Daniel Alfonsetti
Professor Hidalgo
17.831 Data and Politics
11 December 2019

Economics and Election Outcomes

Abstract

Can economic indicators predict when elections will result in changes of the party in power? That is the question I seek to answer in this project. To this end, I collected and merged 3 datasets: one showing gubernatorial election results, one showing GDP by state, and one showing unemployment by state. The overlapping years were from 1977 to 2010. I proceeded to create several derived variables such as moving averages and year to year percent changes for both GDP and unemployment. I then used a variety of methods to see if these economic indicators could predict if upcoming gubernatorial elections would result in the governor's party changing within both 4 year and 1 year time horizons. In addition to some simple t-tests and data visualizations, I trained random forest and logistic regression models using my economic indicators to try to predict such changes. When predicting if the governor's party will change within 4 years, the model that produced the most reasonable results in the training set was a random forest model that had about 69% accuracy on the test set. This is about 3% better than the baseline accuracy of a trivial prediction model that would always output no change. Counter intuitively, the models predicting whether change would occur next year did not perform as well. This was probably due to class imbalance in the training set. The results of this project seem to suggest that GDP and unemployment statistics do have at least somewhat of an affect on election outcomes, but not in a particularly strong or transparent way.

<u>Code Repository</u>
<u>Slide Show</u>

Hypotheses

A downwards trend in GDP and/or an upwards trend in unemployment will increase the chances of a governor party change in upcoming years.

The Data and the Data Cleaning Phase

All of the code corresponding to this section can be found in "/scripts/data_cleaning_and_exploration.Rmd". The raw data can be found in "/data/" and the processed, cleaned, and merged data can be found in "/data_processed/".

I began by extracting information from three different sources. The first source, from Harvard University's "Dataverse" website, contained gubernatorial election data from the 1960s to 2010.[1] I extracted the relevant columns from this dataset ("State", "Year", and "Years Since Other Party"). By adding a few derived variables to the dataset, I managed to create columns indicating whether if for a given year and a given state, the party of the governor would switch within the next four years. The same was done for a single year (next year) outlook.

The second source, taken from the Bureau of Labor Statistics, contained monthly unemployment and participation rate statistics for each state, ranging from 1976 to 2019.[2] I proceeded to average the monthly unemployment and participation rates for each year for each state.

The third data source, taken from the Bureau of Economic Analysis, had yearly real GDP data for each state, but there were two separate tables: one table contained the years from 1977 to 1997 in units of 1997 chained dollars while the other contained the years from 1997 to 2018 and was in 2012 chained dollars.[3] Fortunately, they datasets overlapped in the year 1997. This allowed me to fit a regression model from the 1997 GDP in 1997 chained dollars to the 1997 GDP in 2012 chained dollars. The fitted regression model thereby gave me a general way to convert 2012 chained dollars to 1997 chained dollars. Thus, I converted the table that had data from 1997 to 2018 in 2012 chained dollars to 1997 chained dollars. I then proceeded to merge the two tables which were now in the same units, thus deriving a table that had yearly GDP data from 1977 to 2012, all in 1997 chained dollars. I then added a column to my data indicating year-to-year GDP percent changes. I used percent changes instead of raw differences because some states have much larger GDPs and thus much larger absolute changes than other states.

After processing and cleaning each of the three data sources into formats I could use, I merged them together by state and year. I then proceeded to add trend indicators to each observation because my hypothesis is that economic trends affect party turnover in the governor seat and that standalone numbers like GDP and unemployment for a given state-year observation have little predictive power when taken out of their chronological context. To this end, I calculated 4 and 8 year moving averages for GDP, year-to-year GDP percent change, and unemployment. For each of

[1] https://dataverse.harvard.edu/dataset.xhtml?persistentId=hdl:1902.1/20408
[2] https://www.bls.gov/web/laus.supp.toc.htm
[3] https://apps.bea.gov/regional/downloadzip.cfm

these 3 variables, I also added a corresponding variable that calculated the difference between their 4 and 8 year moving averages. The idea behind 'difference of moving averages' variables is that, if they are negative, that means the 4 year moving average is lower than the 8 year moving average, indicating that - in the case of GDP - GDP is currently in a downtrend and that the state economy may be doing poorly. Here is what my final dataset looked like:

	state	year	change_next_year	change_within_4years	participation_rate	unemployment_rate	gdp_pchange	gdp	gdp_ma4_8_diff	gdp_pchange_ma4_8_diff	uer_ma4_8_diff
1	Alabama	1977	FALSE	FALSE	54.05833	7.316667	NA	59478.40	NA	NA	NA
2	Alabama	1978	FALSE	FALSE	55.32500	6.316667	6.03	63298.00	NA	NA	NA
3	Alabama	1979	FALSE	FALSE	55.10000	7.175000	2.60	64987.60	NA	NA	NA
4	Alabama	1980	FALSE	FALSE	53.76667	8.883333	-0.63	64582.30	NA	NA	NA
5	Alabama	1981	FALSE	FALSE	52.37500	10.533333	1.81	65771.00	NA	NA	NA
6	Alabama	1982	FALSE	FALSE	51.03333	14.083333	-2.55	64134.10	NA	NA	NA
7	Alabama	1983	FALSE	TRUE	52.00833	13.833333	4.71	67305.00	NA	NA	NA
8	Alabama	1984	FALSE	TRUE	54.14167	11.000000	4.51	70481.40	1918.15000	NA	2.469791667
9	Alabama	1985	FALSE	TRUE	55.16667	9.141667	4.50	73801.30	2135.36250	0.17000	1.893750000
10	Alabama	1986	TRUE	TRUE	56.00000	9.708333	1.01	74558.00	3333.83750	1.68750	0.376041667
11	Alabama	1987	FALSE	FALSE	56.59167	8.058333	5.21	78659.10	4463.42500	1.48625	-1.178125000
12	Alabama	1988	FALSE	FALSE	57.00000	7.191667	4.66	82507.50	5229.30000	0.86250	-1.918750000
13	Alabama	1989	FALSE	TRUE	57.36667	7.041667	-0.12	82407.70	5301.31250	-0.05125	-2.007291667
14	Alabama	1990	FALSE	TRUE	57.85000	6.791667	1.62	83766.40	5149.37500	-0.42000	-1.825000000
15	Alabama	1991	FALSE	TRUE	57.12500	7.350000	2.73	86115.50	4662.16250	-0.79250	-1.191666667
16	Alabama	1992	TRUE	TRUE	57.64167	7.566667	4.12	89813.90	4072.20000	-0.87875	-0.668750000
17	Alabama	1993	FALSE	TRUE	58.13333	7.366667	1.48	91167.60	4091.38750	-0.10125	-0.365625000
18	Alabama	1994	TRUE	TRUE	58.86667	6.191667	3.83	94803.10	4319.92500	0.09875	-0.076041667
19	Alabama	1995	FALSE	TRUE	59.75000	5.950000	3.29	98023.90	4876.42500	0.47875	-0.162500000
20	Alabama	1996	FALSE	TRUE	60.20833	5.225000	3.31	101379.30	5408.80000	0.44500	-0.502083333
21	Alabama	1997	FALSE	TRUE	61.22500	4.975000	3.27	104805.20	6018.51250	0.46875	-0.841666667
22	Alabama	1998	TRUE	TRUE	61.11667	4.391667	0.22	105031.76	5917.50750	-0.25875	-0.991666667
23	Alabama	1999	FALSE	TRUE	60.66667	4.733333	3.62	108978.92	5798.33500	-0.28750	-0.968750000
24	Alabama	2000	FALSE	TRUE	60.26667	4.575000	1.60	110749.80	5523.97250	-0.40000	-0.757291667
25	Alabama	2001	FALSE	TRUE	59.15833	5.100000	-0.10	110637.79	4548.34625	-1.04500	-0.442708333
26	Alabama	2002	TRUE	TRUE	58.21667	5.900000	2.82	113844.66	4371.37625	-0.26875	-0.029166667
27	Alabama	2003	FALSE	FALSE	58.22500	6.025000	2.53	116804.25	3980.16500	-0.44625	0.284375000
28	Alabama	2004	FALSE	FALSE	58.50833	5.700000	6.07	124355.43	4509.55625	0.32625	0.506250000
29	Alabama	2005	FALSE	FALSE	58.94167	4.500000	3.28	128575.66	6022.71625	1.17000	0.415625000
30	Alabama	2006	FALSE	FALSE	59.05833	4.066667	1.87	131024.50	7068.58375	0.72625	-0.002083333
31	Alabama	2007	FALSE	NA	58.70833	3.975000	0.58	131794.52	7964.20125	0.61875	-0.419791667
32	Alabama	2008	FALSE	NA	57.16667	5.716667	-0.31	131384.60	7142.14375	-0.73750	-0.558333333
33	Alabama	2009	FALSE	NA	53.18333	10.991667	-4.14	126159.84	4597.93250	-2.08750	0.328125000
34	Alabama	2010	NA	NA	53.25833	10.541667	2.36	129205.23	2223.04375	-1.90750	1.366666667
35	Alaska	1977	FALSE	FALSE	62.85000	9.866667	NA	16933.30	NA	NA	NA
36	Alaska	1978	FALSE	FALSE	63.78333	10.616667	8.22	18450.30	NA	NA	NA
37	Alaska	1979	FALSE	TRUE	64.58333	9.275000	4.70	19360.90	NA	NA	NA
38	Alaska	1980	FALSE	TRUE	64.17500	9.650000	13.72	22439.80	NA	NA	NA
39	Alaska	1981	FALSE	TRUE	64.91667	9.383333	13.24	25865.60	NA	NA	NA
40	Alaska	1982	TRUE	TRUE	65.57500	9.958333	2.55	26542.80	NA	NA	NA

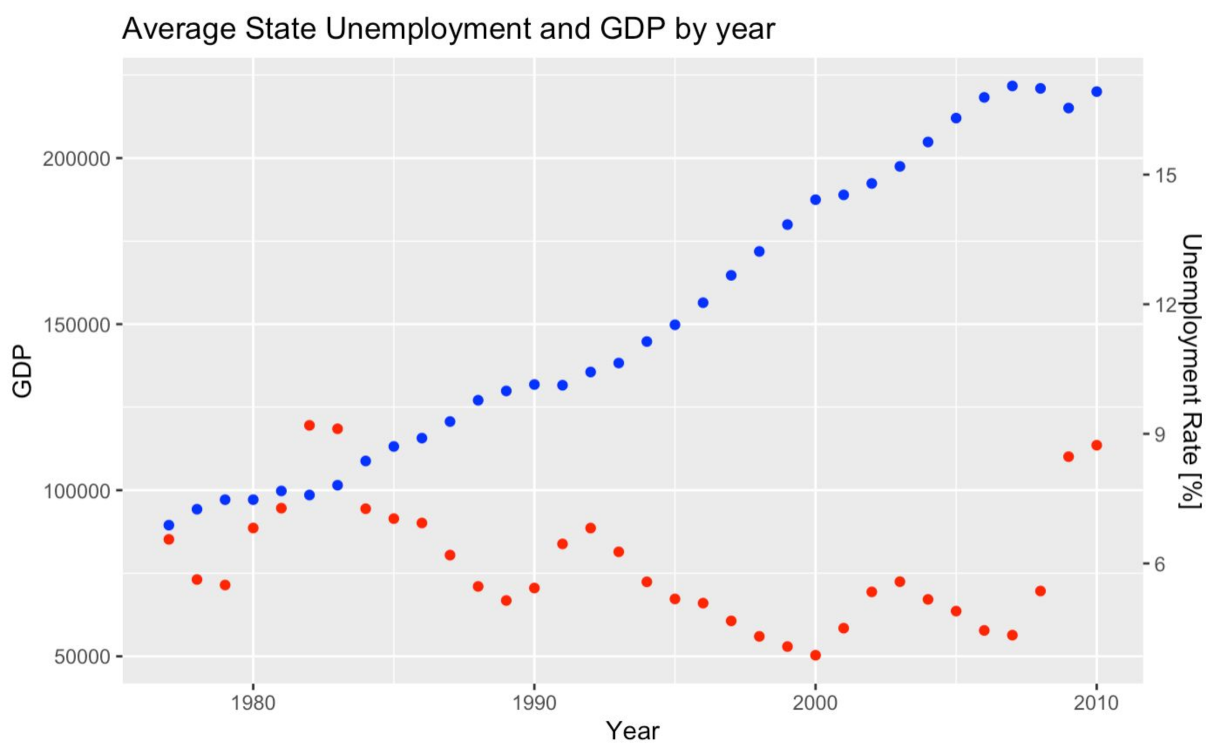
The more esoteric column names are defined below:

- "gdp_ma4_8diff" = Difference between the 4 and 8 year GDP moving averages
- "gdp_pchange_ma4_8diff" = Difference between the 4 and 8 year % GDP change moving averages
- "uer_ma_4_8diff" = Difference between the 4 and 8 year unemployment moving averages

Notice that it is impossible to calculate the 8 year moving average for GDP and unemployment for the first 7 years of observations for any state, and thus impossible to calculate the difference between the 4 and 8 year moving average for the first 7 years.

The same is true for GDP percent change, except the inability to calculate the difference extends an additional year since the raw GDP percent change cannot be calculated until the second observation for a given state. Additionally, observe that in the last 4 years on record for any state, it is impossible to tell (with this dataset alone) whether the governor's party will change within 4 years (see rows 31 to 34). The same logic applies for only the very last year when trying to predict if the party will change next year.

That completed my data wrangling and the construction of all the variables I was interested in. I then randomly split the data into an 80-20 training-test split, saving each into their own files for use later in predictive modeling. Before doing any modeling though, I wanted to verify that my processed data was reasonable looking, and that none of my transformations (such as converting from 2012 chained to 1997 chained dollars) hurt the data integrity. To this end, I plotted GDP and unemployment rates over time, yielding this graph:



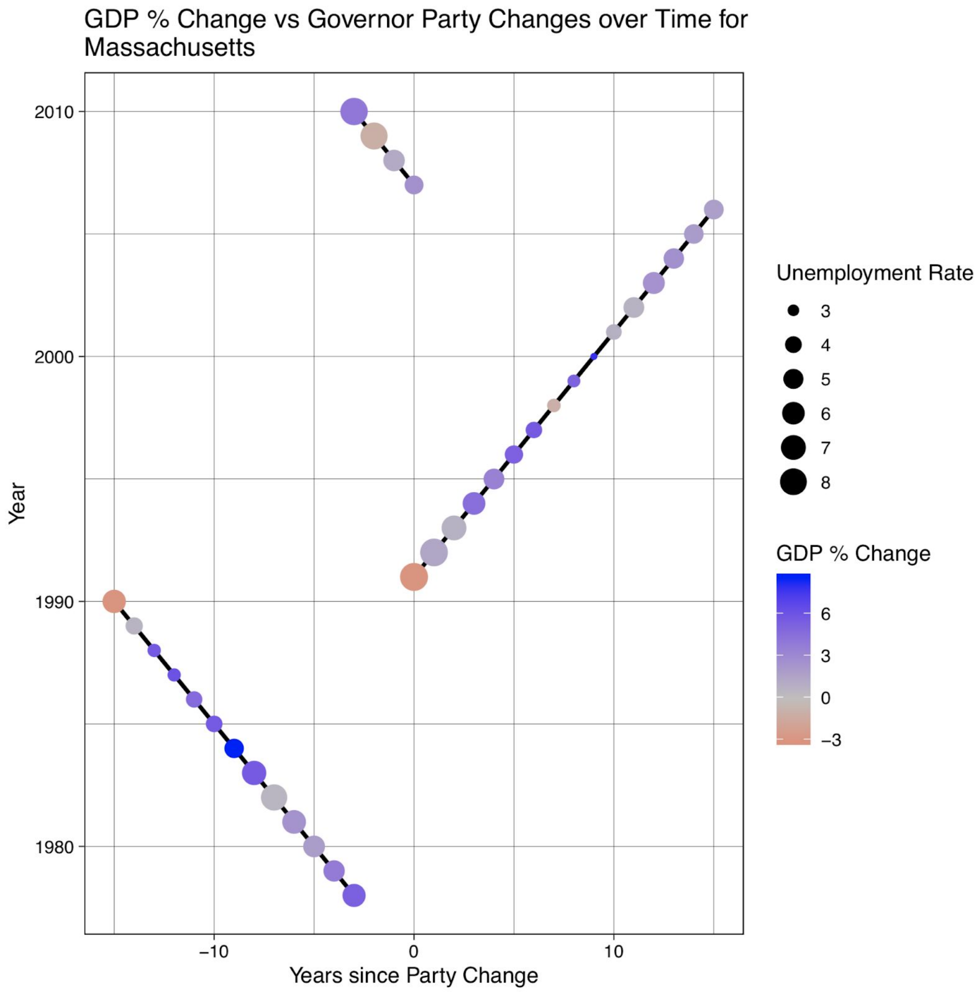
Through some cross referencing with wikipedia, the graph appeared to be consistent with the dates of when US recessions have occurred.[4] For example, you can see a relatively large downtick in GDP and an uptick in unemployment during the late 2000s when the housing crisis hit, as well as similar but smaller trend in the early 2000s

[4] https://en.wikipedia.org/wiki/List_of_recessions_in_the_United_States

after the dotcom bubble burst. Based on this, I was confident that my data was accurate.

Analysis

I began by doing some visual, qualitative analyses on the relation between GDP and unemployment with party changes. I made the following type of plot for each of the 50 states[5]:



Each circle corresponds to one row in the dataset. Years are on the y-axis and on the x-axis is an indication of how long it has been since the last time a governor of the opposite party was in office. If the circles are extending to the left, that shows that the Democrats are in power, and if the circles are extending to the right, that shows that the Republicans are in power. Encoded in the color of the circles is whether or not the GDP from the last year to the current year has increased or decreased.

[5] The plots can be found at "output/unemployment_rate_vs_governor_party_change.pdf".

If my hypothesis were correct, we would expect to see larger (higher unemployment) and/or more red circles (decreasing GDP) right before the switch in parties from 1990 to 1991 and from 2006 to 2007. This holds more so for the 1990 to 1991 switch than for the 2006 to 2007 switch.

After analyzing the analogous plots I made for the other states, I could not discern any obvious relationships that would validate or invalidate my hypothesis. I therefore turned to quantitative methods and ran some simple t-tests on my various economic indicators, partitioning my observations into groups based on whether or not the governor's party will change within four years. I then repeated the same t-tests except I based the groups on whether or not the governor party will change within 1 year instead of 4 years.

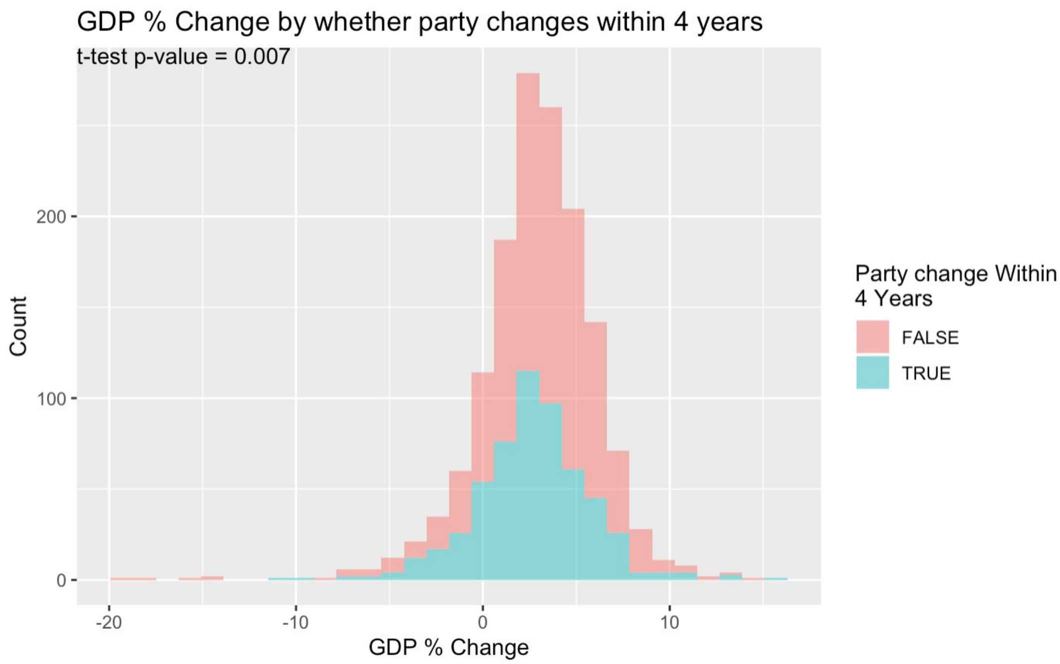
Partitioning by the 4 year time horizon, the unemployment rate between the two groups was statistically significant (p=0.003). However, when the groups were formed based on the one year horizon, there was no significant difference between their unemployment rates (p=0.142). This discrepancy probably stems from the fact that there is a severe class imbalance when partitioning on the one year time horizon (1340 observations where the party doesn't change next year vs 170 observations where it does change).

Performing a t-test on year-to-year GDP % change ("gdp_pchange") when grouping on either the one year or 4 year time horizon yield significant p-values. As seen below, the directionality of the differences is as expected (lower GDP and higher unemployment is associated with parties changing).

change_within_4years <lgl>	mean_unemployment <dbl>	mean_gdp_pchange <dbl>	count <int>
FALSE	5.737158	3.092022	937
TRUE	6.056661	2.635820	573

change_next_year <lgl>	mean_unemployment <dbl>	mean_gdp_pchange <dbl>	count <int>
FALSE	5.830274	3.024276	1340
TRUE	6.080098	2.111361	170

While the p-values are significant in several of the tests described above, plotting the distributions of the groups used in the t-tests was not as convincing, as exemplified in the graph below.



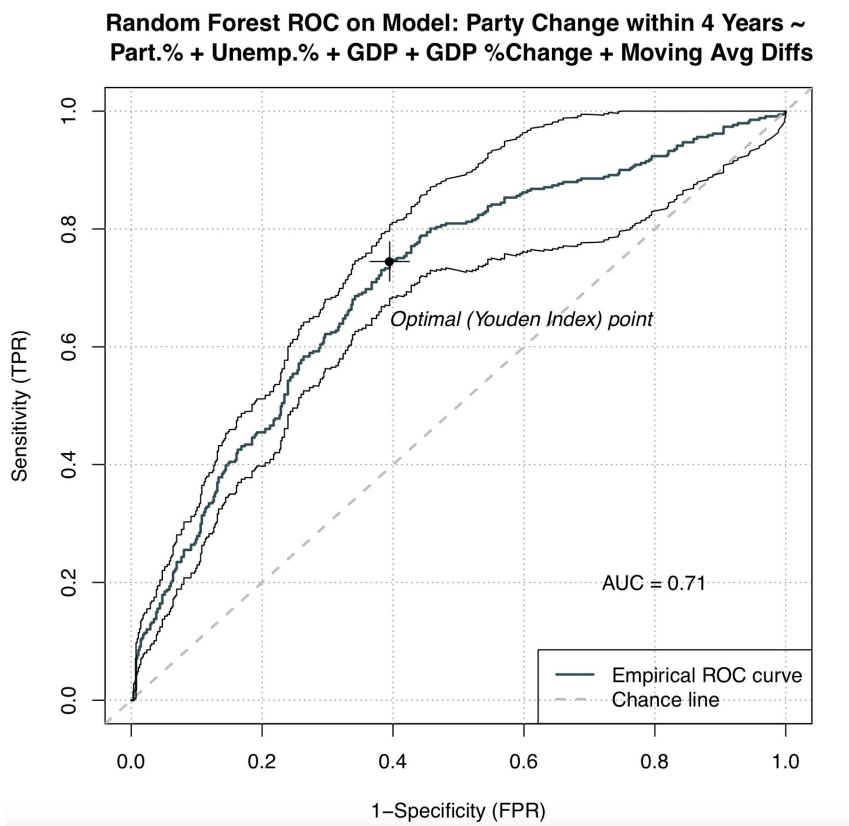
Model Training, Cross Validation, and Model Selection Phase

Regardless of the inconsistency between the t-test results and their corresponding visualizations, I wanted to see if I could train a model to predict turnover. I compared six different model variants. The six model variants came from me having two variables to predict (i.e. party change within 4 years and party change next year) as well as three sets of independent variables that I created. The three sets of independent variables were as follows:

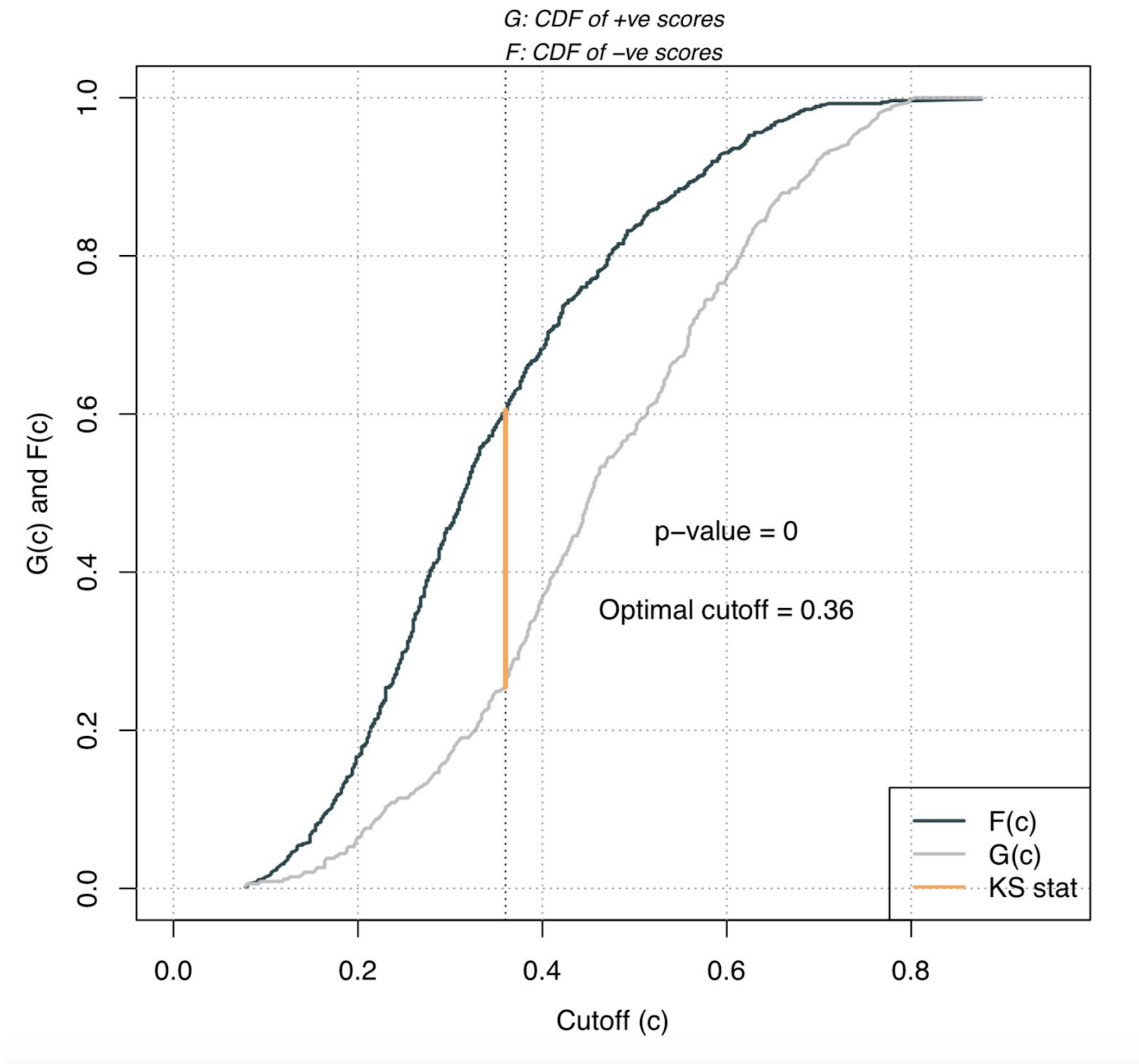
- Set 1:
  - Participation Rate
  - Unemployment Rate
  - GDP % Change
- Set 2: Set 1 variables, plus
  - Avg 4yr GDP % Change - Avg 8yr GDP % Change (Moving)
  - Avg 4yr Unemployment - Avg 8yr Unemployment (Moving)
- Set 3: Set 1 and 2 variables, plus
  - GDP
  - Avg 4yr GDP - Avg 8yr GDP (Moving)

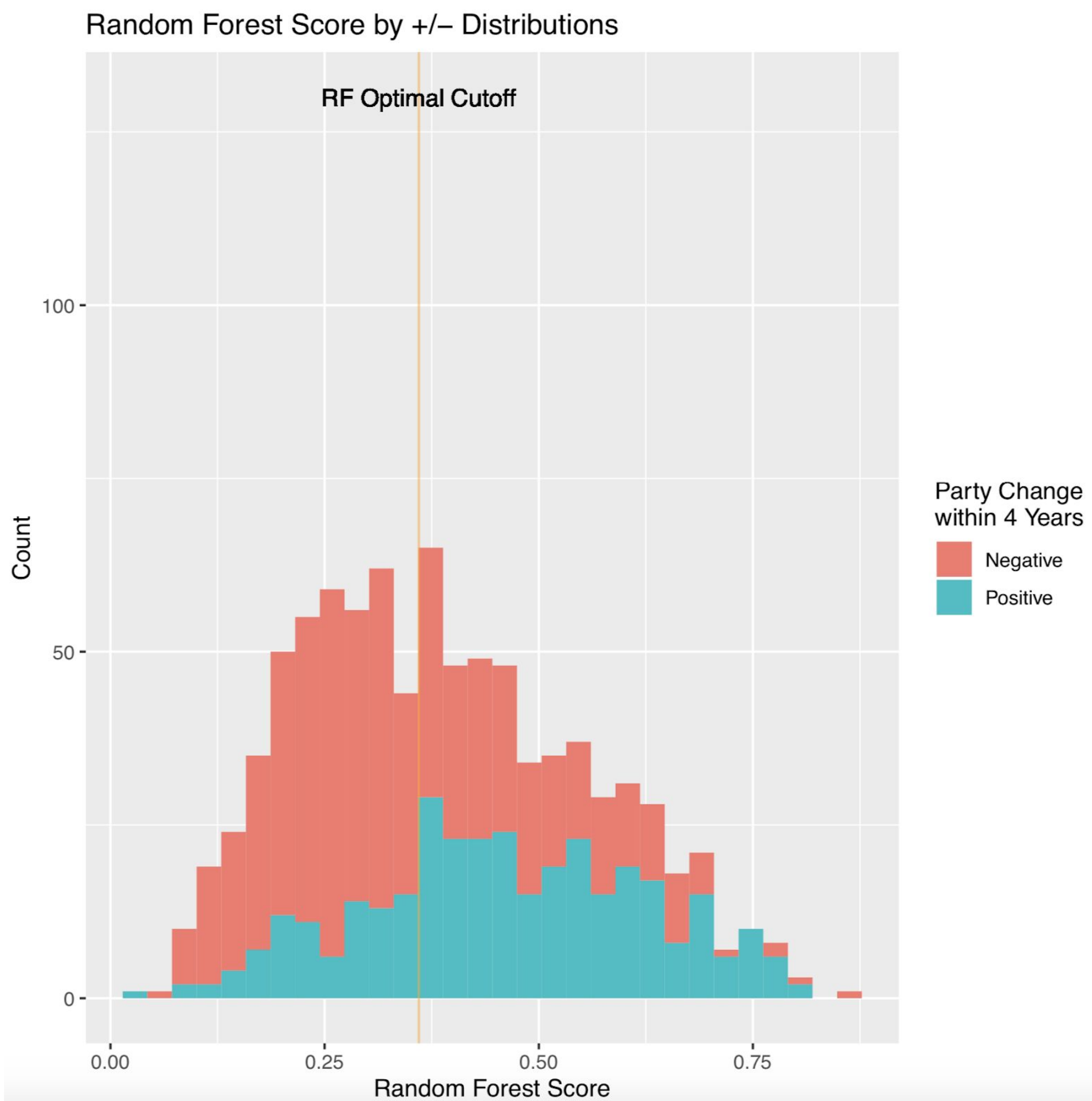
For each of these six model variants, I trained both a random forest and a logistic regression function using the training data I saved after cleaning the data. I then compared their performances using leave-one-out cross validation (LOOCV).[6] I found that the random forests generally had equal or better ROC plots (e.g. larger AUC), depending on the exact model used.

Somewhat unsurprisingly, the random forest with set 3 (i.e. all of the independent variables) did the best, but only when predicting on the four year horizon did the best. With so many variables, I was worried that the model might be overfitting. However, I figured that the leave one out cross validation should have helped reduce the potential for overfitting, so I went ahead and chose that model (model 5) anyways. The model outputs (for all six models) can be found in "/output/cross_val/" and the code training and cross validating each model can be found in "/scripts/". Below is an ROC curve (with 95% confidence intervals), a KS plot, and a histogram showing how well the best performing model (model 5) splits the training data:



6 https://stats.stackexchange.com/questions/61783/bias-and-variance-in-leave-one-out-vs-k-fold-cross-validation/252031#252031 Originally, after I completed my analysis, I was worried that LOOCV would introduce more variance than if I had done k-fold CV, but this discussion seems to suggest that may not be the case.

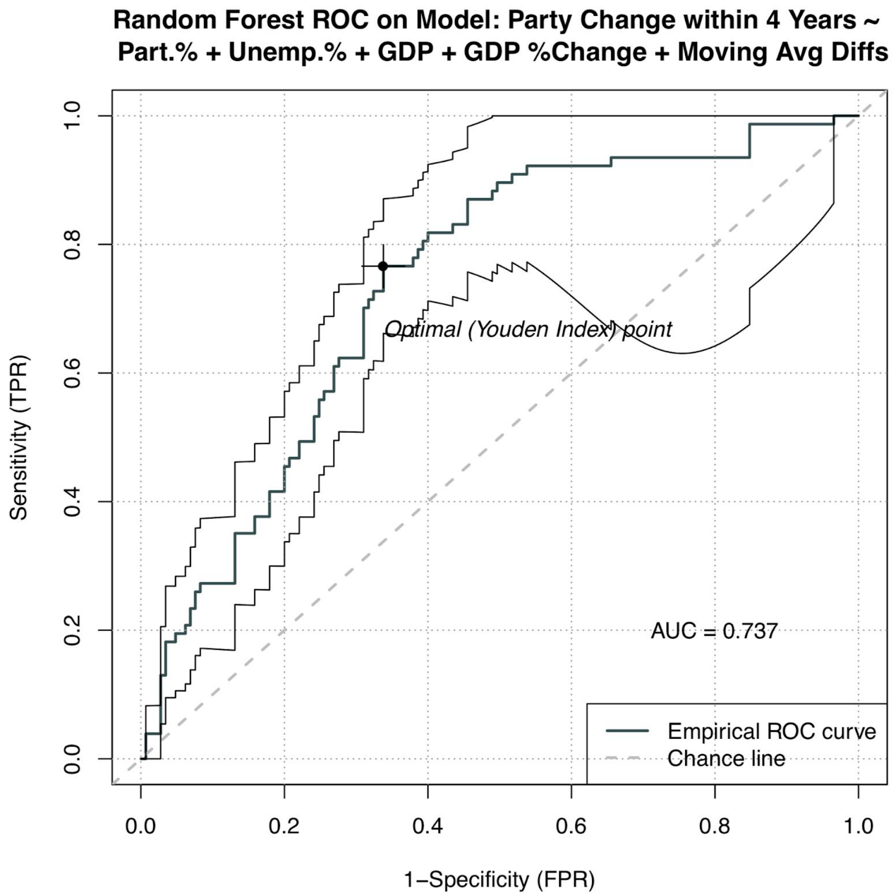

G: CDF of +ve scores
F: CDF of −ve scores
G(c) and F(c)
0.0
0.2
0.4
0.6
0.8
1.0
p−value = 0
Optimal cutoff = 0.36
F(c)
G(c)
KS stat
0.0
0.2
0.4
0.6
0.8
Cutoff (c)



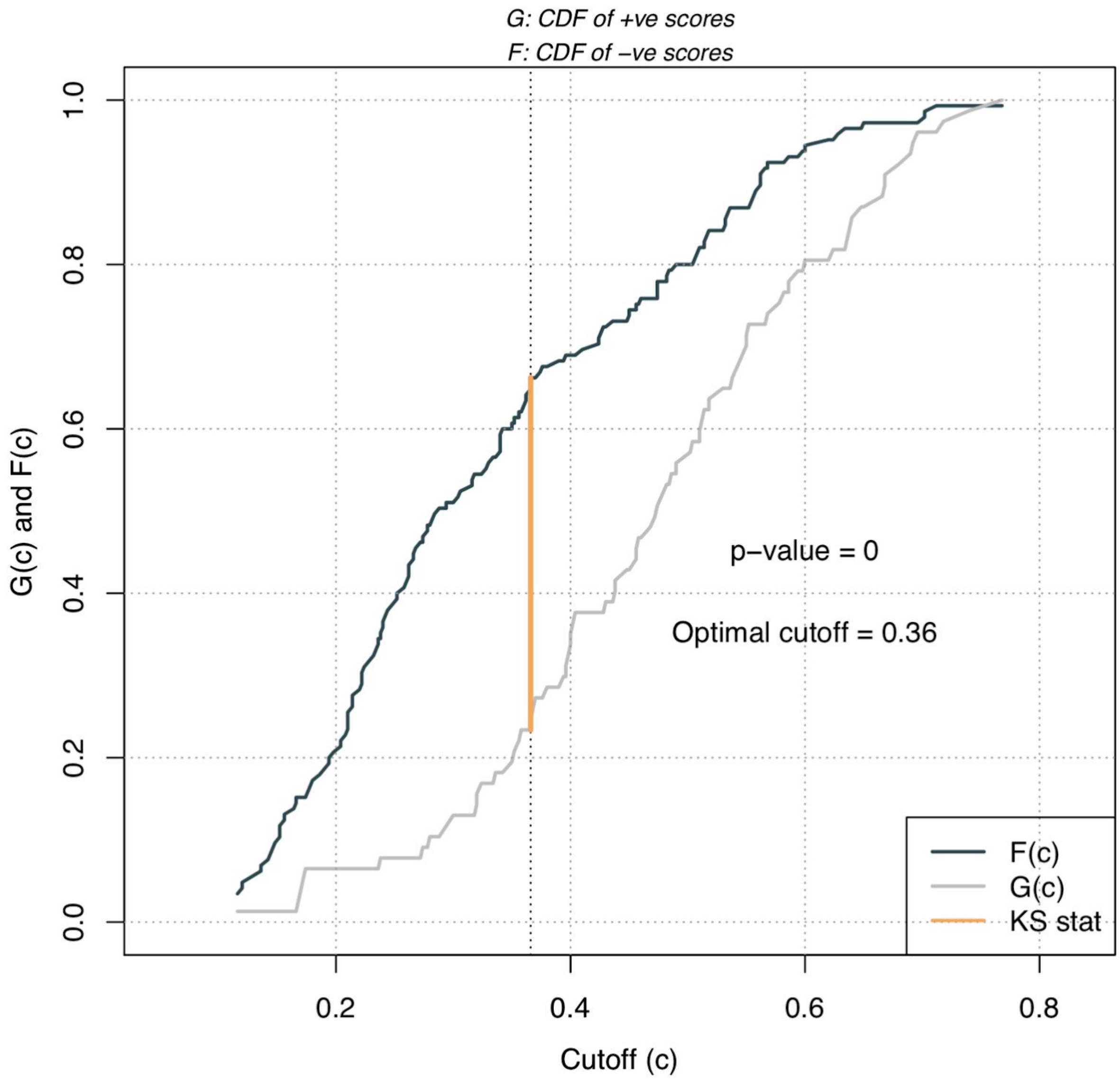
Contrary to expectations, the models predicting on the one year time horizon did worse in general than those predicting four years in advance. The only reasonable explanation for this is due to class imbalance. In the training set, the party “change within 4 years” variable had 547 negative instances and 341 true instances - a somewhat balanced composition - but the party ‘change next year’ variable was much more unbalanced, with 784 negative instances and only 104 true instances.

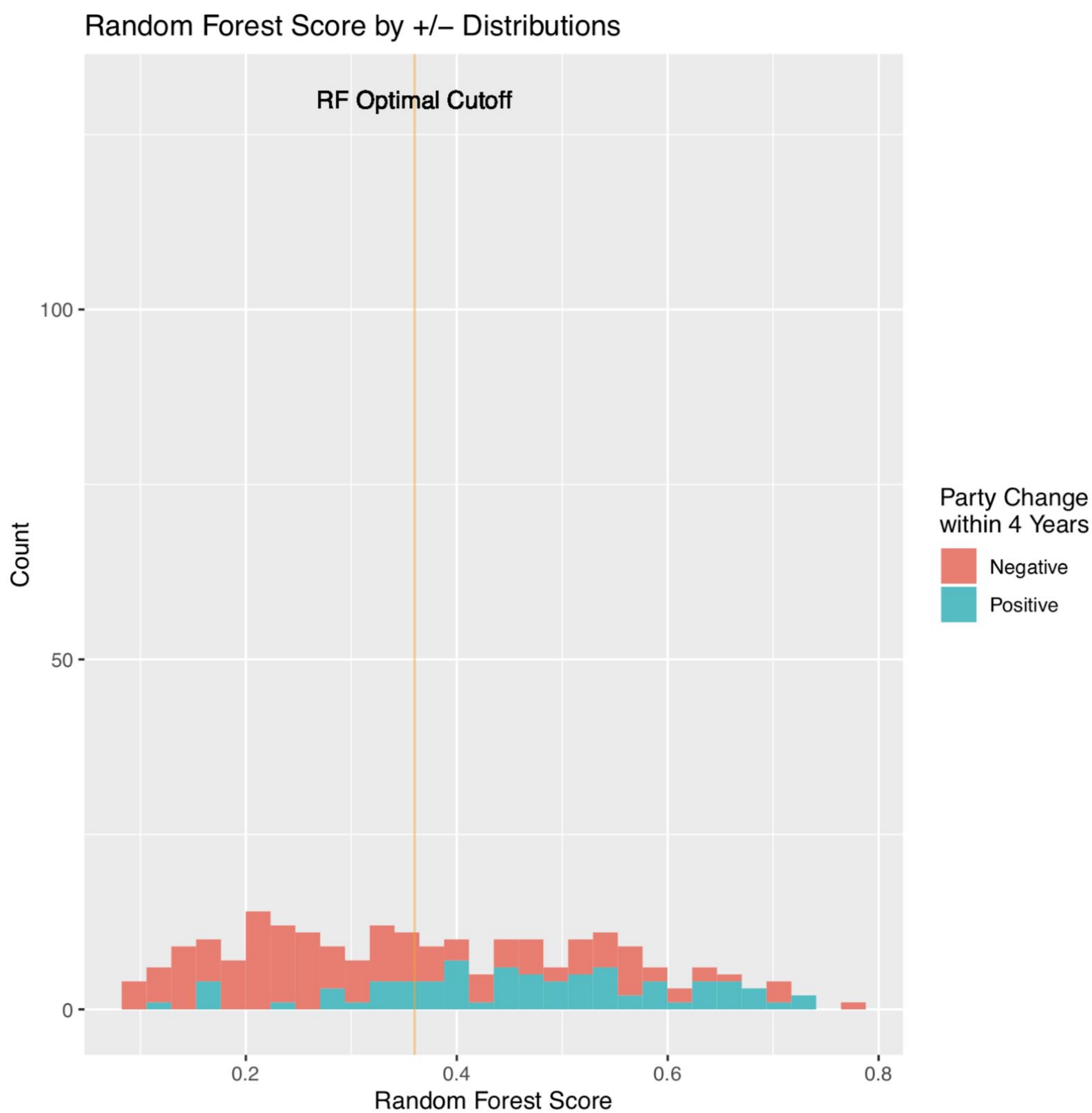
Testing Phase

As explained above, after training and using LOOCV, I chose the random forest model predicting party change on the four year time horizon using all of the independent variables (model 5) as my solution model. I proceeded to test this on the data that was held out for testing at the end of the data cleaning phase.[7] The code for this can be seen in "/scripts/test_selected_model.R". Below is the ROC plot, the KS plot, and a histogram showing how well the model splits the two categories in the testing data.



[7] Note that this test data is separate from the data that was used for cross validation. None of the models had seen this test data before. The cross validation data during the cross-validation and model selection phase came from the training set. In that phase, during each round of LOOCV, the training data was split into a validation set of 1 observation (the 'one left out') and then the rest was used as the *actual* training data for that particularly LOOCV iteration. In a somewhat philosophical sense, if we had just chosen the model that best fit the test data, our test data would have effectively become more training and validation data, and we would have introduce unintentional bias into our analysis. Testing only after we have committed to a model prevents this bias.


G: CDF of +ve scores
F: CDF of −ve scores
p−value = 0
Optimal cutoff = 0.36
F(c)
G(c)
KS stat
G(c) and F(c)
Cutoff (c)
0.0
0.2
0.4
0.6
0.8
1.0
0.2
0.4
0.6
0.8



Using the optimal cutoff[8] of 0.36, the confusion matrix is the following:

	actual	
predicted	Negative	Positive
Negative	92	18
Positive	53	59

[8] The optimal cutoff used on the test data was derived from the training set, *not* the test set. If we had used the optimal cutoff from the test set, we would be fitting to the test set and thus introducing bias.

The accuracy is 68.02%, which is 2.7% above the baseline accuracy of 65.32% yielded when we use a trivial prediction model that assumes all instances are negative instances (i.e. the party doesn't change within 4 years). This is not particularly amazing, but our model does beat the trivial model in the sense that it has a lower false negative rate (type II error):

Party doesn't change, but we predict it to change (type I): (53)/(92+53) = 36.55%

Party changes, but we predict it to not change (type II): (18)/(18+59) = 23.37%

Conclusions

In conclusion, the fact that we can train a random forest model to predict governor turnover with about 3% more accuracy than the baseline model suggests that there is some evidence that economic trends do affect when the party of the governor changes, but the effect is not particularly strong nor does it work through a particularly clear mechanism. Based on the preliminary analyses done alongside the t-tests, the direction of this effect captured by the random forest model is inline with our hypothesis that decreasing GDP and increasing unemployment increases the governor party turnover rate.

Challenges

There were challenges at each stage of the analysis. In the data gathering and cleaning phase, I found that the gubernatorial data table was not particularly intuitive or well organized, so extracting the correct information to use from it was difficult.

Another challenge was that I could only use years for which I had overlapping GDP, unemployment, and gubernatorial election data for. This substantially cut down on the number of observations I could train my models with, since the gubernatorial data went back as far as the 1920s while the GDP and unemployment data only started in the 1970s. Moreover, I had to further reduce the number of observations I could train with because I could not calculate eight year moving averages for GDP or unemployment during the first seven years of each state's economic record. Future directions for the project might therefore include trying to find a way to train the random forest and logistic regression models when some observations have omitted features values.

One of the biggest challenges, but probably the most educational, was making sure the statistical analyses were done correctly and that I was not unintentionally introducing biases into my model selection and testing methodologies. I referenced a lot of stack exchange articles to help me out in this regard.

Future Directions and Improvements

There are several things I would change if I were to redo the analysis. For one, there were many other columns in the gubernatorial data table that I did not use in my analysis, mostly due to a lot of the entries having missing information for those columns. Future work might look at incorporating those other pieces of information into the analysis, such as “gov_pty_change_since_last_budget” which encodes whether the governor party has changed since the last state budget was passed.

Some other things I would like to try include:

- Training models that had a notion of time inherently built into them, such as recurrent neural networks or hidden markov models.
- Add in moving averages and percent changes for participation rate.
- Find GDP and unemployment data going back before the 1970s.
- Test if our findings hold true for presidential elections (find more data for this).
- Correct “years_since_party_change” variable from the Harvard dataverse dataset for independent governors. Currently, each year an independent is in office, the variable is reset to 0.

References

1. https://www.investopedia.com/terms/c/civilian-labor-force.asp
2. https://towardsdatascience.com/random-forest-in-r-f66adf80ec9
3. https://machinelearningmastery.com/difference-test-validation-datasets/
4. https://en.wikipedia.org/wiki/Cross-validation_(statistics)#Leave-one-out_cross-validation
5. https://stats.stackexchange.com/questions/51416/k-fold-vs-monte-carlo-cross-validation
6. https://stats.stackexchange.com/questions/12412/how-do-you-generate-roc-curves-for-leave-one-out-cross-validation
7. https://stats.stackexchange.com/questions/90288/in-k-fold-cross-validation-does-the-training-subsample-include-test-set
8. https://towardsdatascience.com/why-linear-regression-is-not-suitable-for-binary-classification-c64457be8e28
9. https://towardsdatascience.com/an-implementation-and-explanation-of-the-random-forest-in-python-77bf308a9b76
10. https://elitedatascience.com/overfitting-in-machine-learning
11. https://dataverse.harvard.edu/dataset.xhtml?persistentId=hdl:1902.1/20408
12. https://www.bls.gov/web/laus.supp.toc.htm
13. https://apps.bea.gov/regional/downloadzip.cfm
14. https://stats.stackexchange.com/questions/61783/bias-and-variance-in-leave-one-out-vs-k-fold-cross-validation/252031#252031
15. https://en.wikipedia.org/wiki/List_of_recessions_in_the_United_States