Candidate of Change?

Did Unemployment Trends across America Contribute to the Rise of Donald Trump?

Laith Barakat
University of Cincinnati MS-BANA Capstone
First Reader – Dungang Liu
Second Reader – Ed Winkofsky

Table of Contents

Abstract

In the aftermath of a particularly polarizing United States presidential election cycle of 2016, political pundits and commentators have been resolutely torn on the factors leading to the outcome. While many theories have been tested, one compelling potential reason for Donald Trump's win in 2016 piqued interest for the research of this paper: that household economic performance and work force indicators significantly swung previously Democratic-leaning counties to vote Republican. This paper attempts to build simplistic logistic regression models around some such economic indicators at a county level. Once a complete model is achieved, the paper displays techniques that could be used as a foundation for more extensive and complex models, as well as an interesting application of basic statistical modeling within the political sphere.

Introduction

The United States presidential election of 2016 was one of the most polarizing and contentious elections in recent US history. Republican candidate Donald Trump, a newcomer to the political scene and previously a businessman and reality television star, had beaten Democrat Hillary Clinton, a career establishment politician and former First Lady. This unprecedented victory by Trump was underscored by the fact that Trump had not won the popular vote; his victory was won purely within the Electoral College. This victory is one that scholars and those within the political industry of the US have analyzed and speculated extensively since election night of 2016. Why did Americans trust such a newcomer over an experienced candidate with the highest executive job in the country? Amid claims of Russian internet propaganda spreading and a groundswell of support from irregular voter groups around the country, there seemed to be no clear answer.

Yet, as the first term of the Trump Administration was executed, another newcomer to the political world emerged, this time from the same party as Hillary Clinton and those stumped by her loss: Andrew Yang. Yang, a former corporate lawyer, businessman, and non-profit executive, launched a campaign for president in 2017 based on a premise previously not entertained by the rest of the Democratic establishment: that Donald Trump had won because of underlying economic problems within America, related closely to the automation of labor. Yang commented on the current political administration in the Democratic primary debate on February 7, 2020:

> *"Trump is not the cause of all of our problems and we are making a mistake when we act like he is. He is a symptom of a disease that has been building up in our communities for years and decades. It's our job to get to the harder work of actually curing the disease...Our communities have been disintegrating beneath our feet. That's why Iowa, a traditional swing state, went to Trump by almost 10 points. That's why Ohio, a traditional swing state, is now so red that I am told we are not even going to campaign there. These communities are seeing their way of life get blasted into smithereens. We have automated away four million manufacturing jobs and counting, we are closing 30 percent of New Hampshire stores and malls and Amazon, the force behind that, is literally paying zero taxes. These are the changes that Americans are seeing and feeling around us every day."*

https://www.newsweek.com/andrew-yang-donald-trump-symptom-disease-democratic-party-new-hampshire-debate-1486359

Problem Description

This compelling argument of economic disadvantages contributing to political results that Yang posits is the premise of this paper: the research done intends to pinpoint whether job loss and basic economic indicators (not directly related to automation) across America contributed to Donald Trump's ability to flip counties from Democratic-leaning to Republican-

leaning. Trump's message during his campaign was certainly a populist one, in which a defining theme was that the old jobs of America within the Rust Belt (heavily industrial northern Midwest of US) would be brought back from outsourcing and globalization (https://www.politico.com/story/2016/06/full-transcript-trump-job-plan-speech-224891). Indeed, Yang's rationale for job loss differs vastly from Trump's: automation and outsourcing are two quite massive macroeconomic concepts, and a comprehensive comparison of these phenomena is well beyond the scope of this paper. For this reason and others[1], this paper focuses only on general job loss at a county level in relation to election results.

The datasets employed by this analysis are open source and publicly available[2,3]. The first dataset of interest is the election result data, available from the Dataverse of Harvard. This data is presented as county (and county equivalent) level election results for every presidential election in the United States from the year 2000 through 2016. The other dataset is the labor statistics by county, publicly available from the Bureau of Labor Statistics. This data is published yearly. These datasets, and the proper transformation and joining of them, will be discussed more extensively in the subsequent "Exploratory Data Analysis" section.

Exploratory Data Analysis

The analysis taken in this research is undertaken in the statistical software R. In R, the data is transformed and combined; some initial visualizations for understanding the data are generated; generalized logistic regression models are created and assessed for strength. The import, transformation, and combination of the datasets are discussed in this section.

Labor Datasets

Upon import of the labor force data, one can observe firstly that the force data is by county and is the annual average of each year. Since these are individual datasets for each year, the years 2012-2016 were elected to import for labor trends and append all of the data. Researching as far back as 2012 gives perspective on the full political landscape leading up to the election in 2016 – Donald Trump's political presence before 2012 was minimal, so the juxtaposition of his message and labor trends prior to this year could be inferred as unclear in the mind of the voter. The data dictionary for each Labor Force dataset is identical as follows:

Element	Type	Description
LAUS Code	Character	Dataset-specific County code for each county
State FIPS Code	Character	Standardized Federal Information Processing Standards Code for the state

[1] Another primary reason for this project's focus on general job loss was a cost-prohibitive one: while a dataset for automation-exclusive job loss by US county does exist, it is outside of the budget of this project. Future work could include a more in-depth analysis of this.
[2] Election results dataset: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/42MVDX
[3] Bureau of Labor Statistics datasets: https://www.bls.gov/lau/#cntyaa

County FIPS Code	Character	Standardized Federal Information Processing Standards Code for the county
County Name/State Abbreviation	Character	Name of County and State
Year	Character	Year
Labor Force	Int	Total labor force in county
Employed	Int	Employed total of the labor force in county
Unemployed	Int	Unemployed total of the labor force in county
Unemployment Rate (%)	Int	Rounded percent unemployed from the labor force (unemployed / labor force)

For the purpose of this research, we will be using FIPS code as the primary key to link the election data with these labor datasets. Upon import of these datasets and proper binding of the separate sets, the master labor dataset gives all 50 states' county and county-equivalent labor statistic averages over the years 2012-2016[4].

Election Results Dataset

The Election results dataset from the Harvard Dataverse has a data dictionary described below, with 50,524 observations:

Element	Type	Description
Year	Int/Character	Year of election results reported
State	Factor	State of reported election results
State_po	Factor	Abbreviated state of reported election results
County	Factor	County of reported election results
FIPS	Int/Character	Standardized Federal Information Processing Standards Code for the state and county
Office	Factor	Election office for reported results; "president" for all
Candidate	Factor	Name of Candidate for reported results
Party	Factor	Name of party for reported results
Candidatevotes	Int	Votes for candidate
Totalvotes	Int	Total votes in the county
version	Int	Version of dataset

After filtering results and eliminating unnecessary columns, the resulting data frame has FIPS code, year, state, county, 'Barack Obama', 'Donald Trump', 'Hillary Clinton', 'Mitt Romney', and Other[5]. Therefore, we now have a table with each county's election results by year (for 2012 and 2016) and showing the number of votes for each candidate in each observation.

[4] The Bureau of Labor Statistics' datasets also include the 78 municipalities of Puerto Rico.
[5] Each of the elements with the names have the number of votes for each candidate

Combining Datasets

To combine these datasets, the data was needed to optimize the foreign key FIPS for each table and commit a full join on the identical FIPS keys. The resulting table provides all of the relevant data, but the years were still needed to be spread across the labor statistics so that the data is long. With multiple columns needed to spread the values through, a custom spread function is created to spread properly. The Year variable could then be spread through the labor force, employed, and unemployed fields to transform the data. This result table provides each county as one observation, with 2016 votes for each candidate, 2012 votes for each candidate, and labor statistics for 2012-2016 in each observation. Once these transformation steps are completed, one has the proper data frame to proceed with analysis.

Binary Variables and Data Exploration

Initializing new variables is an important next step before proceeding; it was assumed that a county "win" is constituted by a higher vote count for one candidate over the other[6]. Therefore, there are three important binary variables to initialize:

A Trump win in a county: If more Trump votes than Hillary votes, 1, otherwise, 0

An Obama win in a county: If more Obama votes than Romney votes, 1, otherwise, 0

A Trump county flip: If Obama won a county in 2012, but Trump won in 2016, 1, otherwise, 0

It was then intended to check the aggregate values for popular vote of each candidate. Intuitively, it would follow that the sum of all of the county votes would approximately equal the official popular vote tally for each candidate. Using *tidyverse*'s *summarize()* function, the total vote tallies based on the data are below:

Candidate	Data Vote Sum	Official Popular Vote Count[7]
Mitt Romney	60,748,718	60,933,504
Barack Obama	65,861,582	65,915,795
Hillary Clinton	65,851,549	65,853,514
Donald Trump	62,980,433	62,984,828

The data figures are all close to the official popular vote, so analysis could proceed. It has been theorized that the discrepancies are due to a combination of expatriate (non-state affiliated votes) and NA values within the election data. There is reason to focus on the change of unemployment in each county from year to year. Thus, another variable was initialized:

[6] Of course, this does not account for third party votes or the nuance of the actual American voting system; what is being targeted is general population sentiment toward Trump messaging and campaigning on a county level. If a county was previously an Obama-leaning county and becomes a Trump-leaning county, the research hopes to tease that out of the data and identify that as an example of Yang's message quoted in the introduction.

[7] https://en.wikipedia.org/wiki/2012_United_States_presidential_election

unemp_growth_12to13 = (2013_Unemployed - 2012_Unemployed) / 2013_Force

This variable was duplicated for the following years, so that four variables were created: 2012 to 2013 unemployment growth, 2013 to 2014 unemployment growth, 2014 to 2015 unemployment growth, and 2015 to 2016 unemployment growth. These are percentage growth variables, as huge outliers and differences in population size are recognized across counties. The summaries for these variables are Figure 1 in the Appendix. It can be observed that none of these variables have a positive mean or median for the unemployment growth rate. Overall, it can be inferred that total unemployment in this 5 year period was on a favorable trend throughout the United States. This is not conducive to the research assumption; however, it does not disprove it.

Another piece of exploratory analysis that the research attempted to seek was the percent of counties won by Trump and Obama, as well as the percent of counties flipped by Trump. These results are shown in Figure 2 of the Appendix. It is apparent that Trump won a much higher percentage of counties than Obama. This, when checking against official election results[8] , displays an accurate account of the vote breakdown. This staggering difference in the percent of counties won by each of the two winners in 2012 and 2016 (22.6% for Obama[9] versus 83.9% for Trump) demonstrates the divide commonly experienced in modern American politics: Democratic candidates, on the whole, win a lower percentage of the counties, but win more of the more populated counties. Conservatives, on the contrary, deliver across less densely populated counties and swallow up many of more counties on the whole.

Histograms of the unemployment growth variables are displayed in Figure 3 of the Appendix. The pairwise correlation of each of these unemployment growth variables may also require to be checked. The pairwise correlation matrix is shown in Figure 4 of the Appendix. It is safe to say that none of the correlation coefficients are unreasonably high, and that these variables will not experience much multicollinearity.

Logistic Regression Models

Multiple approaches were taken during the regression research phase. The overall goal of the project became more largely scoped as it was realized that multiple variables could be employed to indicate employment health on a county basis. It became clear that observing counties that Trump flipped could be influenced by more than just growth or decline in employment over time (the overall economic and earnings health of a county can be summarized by many indicators, and this project is a proverbial tip of the iceberg in terms of

[8] https://web.archive.org/web/20161207142233/http://bigstory.ap.org/article/fb5a5f7da21d460bbffb6985cb01cb2c/trending-story-clinton-won-just-57-counties-untrue

[9] Article to verify 2012 result figures: http://www.nbcnews.com/id/50073771/t/obama-won-record-low-share-us-counties-he-won-them-big/#.XpYDWMhKjic

analysis that can be done in that regard). The following approaches were used as separate models, and then all combined into one final model:

1. the unemployment growth variables that were created to build generalized linear models to predict counties that Trump flipped
2. a flat unemployment rate by county in the election year to predict counties that Trump flipped
3. indicator variables for each state (dummy variables) to predict counties that Trump flipped due to which state the county is in
4. an overall average income by county for 2016 to predict counties that Trump flipped

Approach 1 utilizes stepwise variable selection and testing set cross-validation to determine the best models. Since there is only a single variable in question for approaches 2, 3, and 4, no variable selection was utilized; however, test set cross-validation was still employed to determine if models were significant. Multiple model link types, namely logit, probit, and loglog, were explored for each approach.

Unemployment Growth over Time Models

This first set of models was conducted on the unemployment growth rate variables previously mentioned. Forward stepwise variable selection was conducted to limit the number of variables included, employing both AIC and BIC as selection criteria for thorough selection. In both the AIC and BIC selection cases, a model with all four unemployment growth rate variables was selected as the optimal model. Interestingly, only the 2014-2015 unemployment growth variable was statistically significant in terms of a Wald test p-value approach. However, it was elected that all of the variables, regardless of significance, be included, since multicollinearity had been proven to be of little influence in this data.

The results of this model were unencouraging. The ROC curves for both the training data (randomly sampled 80% of the full dataset) and the testing data (the remaining 20% of the dataset) are shown in Figure 5 of the Appendix. The area under the curve (AUC) estimates for these curves are roughly 56% and 65%, respectively. These results are conclusively below the standard 70% AUC values that would be necessary for a “good” model. When reconstructing the models under probit and loglog link functions instead of standard logit, the AUC values do not improve and are similar values.

Unemployment Rate by County at Election Year Model

The next approach uses only one variable, the unemployment rate based on the labor statistics in the year 2016 for each county. In theory, this variable could represent the general sentiment of each county’s economic health. Counties with higher unemployment in the election year could, based on this project’s hypothesis, be more highly susceptible to the Trump campaign’s messaging. Yet, this hypothesis can seemingly be rejected: the AUC estimate for the training dataset is around 48%, a figure that is quite low, and the estimate for the testing dataset is 60%. Once again, changing the link functions to probit and loglog functions do not

change the AUC values significantly. The ROC curves for the logit model training and testing sets are contained in Figure 6 of the Appendix.

State Model

It was recognized that a model that predicts whether a county was flipped by Trump based on the state it is contained in is not a groundbreakingly insightful model, especially given the relative partisanship of state-level voting trends in the United States. Yet this paper hopes to achieve at least one model of acceptable accuracy throughout the course of this project. The factor variable that identifies each county's state being the sole predictor yields a model with much better performance than any previous models. The AUC for the training dataset is 85%, and the AUC for the testing dataset is 77%. These AUC values are much more indicative of a well-performing model (Figure 7 of Appendix).

Of course, a model that predicts a Trump-flipped county based on the state brings some insight: among the most significant factor variables in the model are the indicator variables for Iowa, Michigan, and Wisconsin – swing states that Trump notably flipped in 2016. It is recognized that this is a redundant insight; nevertheless, state-level election results are being teased out of the county-level data, which is a positive outcome from the research.

Median Income in 2016 Model

The final model standalone included in the research scope was a model built with the Trump flip indicator variable as the response variable, with county median household income as the predictor. The figures for county level median household income were not provided in the initial two datasets used, but were found within the US Census Bureau's openly available datasets[10]. This data was easily joined through the FIPS key and included in the working master dataset. The theory behind using this predictor variable was to glean whether economically disadvantaged or lower-income counties had a higher susceptibility for being flipped by Trump.

The results of the simplistic logistic regression model suggest otherwise: the predictor variable yielded a statistically insignificant p-value, and the AUC of the ROC curve for the training data was in the 48% range (Figure 8 of Appendix). Conclusively, this general hypothesis is not supported by simplistic models – there is not initial evidence to say that lower income counties had higher likelihood to flip from Democratic-leaning voting trends to Republican-leaning trends.

Composite Model

Once each individual model has yielded little to no results, a combined model with all of the variables included can be tested. This model is constructed similarly to all of the rest, including unemployment growth rates for all four years, unemployment rate in 2016, indicator variables for each state, and median income of each county. Interestingly, the results are much

[10]Median income dataset: https://www.census.gov/data/datasets/2016/demo/saipe/2016-state-and-county.html

more successful. The training AUC is 85%, and the testing AUC is 82%. The ROC curves are included in the Appendix under Figure 9.

This model is a combined model that can be used to proceed to find an optimal cutoff probability for the prediction. First using the naïve method to determine the probability cutoff of .07, the confusion matrix is listed in Figure 9. Using the grid-search method and determining a probability cutoff of .16, the confusion matrix is listed below the grid-search plot in Figure 9. The training misclassification rate for this probability cutoff is 0.13.

Conclusions

The results of the regression models resoundingly signal one comprehensive conclusion: vast generalizations about election trends on a national scale can often not be easily translated in simplistic, one-dimensional statistical models, but can be achieved through a combination of multiple variables and factors. The level of complexity in the United States' voting system, combined with the diversity of demographic and socioeconomic influences on voting, necessitate models that match those levels of complexity. Andrew Yang's original statement about job loss and economic disadvantage being main contributors to Donald Trump's historic election win was not disproven by this project, but the simplistic, one-variable models and methods used within the scope of this project certainly do not display strong support for a systemic trend in that direction. However, when a robust, composite model is generated using a combination of variables, it can be seen that there is conclusive evidence that some combination of unemployment rates, states, and income figures can inform and predict whether Donald Trump flipped a county from a Democrat-voting county to a Republican-voting county.

Future Work

This section serves as brainstorm for other offshoots of projects that can be explored from this data or beyond what could be accomplished with the project's scope. One primary action that could be taken by researchers is to investigate different data sources and incorporate new data, such as leadup polls to the election, more nuanced economic indicators, and demographic indicators within voting trends. Of course, more nuanced models could also be employed: a composite model based on multiple sets of variables, more complex methods like classification trees, and more detailed analysis (perhaps on a state-by-state basis) could be undertaken as well. These future actions could inform a more highly performing end model that more aligns with the original hypothesis and with Yang's statements and outlook. A last suggestion for any interested reader is to explore similar projects undertaken by other researchers, especially one Oxford Review study that demonstrates the conclusion at which Yang arrived[11].

[11] https://academic.oup.com/oxrep/article-abstract/34/3/418/5047377

Appendix

Figure 1: Summary statistics of Unemployment growth rate variables

unemp_growth_12to13	unemp_growth_13to14	unemp_growth_14to15	unemp_growth_15to16
Min. :-0.04292	Min. :-0.06518	Min. :-0.04077	Min. :-0.03990
1st Qu.:-0.00940	1st Qu.:-0.01676	1st Qu.:-0.01123	1st Qu.:-0.00610
Median :-0.00470	Median :-0.01106	Median :-0.00735	Median :-0.00313
Mean :-0.00561	Mean :-0.01203	Mean :-0.00739	Mean :-0.00296
3rd Qu.:-0.00165	3rd Qu.:-0.00635	3rd Qu.:-0.00347	3rd Qu.: 0.00014
Max. : 0.05881	Max. : 0.01170	Max. : 0.03577	Max. : 0.03241
NA's :50	NA's :50	NA's :50	NA's :50

Figure 2: Aggregate percentage of counties that Trump won, counties that Obama won, and counties that Trump flipped

CountyPercentTrumpWon	CountyPercentObamaWon	CountyPercentTrumpFlip
0.8386076	0.2262658	0.07181272

Figure 3: Histograms of Unemployment growth variables

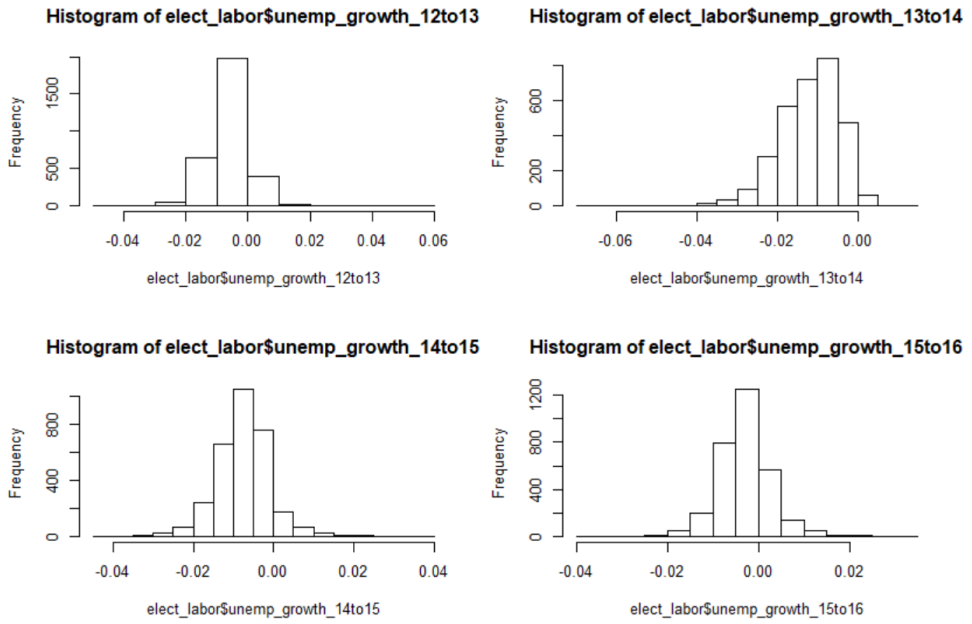
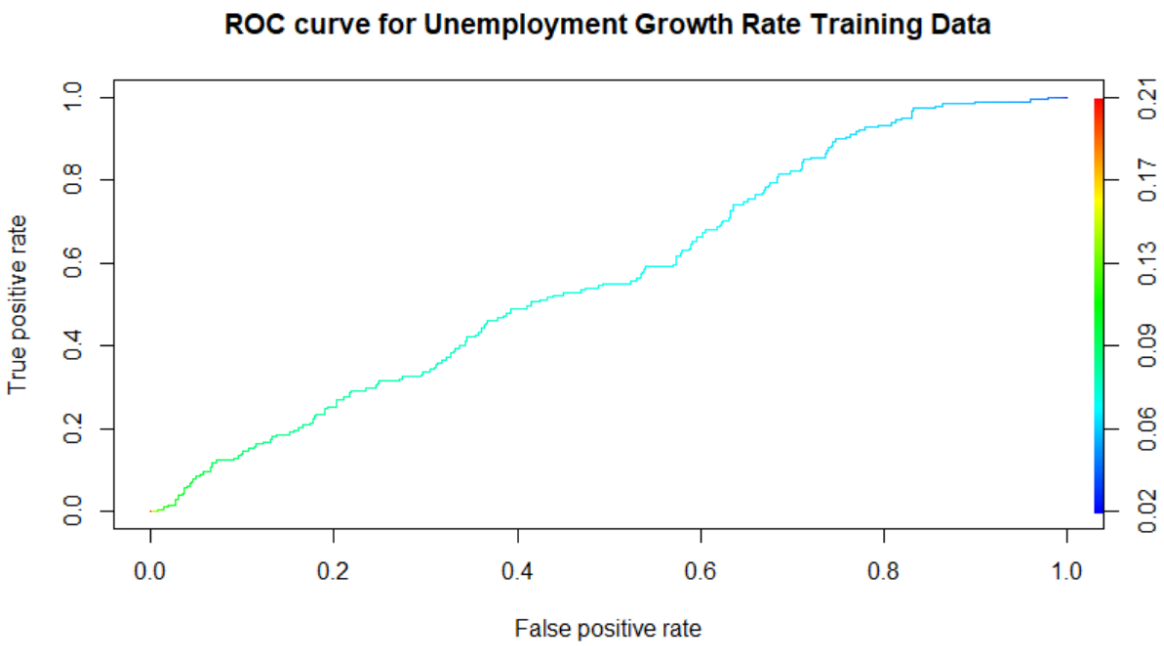


Figure 4: Correlations of Unemployment growth variables

	unemp_growth_12to13	unemp_growth_13to14	unemp_growth_14to15	unemp_growth_15to16
unemp_growth_12to13	1.0000000	0.1155569	0.1058161	0.3507235
unemp_growth_13to14	0.1155569	1.0000000	0.5292753	0.1890059
unemp_growth_14to15	0.1058161	0.5292753	1.0000000	0.4076166
unemp_growth_15to16	0.3507235	0.1890059	0.4076166	1.0000000

Figure 5: ROC Curves for Unemployment Growth Rate models



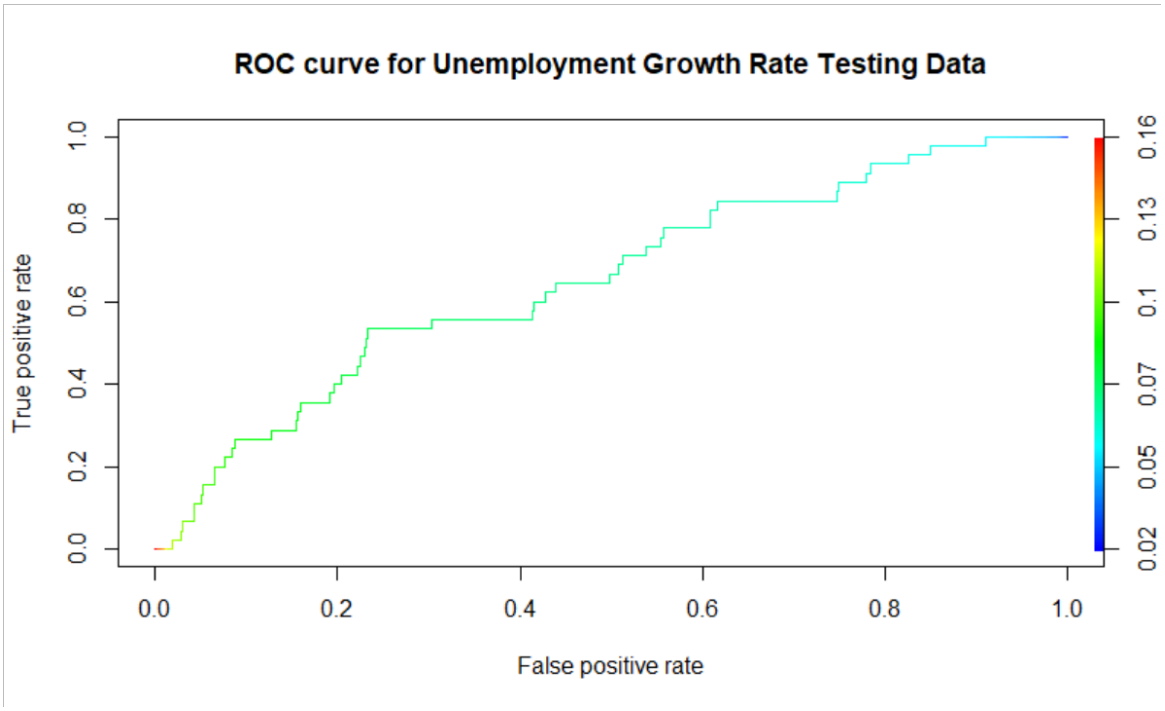
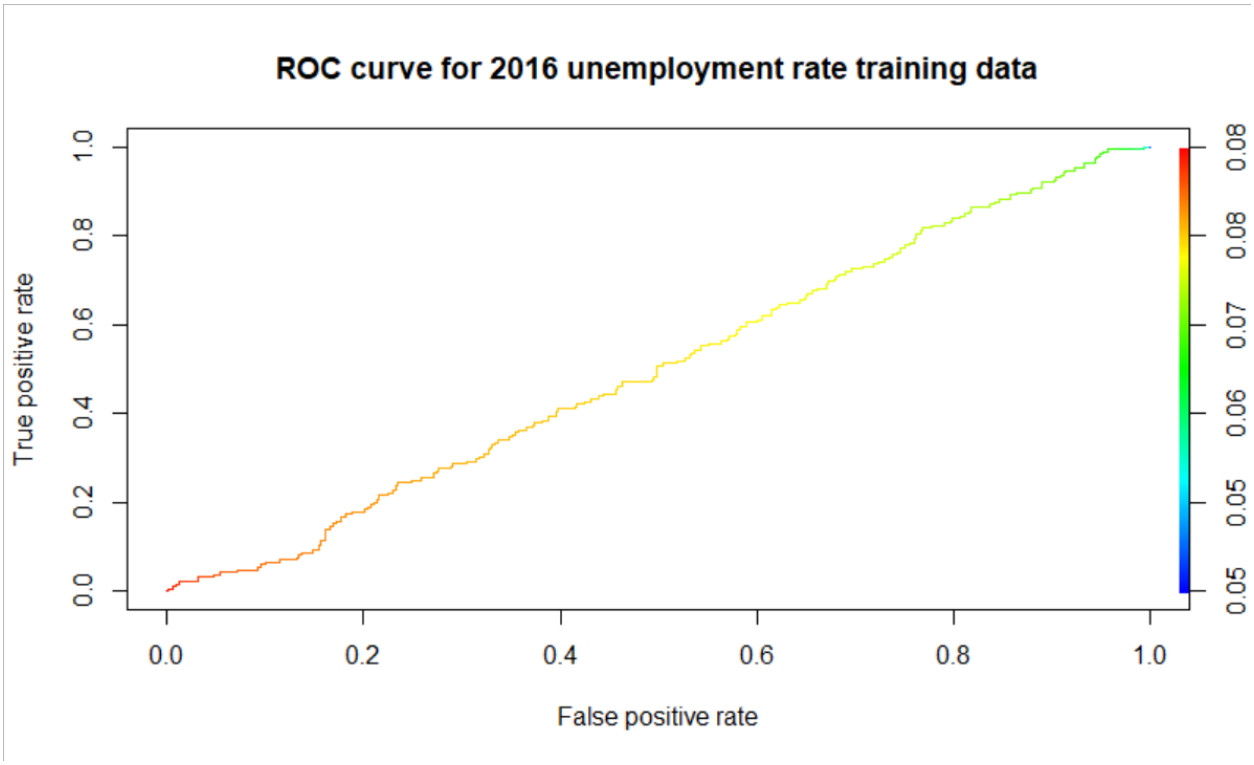


Figure 6: ROC Curves for 2016 Unemployment Rate Models



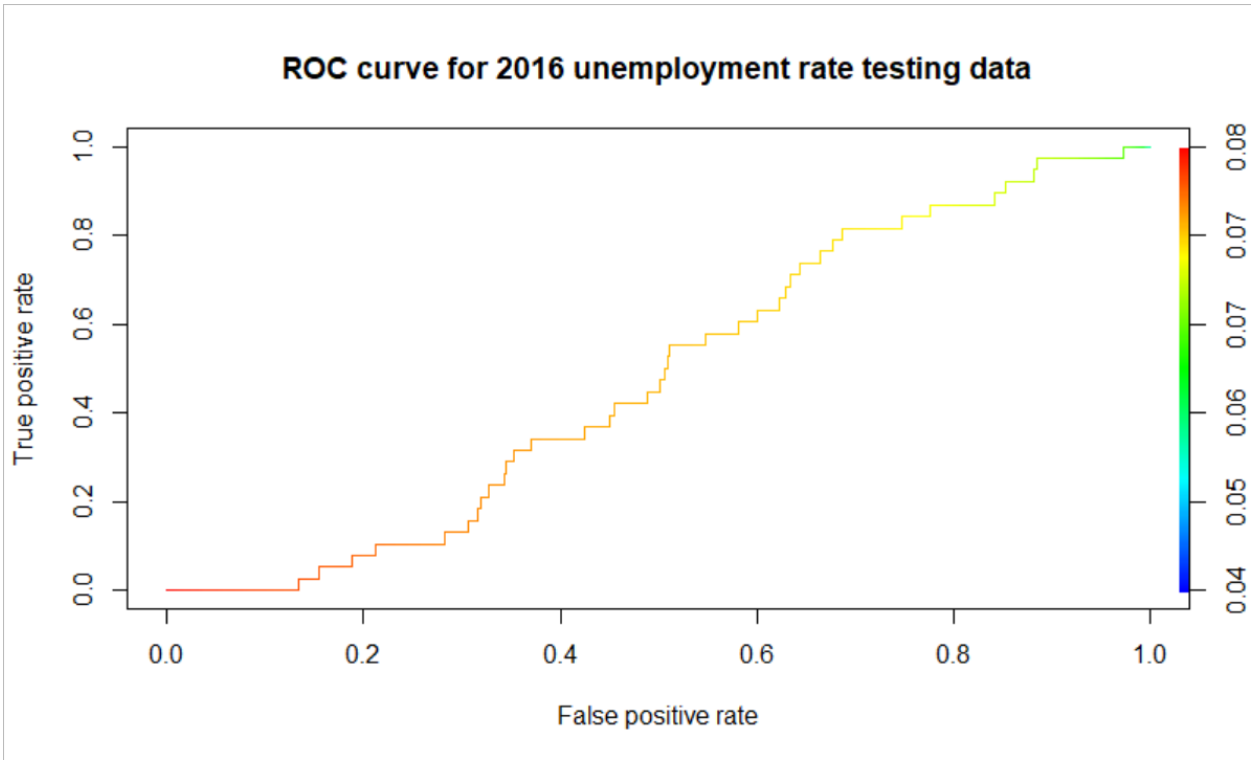
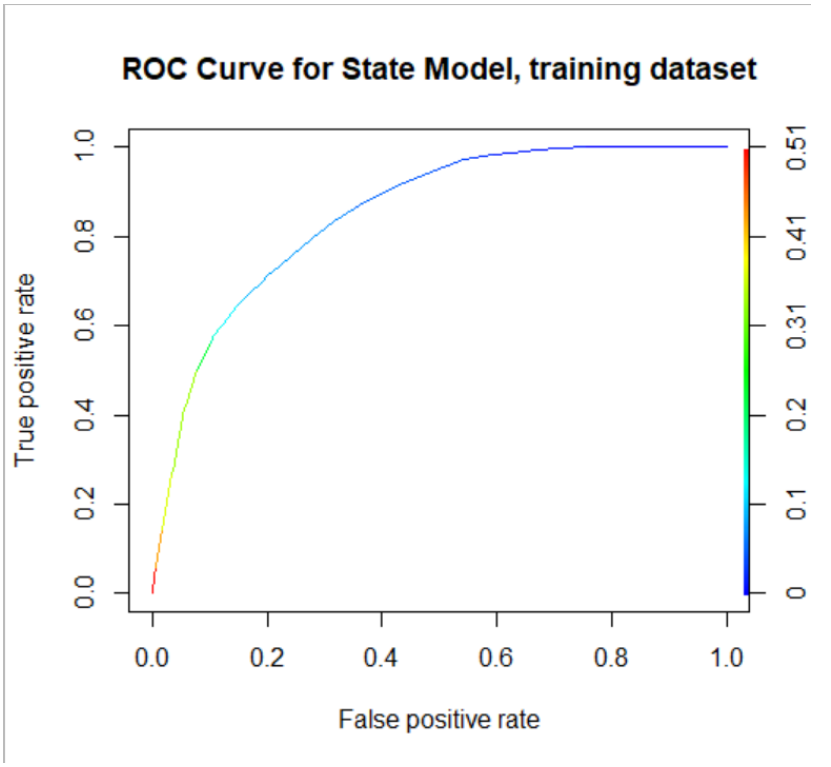


Figure 7: ROC Curves for State Model



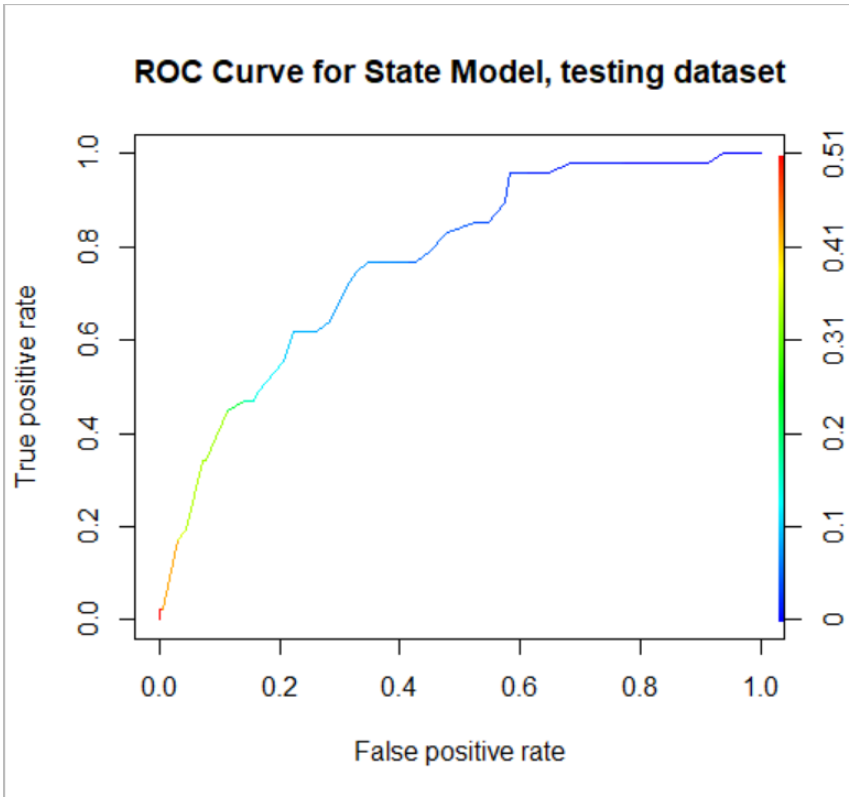
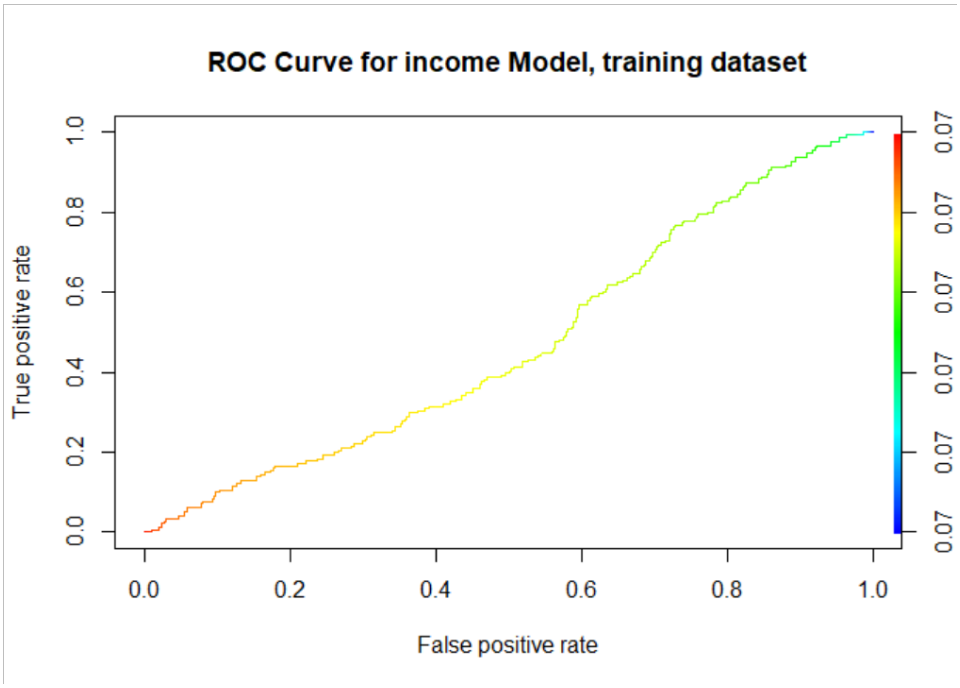


Figure 8: ROC Curves for Income Model



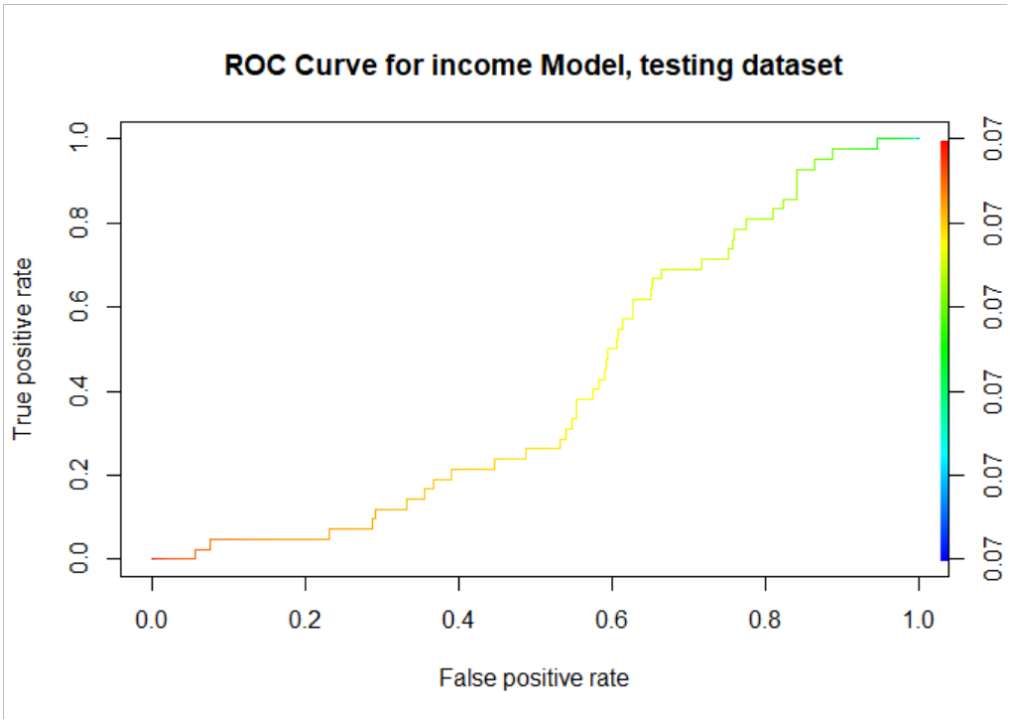
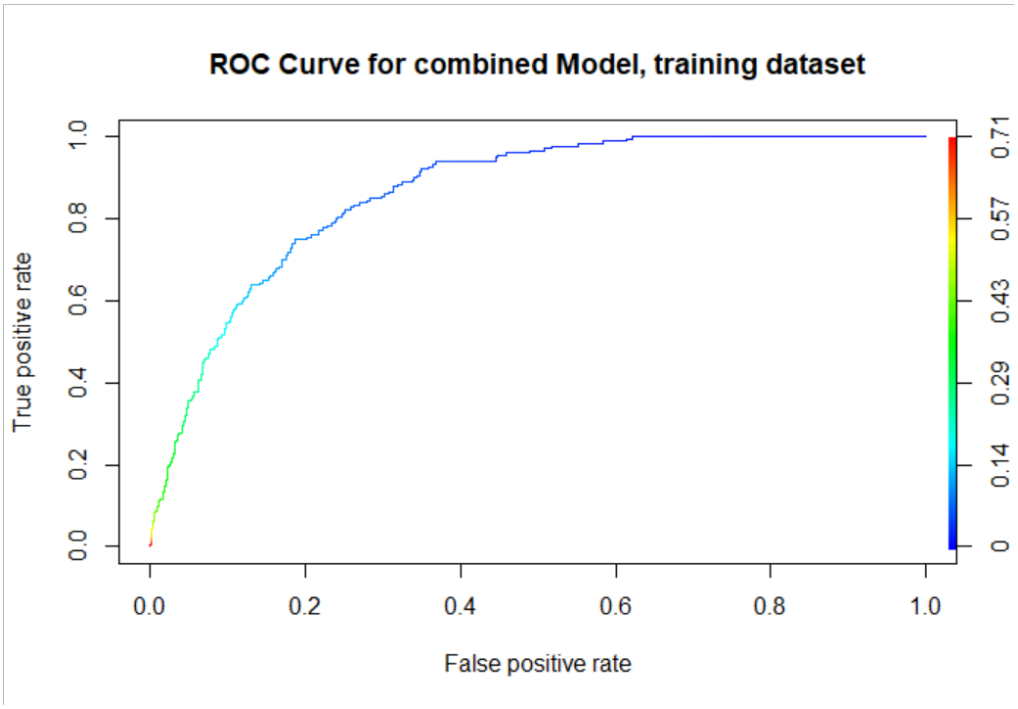
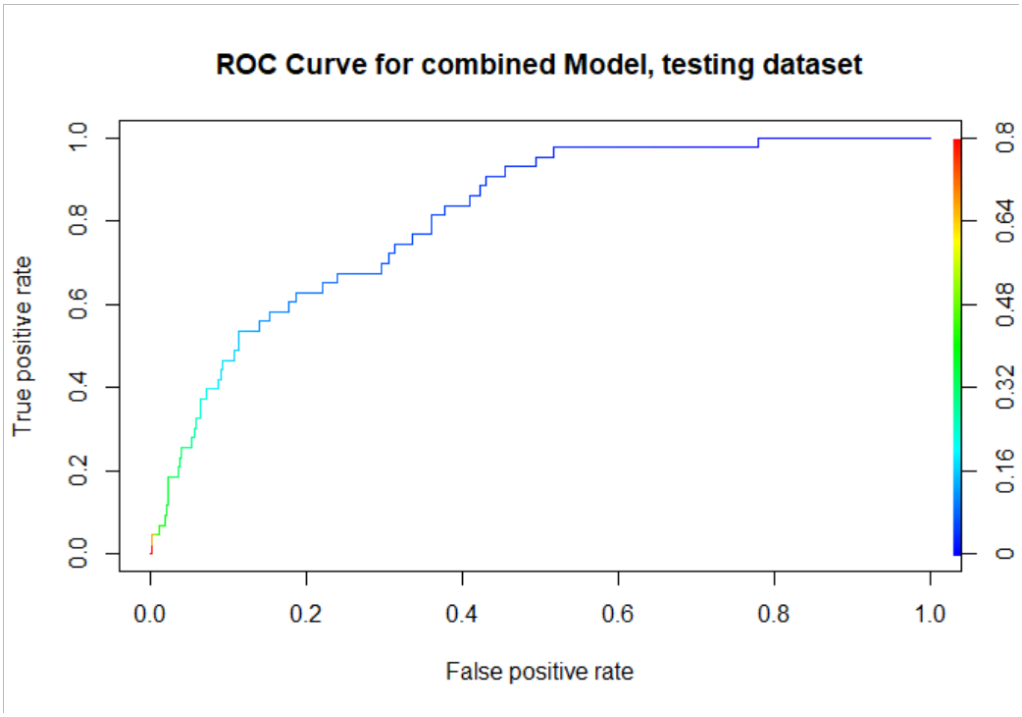


Figure 9: ROC Curves & Confusion Tables for Combined Model

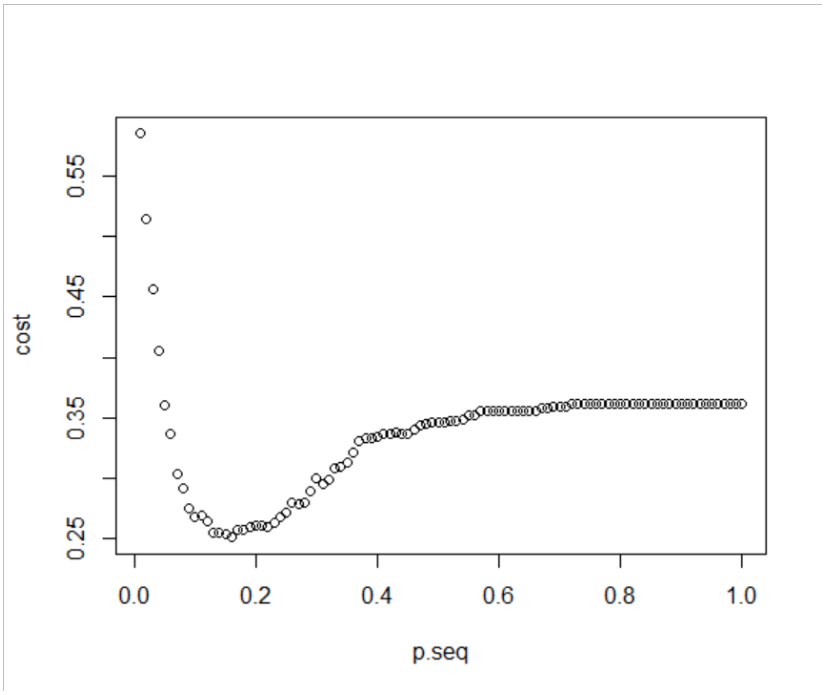




Naïve choice cut-off probability: cutoff of .07

	Predicted	
True	0	1
0	1717	594
1	32	148

Grid search cut-off probability: cutoff of .16



True	Predicted 0	1
0	2061	250
1	75	105