Using Machine Learning to Examine the Link Between Energy Adoption and Political Affiliation

Hajime Alabanza

Contents

1 Introduction **2**

2 Descriptive Statistics **2**

3 Choice of Methodology **4**

3.1 Logistic Regression 4

3.2 Random Forest 6

4 Conclusion **7**

References **8**

1 Introduction

In recent years, climate change has polarized American politics. While there are numerous reports claiming that red states lead the renewables charge, others assert that republican lead states are too reliant on fossil fuels (Moore, 2015), (Balaraman, 2017). Through multiple classification methods, I will ascertain whether a collection of 10 energy sources can determine if a state is red (voters predominantly in support of Republican Party) or blue (voters predominantly in support of Democratic Party). A successful model will reveal whether energy portfolios truly differ between red and blue states. Ultimately, the objective of this study is not to point fingers at any party. Instead, I hope that meaningful results lead to renewable energy proliferation.

2 Descriptive Statistics

In this study, 2016, state-wide data (50 states + District of Columbia) is assembled from the U.S Energy Information Agency (EIA) (Administration, n.d.-b). 10 energy sources will first be considered as covariates in this study: coal, natural gas, hydroelectric conventional, nuclear, petroleum, solar thermal and pv, wind, wood and wood derived fuels, other biomass (waste, biofuels, etc.), and other energy sources (purchased steam, tire-derived fuels, etc.)(Administration, n.d.-a). These resources provided nearly 100% of electricity generation in the United States and their utilization is illustrated in Figure 1:

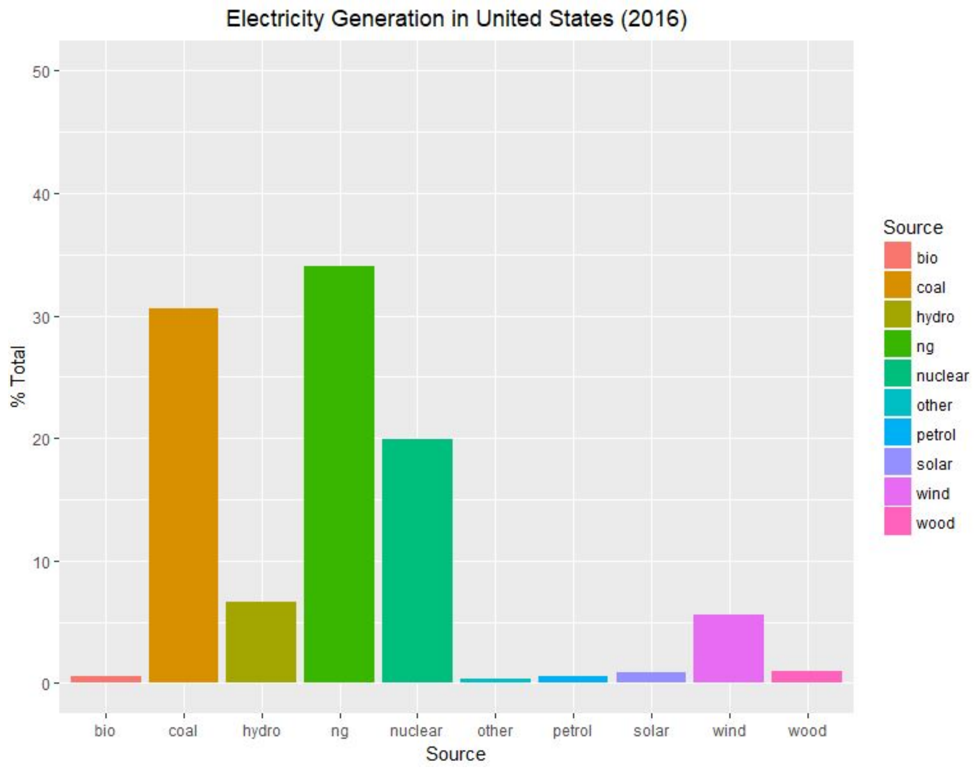

Figure 1: Two sources, geothermal and pumped storage, were removed from this study due to lack of data. Various sources show that they make up less than 1% of the energy mix (Administration, n.d.-a)

Electricity generation was dominated by three sources: natural gas, coal, and nuclear energy (70% of the total mix). In terms of renewable sources, hydroelectricity and wind led the way. Despite this, there are many states that are taking measures to expand renewables. The clean energy landscape is illustrated by Figure 2:

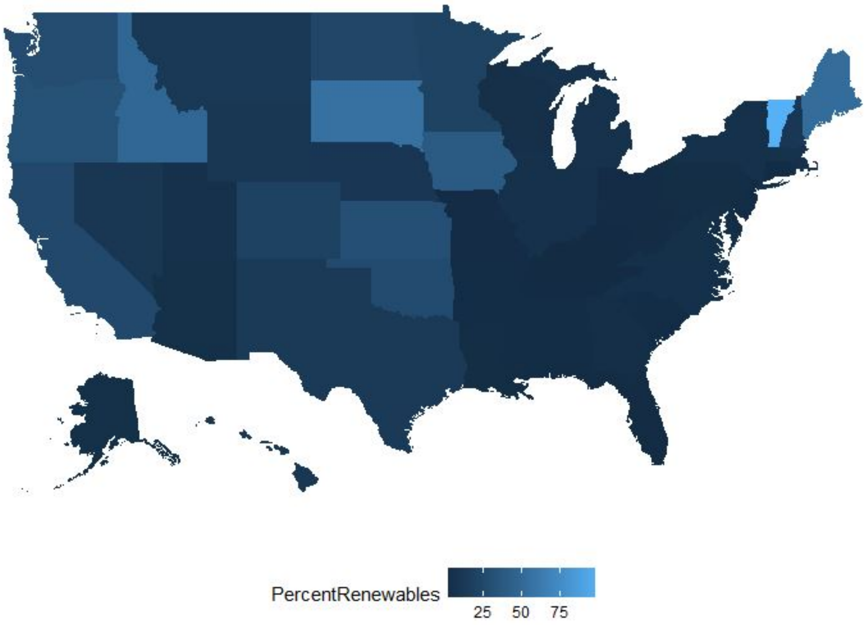

Figure 2: The western portion of the United States seems to have adopted more renewables (excluding hydroelectricity) into their mix, while the south/southeastern US relies on conventional sources.

Next, I dig deeper and examine whether energy profiles differ by political affiliation. Figure 3 breaks down electricity generation per customer by red and blue states:

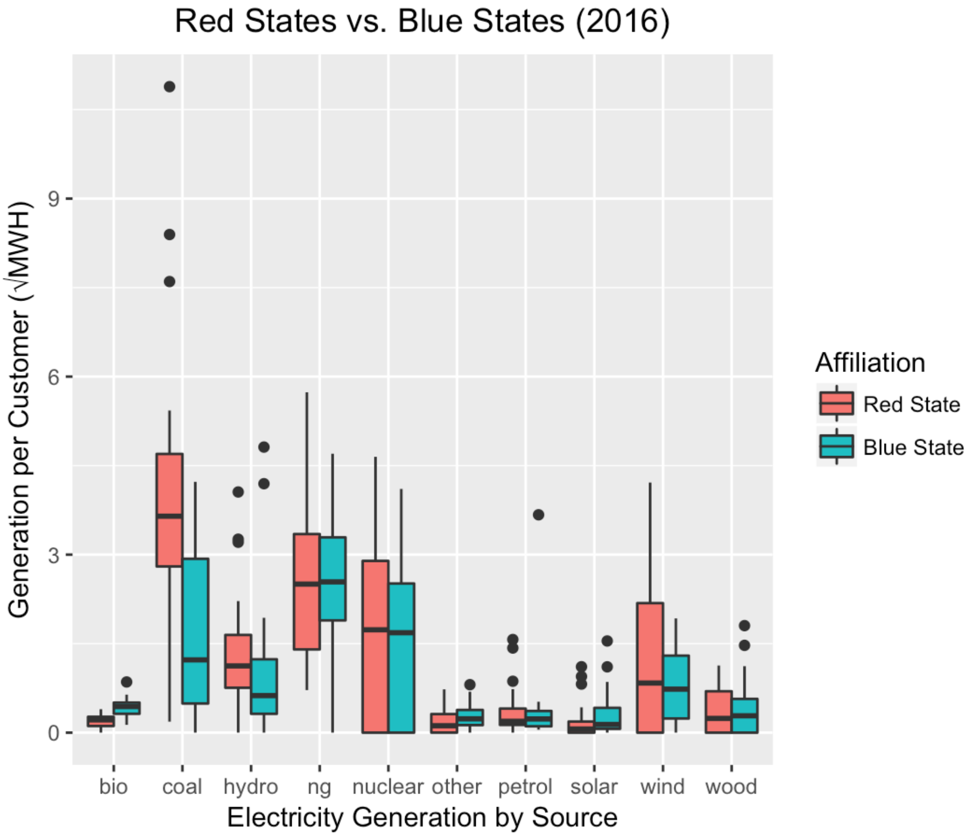

Figure 3: Red and Blue states determined by the results of the 2004, 2008, and 2012 U.S presidential elections(MIT, 2016). For example, Nevada was classified as a blue state as it was carried by Democrats in two out of three elections, in that time frame. In total, there are 24 red states and 27 blue states.

Above, we see that the median generation per customer (note, that I took the square root of MWH to further scale the covariates) for sources such as biomass (bio), coal, and hydroelectricity (hydro) differ a noticeable amount, while other sources look to be at similar levels for both red and blue states (albeit variance for red states seems to be much greater than blue states). Ultimately, this is an indication that these three sources could differentiate red states from blue states.

To get a closer look at whether red differs from blue, I conduct PCA. Through this, we hope to see whether 1) red or blue states form separate clusters 2) states are associated with a particular energy source(s):

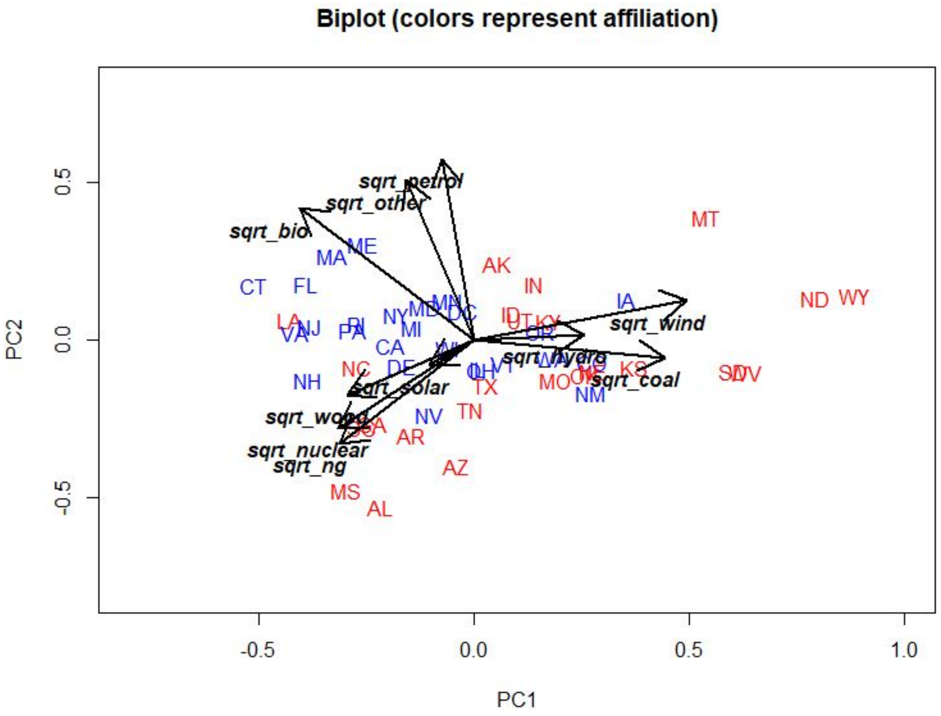

Figure 4: PC1 and PC2 explain nearly 45% of the variation in the data

Although there is a good amount of overlap, Figure 4 shows general separability between affiliations. In particular, many blue states hover around the covariate bio while red states seem to be concentrated around natural gas and coal. In summary, data exploration suggests that red states and blue states have different energy profiles. However, this cannot be taken at face value—it merely forms a hypothesis. In the next section, I build classification models to arrive at formal conclusions.

3 Choice of Methodology

In the subsequent sections, I implement two classification methods to determine whether a state is red or blue based on its energy profile: 1) Logistic Regression 2) Random Forest. Again, my goal is to classify whether a state is red or blue given its energy mix. All methods have their advantages. For example, random forest models work well in high dimensional spaces, like this study, and can help to pinpoint the most important energy sources associated with each political affiliation. A logistic regression provides parameter estimates, which allows for precise interpretation. Most importantly, I employ both methodologies in order to see if I get consistent results. That is, if both methods produce similar results, there is strong evidence that the underlying model is valid.

3.1 Logistic Regression

In this section, I start off with the following logistic regression model to answer my research question:

$$logit(\pi) = \beta_0 + \beta_1 solar + \beta_2 coal + \beta_3 ng + \beta_4 wind + \beta_5 bio + \beta_6 hydro + \beta_7 wood + \beta_8 petrol + \beta_9 nuclear + \beta_{10} other \tag{1}$$

where π =Pr(y=1$|solar, coal, ng, wind, bio, hydro, wood, petrol, nuclear, other$) with blue state $(y = 1)$ and red state$(y = 0)$

As a reminder, covariates are expressed here as generation per customer in $\sqrt{MWH}$ (and all future models). Next, to maximize the precision of the parameter estimates, I implement model reduction techniques (via stepwise model selection and likelihood ratio test). In doing so, I arrive at a more parsimonious model:

$$logit(\pi) = \beta_0 + \beta_1 coal + \beta_2 bio \tag{2}$$

Finally, using Cook's Distance, I identify influential observations to see which states deviate most from the model. Results are displayed below:

Table 1: Results

	Dependent variable:
	Affiliation
coal	-0.525^{*}
	(0.297)
bio	10.531^{***}
	(3.629)
Constant	-1.689
	(1.504)
Observations	51
Log Likelihood	-19.494
Akaike Inf. Crit.	44.989
Note:	*p<0.1; **p<0.05; ***p<0.01

We can see that just sources coal and bio are statistically significant (although coal is on the border with p≈0.07), indicating that we only have evidence that these covariates are meaningful in classifying red and blue states. Furthermore, the coefficient for bio is over 20 times larger than coal. This indicates that bio is a much more important variable in classifying political affiliation. More generally, we can say that an increase in bio power (holding coal constant), increase the odds of a state being classified as blue while an increase in coal power (holding bio constant) leads to a decrease in the odds of being classified as red (reference state).

Table 2: Logistic Regression: Performance

Model	Misclassification Rate
Losistic Regression	0.14

The logistic regression model performs quite well with an accuracy of 86%, but still misclassifies 7 states: Alaska (AK), Colorado (CO), Idaho (ID), Illinois (IL), Nevada (NV), North Carolina (NC), and New Mexico (NM). Using Cook's Distance, CO, ID, and NM appear to deviate most from the model. It turns out that CO and NM, both blue states, generate very little electricity from bio and large amounts from coal compared to their counterparts. Similarly, ID, a red state, generates more bio and less coal in relation to other red states. Figures 5 and 6 illustrate this where the blue horizontal line is the average electricity generated for that particular source by blue states and the red horizontal line is the same, but for red states:

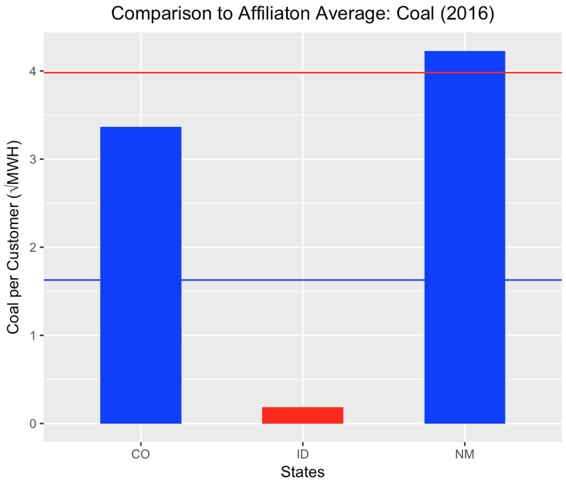

Figure 5: Percentile Rank: CO 69, ID 8, NM 81

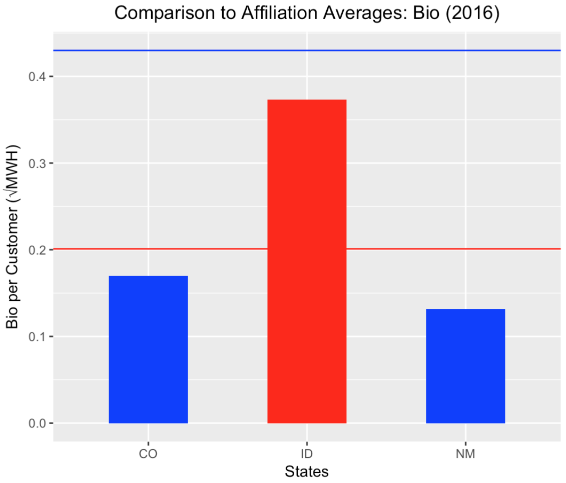

Figure 6: Percentile Rank: CO 20, ID 59, NM 16

In summary, the logistic regression performs well and is consistent with the hypothesis that blue states generate more electricity from bio and red states coal. Close examination of my model showed misclassication occurs when it comes across blue states generating large amounts of coal/little bio and red states generating a lot of bio/little coal.

3.2 Random Forest

In this section, I build a random forest model to see whether I can improve classification accuracy. As mentioned before, random forest could perform better if separation between red states and blue states is non-linear. To get started, I create a model containing all covariates. Next, I find the number of trees in which the error stabilizes, ntree=200. Finally, through tuning, I find that the the optimal number of covariates to choose from for splitting is three. The results are displayed below:

Table 3: Random Forest: Performance

Model	Misclassification Rate
Random Forest	0.17

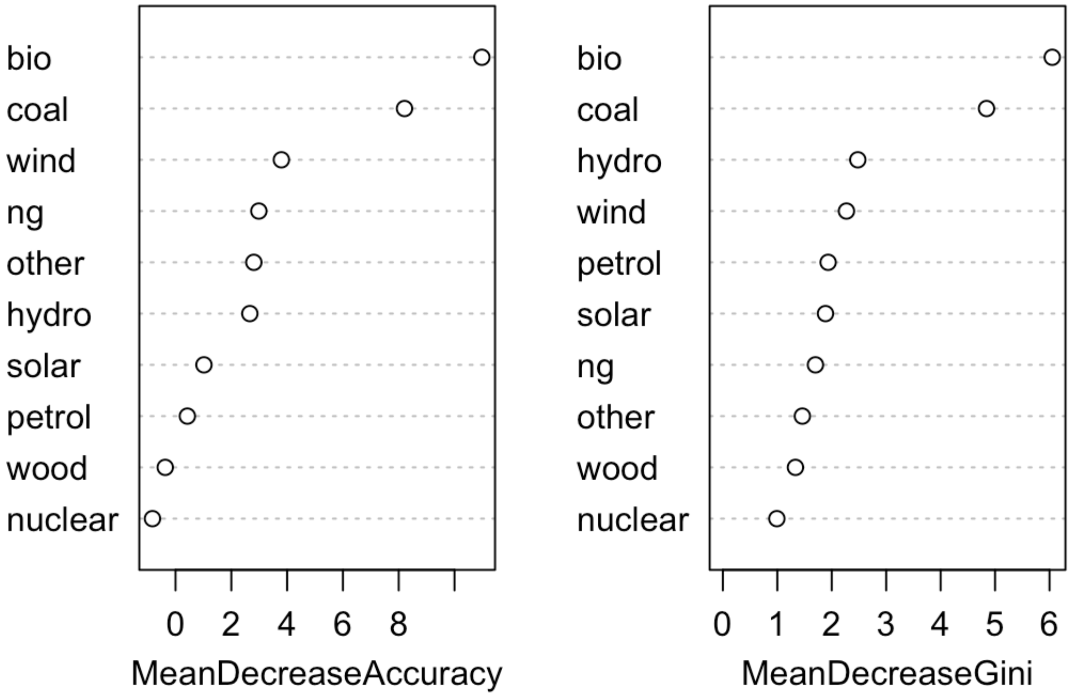

Figure 7

In terms of accuracy, the random forest model perform a bit more poorly compared to the logistic regression model. However, information wise, it conveys the same results. For instance, the mean decrease accuracy plot in Table 7 indicates that bio and coal are the most important variables for classification. That is, removing them from the model would greatly hinder accuracy. Moreover, the mean decrease in gini also attributes high importance to bio and coal since splits on these covariates lead to purer nodes. In conclusion, the random forest model does not add much extra information to the logistic regression model. However, because the results of the two models were so similar, it does provide further evidence that coal and bio differentiate blue states from red states.

4 Conclusion

Table 4: All Models:Performance

	Model	Misclassification Rate
1	Logistic Regression	0.14
2	Random Forest	0.17

In summary, both models yield similar conclusions, which provides strong evidence for the underlying model. In terms of performance, Table 4 shows that the logistic regression model performs slightly better than the random forest model. Ultimately, it seems that energy portfolios do a decent job (average misclassification rate of 15% for all models) in classifying whether a state is red or blue. Misclassification typically arises when the relationship between coal and bio is contradicted as was seen by states like Colorado, New Mexico, and Idaho in the previous section. A deeper dive indicates that the model does a poor job classifying states that either generate bio and coal power near the national average (Illinois, North Carolina) or generate little power from both sources (Nevada). In conclusion, this study makes a strong case that blue states generate more electricity from biomass than red states. It can be argued that red states generate more power from coal compared to blue states, but the argument is a bit more tenuous considering the significance level for the covariate *coal* in Table 1 of the logistic regression model.

References

Administration, U. E. I. (n.d.-a). Retrieved from https://www.eia.gov/electricity/data/state/

Administration, U. E. I. (n.d.-b). Retrieved from https://www.eia.gov/electricity/monthly/current_month/epm.pdf

Balaraman, K. (2017). *Red states rank among renewable energy leaders.* EE News. Retrieved from https://www.scientificamerican.com/article/red-states-rank-among-renewable-energy-leaders/

MIT. (2016). *U.s. president 1976–2016.* Harvard Dataverse. Retrieved from https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/42MVDX

Moore, S. (2015). *War on coal is war on red states.* Forbes. Retrieved from https://www.forbes.com/sites/stevemoore/2015/08/07/170/#6df9e8c120ed