s energy mix. Natural gas is a fossil fuel, but its emissions are so small (relative to coal), that even most progressives are accepting of it for the near-mid future.

The first step in this data frame’s main use was to group by state and input type. From here, I sorted and found the max generation associated with a given input type for each state. This was a big assumption on my part, both because this was a sum of all years, meaning states that had recently switched to cleaner energy may still be categorized in the coal category, and it didn’t include the second biggest generation. Regardless, I moved forward because I needed some form of analysis, and those would certainly do. The dataframe I performed analysis on was a surprisingly good representation of what is going on, even with all the assumptions.

I ran a simple K-nearest neighbors model to see if maybe states in need of more generation used a certain generation type more often. K-nearest neighbors would give me an estimate of how many neighborhoods I had, which was explanatory in and of itself. I ran a grid search cross validation and found my n\_neighbors to be 3-5. I went with 5 to provide enough room for the model to capture the various generation types. I considered bootstrapping to get more data, but I didn’t want to paint a skewed picture because the United States is pretty static. My accuracy ranged from .4 to .7, which, considering how little data the model got to train on, I’m happy with.

I then merged a political affiliation dataframe onto dataframe 2. I got this data from the majority in the House of Representatives for each state, which I figured would be better because the “sample size” of representative is generally bigger than those of senators or electoral college swing. I ran KNN again, this time with political party included, and found really similar results.

Wanting to be thorough I ran a hypothesis test on whether or not the two test yielded different results. I took a difference of means between the two, permuted them 1000 times, and took the difference of means again on each of those samples. I counted the samples that were bigger than the original difference of means, and ultimately failed to reject (including the political affiliation didn’t help in predicting generation).

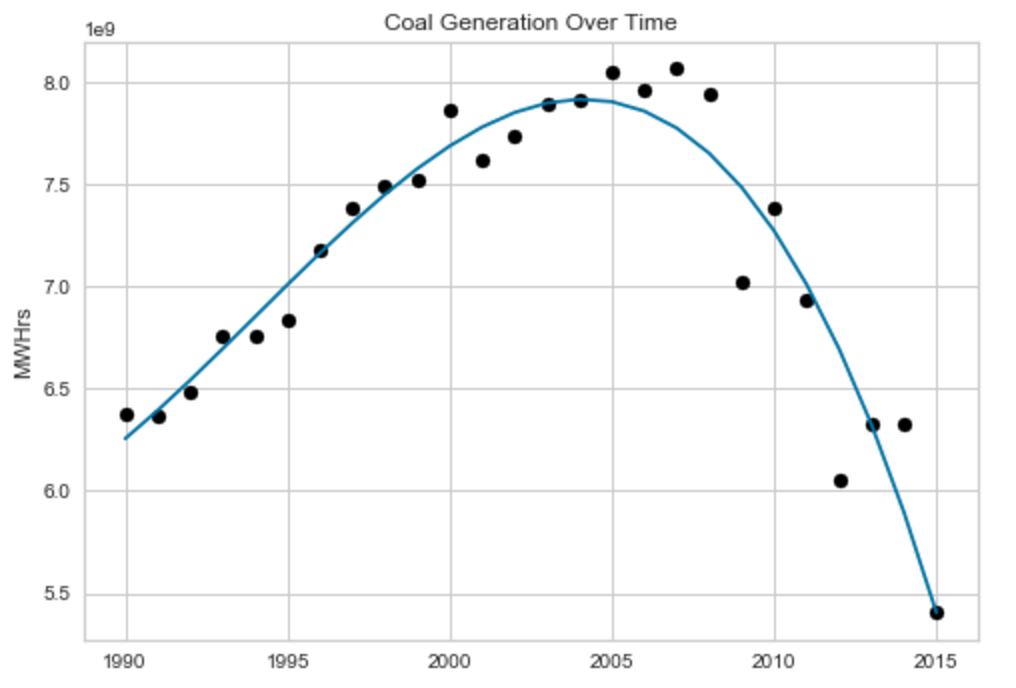
I did another test to determine whether a Random Forest could do a better job than our K-nearest neighbors on predicting based on quantity of generation alone. I found that there is a significant difference between the two, with the Random Forest doing better, at the 85% level.

The last piece of analysis I did was sort of a counter to this, can generation and political climate at least predict whether a state will or will not use rely on coal. I used a Logistic Regression this time to classify, and ran a grid search cross validation. Because of how the splitting works, my C values were across the board from 3-4, but my penalty was always L1. Ultimately our accuracy ended up being ~.60, which again, isn’t all that bad considering the assumption made and the sparsity of our data.

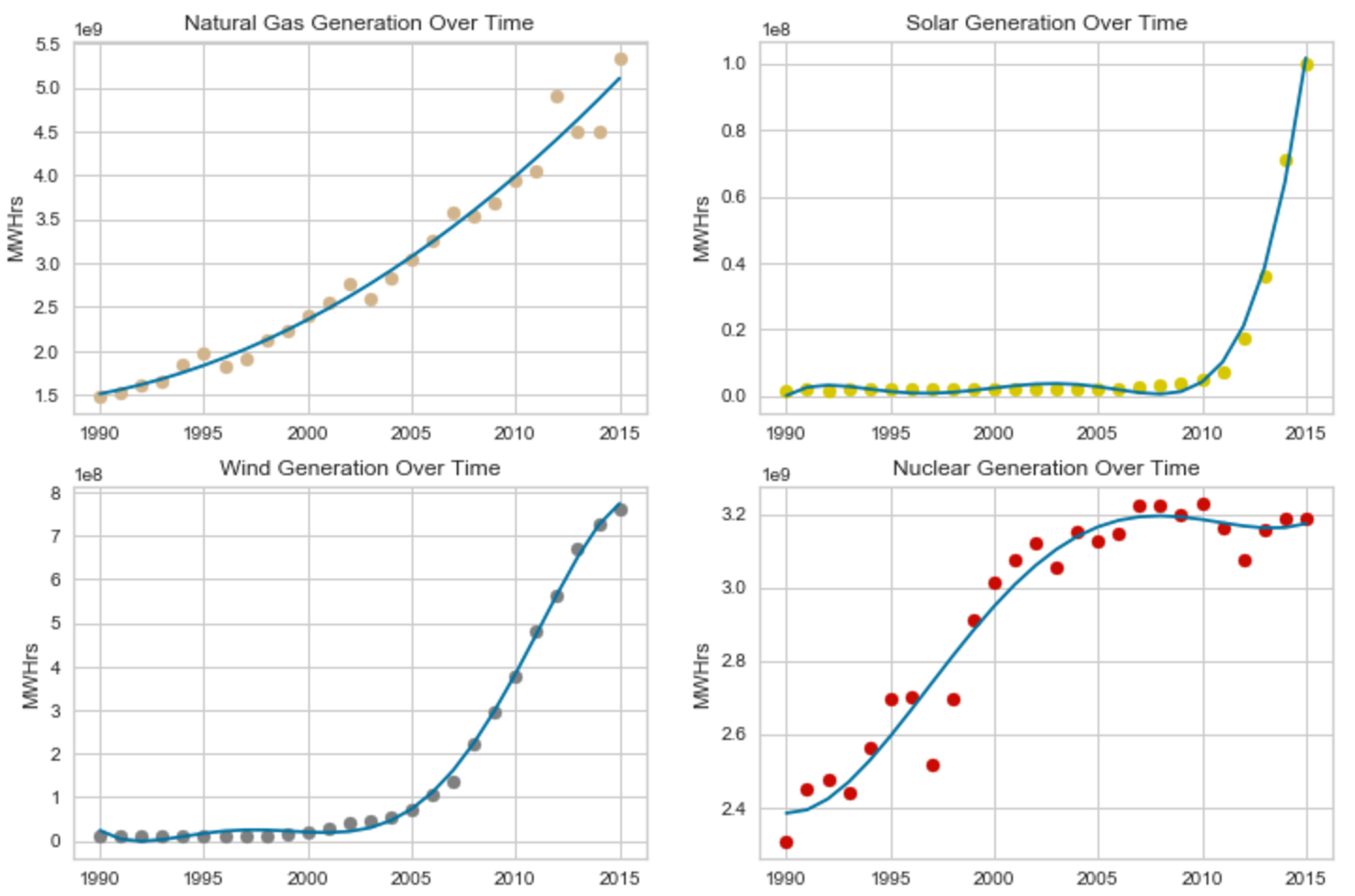
I ran this again, but with just political inclination as a predictor, and I actually found that it did a pretty decent job of predicting. With an L1 penalty, and a C of ~.5, I got an accuracy of .71. I ran it a few times and found similar results, so it appears that political affiliation may indicate how willing a state is when it comes to the idea of using coal, but not the other way around (Liberal states are likely opposed to coal use, while conservative states are indifferent).

Wanting to test the efficacy of various algorithms, I ran another binary classifier. I chose a Support Vector Machine (C = 1, gamma = 1), and tested if it was better than our Logistic Regression. It outperformed at an 85% confidence levels, which I think is conclusive enough to say it did better.

Although the analysis has yielded some surprising results, potentially the most surprising and comforting news (at least to progressives), is that coal is not only declining at a positive rate, but that it has been for decades. This is not political in nature, as George W. Bush was in office for a number of those years, but economic. With the proliferation of Natural Gas through fracking and shale exploitation, coal is being beat out by the cheaper, cleaner, more effective natural gas hitting the market.



We also see that virtually every other form of generation is taking its place in the mix. Very good news for other producers, but bad news for coal workers.



One thing to note: Nuclear was climbing rapidly until Fukushima happened. It lost a lot of traction with this events, and even though it is actually the safest form of large scale power generation, panic tends to spread, and scapegoats are made. Fukushima was considered to be the perfect storm of bad luck, and terrible design, but that did not stop the media from massacring it. We are only now starting to see a recovery of Nuclear, but time will tell how well it fares with the rapid improvements in other, less radioactive clean energies.

My take away messages are as follows:

* Coal’s days are numbered, so if you’re in that space, it may be time to start exploring other opportunities.
* Some of the other opportunities are outlined in my second takeaway, which is that renewable/natural gas is exploding right now. There is a positive slope associated with all of these forms of generation, and a positive slope on that positive slope, so there is certainly room to grow.
* The last takeaway would have to be that while liberal states are going to dislike coal, conservative states may as well. There is some predictive power in figuring out if a state likes coal based off of its political affiliation, but the actual accuracy scores are too low to be trusted.

Having more data would have made this analysis much more viable, but even with the limited amount we had, there were still things learned. It’s a tough problem because there is a limit to how much data we can collect; there are only 50 states. Breaking things down by county or district could have been beneficial (had we had access to that data). Of course we can’t rely on these models, but they help us not only understand the landscape better, but they give us an idea of what the next steps should be in figuring out what energy is doing in the United States. Nothing is as clear cut as the media makes it out to be, mainly because no one wants to do any hard thinking, they just want the headline. Our job as data scientists is to give them a headline that depicts the truth while we work to make sure those headlines are reliable and accurate.