Final Project Report - Unemployment Rate

Ashley Clarke

Due: December 19, 2021 at 11:59pm

Contents

Executive Summary

Problem

The Federal Reserve of Economic Data (FRED) releases two reports monthly: > Employment Situation > Gross Domestic Product

While these reports are intended to give an overview of the current economic situation, there are thousands of variables in each report. This report intends to identify which variables are predictive of the U.S. unemployment rate. In addition to the variables found in these reports, inflation and the federal funds rate were added as predictors.

Instead of looking at classic predicitors of unemployment such as whether the U.S. is in a recession or stock market prices, this report examines factors such as number of employees by industry, hours worked by industry, levels of federal aid, farm output, net lending, real consumption, etc.

This report predicts unemployment rate based on five different methods: least squares regression, Ridge regression, Lasso regression, Random Forest regresssion, and Boosting.

Federal Reserve Economic Data: https://fred.stlouisfed.org/release?rid=50

The Burea of Economic Analysis: https://www.bea.gov/data/gdp/gross-domestic-product

Data

I first downloaded U.S. unemployment data, which was first available in June of 1954.

The Federal Reserve of Economic Data (FRED) releases two reports monthly:

Employment Situation

Gross Domestic Product

Brief description of data. January 1960-November 2021, x amount of variables pulled from Fred, added inflation, and federal funds rate. Caution that date points might not be independent, but excluded date in analysis

Analysis

Overview of analysis conducted: pulled all data, remoed highly correlated variables, EDA, regression, tree based, model evaluation

Conclusions

Main conclusions of the analysis, including takeaways for stakeholders Which variables have most importance? Anything surprising?

Introduction

Background information

Coronavirus disease (COVID-19) has had a devastating global impact, with a cumulative total of 149,987,772 confirmed cases and 3,157,594 deaths worldwide as of April 28, 2021.[1] About a fifth of these cases have been in the United States, with a recent count of 32,551,440 cases and 582,668 deaths.[2] With these staggering numbers still increasing despite recent large-scale vaccine rollouts, it is of vital importance to utilize various data sources to understand both the progression of COVID-19 thus far as well as the highest risk factors for contracting COVID-19. Furthermore, a thorough analysis of COVID-19 rates and predictive factors may

[1] Coronavirus Cases: Worldometer. (n.d.). https://www.worldometers.info/coronavirus/.
[2] Ibid.

help inform strategies to improve public health policies that could mitigate the negative impact of a future pandemic, which many scientists say is not a matter of if but of when.[3]

Past research has shown that infectious diseases are influenced by a variety of factors. Obesity, for instance, is associated with a higher likelihood of contracting influenza A, and seasonal temperature changes have shown to be predictive of the 2003 severe acute respiratory syndrome (SARS).[4] The CDC is currently in the process of identifying potential risk factors for severe COVID-19 illness,[5] and some that have already been identified include heart disease, diabetes, and pregnancy.[6] Yet despite these efforts, there is still much to be learned. Specifically, there is still insufficient research to explain the differences in COVID-19 susceptibility and mortality that exist not just on the individual level but also on broader population levels.

Analysis goals

Given our knowledge of the capacity for a variety of factors to influence infectious disease spread as well as the fact that different counties in the US have differing levels of baseline health factors, we sought to investigate how rates of COVID-19 cases and deaths across the US are affected by various measures of community health. Specifically, we were interested in which kinds of factors (e.g., clinical, behavioral, health)—and which specific variables—are most predictive of deaths per cases (also known as case fatality rate).

Significance

We hope that our analysis will contribute to the growing body of research on COVID-19 risk factors by expanding our understanding of COVID-19 and supporting efforts to mitigate the risk of future pandemics. Our results also shed light on the importance of analyzing social determinants of health in efforts to improve health outcomes.

Data

Data sources

The data used in this report stems from two reports:

Employment Situation:: Federal Reserve Economic Data: https://fred.stlouisfed.org/release?rid=50

Gross Domestic Product: The Bureau of Economic Analysis: https://www.bea.gov/data/gdp/gross-domestic-product

The data was collected using fredr, which provides a complete set of R bindings to the Federal Reserve of Economic Data (FRED) RESTful API, provided by the Federal Reserve Bank of St. Louis. The fredr package allowed me to search for and fetch time series observations as well as associated metadata within the FRED database. Since FRED organizes their data using variable ids, I downloaded time series observations from all variable ids in the Employment Situation and Gross Domestic Products reports, which represents over 2000 variables, from June 1954-November 2021. Additonally, I downloaded the U.S. unemployment rate (response variable), inflation rate, and federal funds rate. Before cleaning, the data set consisted of 804 observations and 2004 variables.

[3]Robbins, J. (2021, January 4). Heading Off the Next Pandemic. Kaiser Health News. https://khn.org/news/infectious-disease-scientists-preventing-next-pandemic/.

[4]Tian, T., Zhang, J., Hu, L., Jiang, Y., Duan, C., Li, Z., ... & Zhang, H. (2021). Risk factors associated with mortality of COVID-19 in 3125 counties of the United States. Infectious diseases of poverty, 10(1), 1-8.

[5]Centers for Disease Control and Prevention. (n.d.). Assessing Risk Factors for Severe COVID-19 Illness. Centers for Disease Control and Prevention. https://www.cdc.gov/coronavirus/2019-ncov/covid-data/investigations-discovery/assessing-risk-factors.html.

[6]Centers for Disease Control and Prevention. (n.d.). Certain Medical Conditions and Risk for Severe COVID-19 Illness. Centers for Disease Control and Prevention. https://www.cdc.gov/coronavirus/2019-ncov/need-extra-precautions/people-with-medical-conditions.html.

Due to the size of the data set, the data set takes around 5-10 minutes to download from FRED. Also, since the fredr package has a limit of 120 requests / minute, it might take longer than expected for the data to load.

Data cleaning

There were three critical isues that had to be resolved during the data clearning phase

(1) Features have not been reported all years

(2) Features are reported in different time increments: monthly, quarterly, and yearly

(3) Many of the features are highly correlated with each other

1. Timeframe Issues: Not All Features Available Since 1954

While unemployment rate has been reported monthly since June 1954, many other features have not been reported for the entire timeframe. Additionally, certain metrics have not yet been reported for 2021. Therefore, only observations from January 1960 to Decemember 2020 were kept in the data set. If a feature has not been reported since 1960, it was dropped.

2. Time Between Reports: Not All Features Reported Monthly

While both the Employment Situation and Gross Domestic Product reports are released monthly, not every feature is updated monthly. Many features are reported either quarterly or yearly. This issue was identified by examining the number of observations per feature. I noticed that many features only had 61 complete observations and 244 observations.

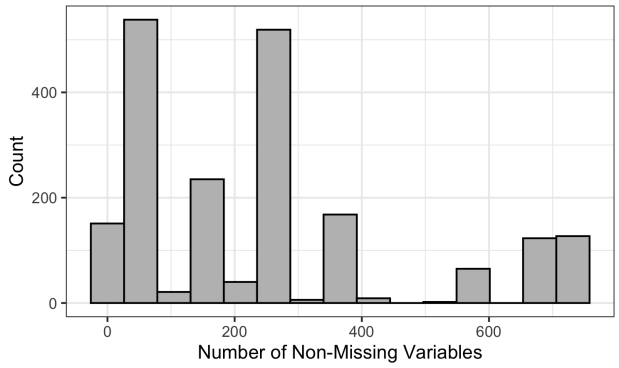

Figure 1: Histogram of the number of observations per feature that are not NA/ missing. Can see peaks at 61 (number of years) and 244 (number of quartes)

The makes sense because the time frame corresponds to 61 years and 244 quarters (3 months each). To impute the missing values for yearly data, I set every month in the yearly equal to the yearly metric. For missing quarterly data, I set the 2 next two months equal to the quarterly metric. While I recognize that this is not a perfect way to impute these features, I believe it is better than dropping columns or rows. After imputing missing values, I dropped all columns with NA values, which left me with 831 features.

3. Repeat features: Some features represent the same metric with minor adjustments

FRED makes many minor adjustments to metrics and reports them as seperate features. For instance, both seasonally adjusted and non seasonally adjusted numbers are reported monthly for most metrics. To eliminate double counting variables, if a variable and another variable have a higher than 0.9 correlation with each other, I only kept one of the variables. After removing variables that are highly correlated, there are 160 remaining features.

Standard Deviation Equal to Zero

I calculated the standard deviations of all variables. If a variable had a standard deviation of 0, I removed it because it is a meaningless feature

Cleaned data set

Observations: The cleaned data set has a total of 732 observations, corresponding to each of the 732 months between January 1960 and December 2020.

Response Variable: Unemployment Rate (UNRATE) is the response variable. The unemployment rate represents the number of unemployed as a percentage of the labor force. Labor force data are restricted to people 16 years of age and older, who currently reside in 1 of the 50 states or the District of Columbia, who do not reside in institutions (e.g., penal and mental facilities, homes for the aged), and who are not on active duty in the Armed Forces. The response variable is reported monthly and is seasonally adjusted.

Documentation: https://fred.stlouisfed.org/series/UNRATE

Explanatory Variable - Reports The cleaned data set includes 160 features, and documentation of each feature can be found at the end of this document (attach link).

Explanatory Variables - Additioanl

Inflation:

Reasoning: According to economic theory, as unemployment rates fall, the rate of inflation rises. This has been formalized according to what is known as "the Phillips Curve."

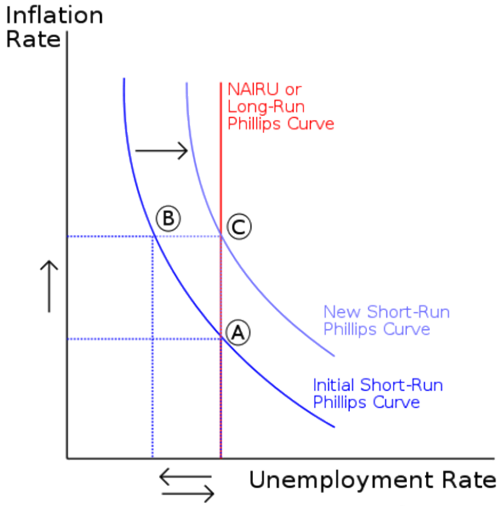

Figure 2: Phillips Curve

Inflation (FPCPITOTLZGUSA) as measured by the consumer price index reflects the annual percentage change in the cost to the average consumer of acquiring a basket of goods and services that may be fixed or changed at specified intervals, such as yearly. The Laspeyres formula is generally used. This metric is not seasonally adjusted and is recorded annually.

Documentation: https://fred.stlouisfed.org/series/FPCPITOTLZGUSA

Federal Funds Rate:

Reasoning for adding: It is thought that the unemployment rate and federals funds rate have a negative contemporaneous relationship. I exepect that when the unemployment rate is at its highest, the federal funds rate will be at its lowest. This likely happens because there is a lower federal funds rate in a weak economy.

Source: https://minds.wisconsin.edu/bitstream/handle/1793/77330/Federal%20Funds%20Rate.pdf?sequence=1&isAllowed=y

The federal funds rate is the interest rate at which depository institutions trade federal funds (balances held at Federal Reserve Banks) with each other overnight. When a depository institution has surplus balances in its reserve account, it lends to other banks in need of larger balances. In simpler terms, a bank with excess cash, which is often referred to as liquidity, will lend to another bank that needs to quickly raise liquidity.

Documentation: https://fred.stlouisfed.org/series/FEDFUNDS

Data allocation

I used an 80-20 split, such that the training dataset consists of 80% of observations and the test data set consists of 20% of observations. The same train-test split was used for each class of methods. Additionally, colinearlity analysis and data exploration used solely the train data set. Thus, there are 585 test observations and 147 train observations.

Data exporation

0.0.1 Response Variable

First, I looked at the response variable's distribution. As seen in the histogram of unemployment rate variable (Figure 3), the data appears to be right-skewed, with some months having a unemployment rate that exceeds 10%. The median unemmployment rate is 5.6%. Next, I looked at which years have the extreme unemployment rates and determined that those months corresponded to recessionary periods.

The sorted data (Figure 4) shows that aggregated across each year, the highest unemployment rates were 1975-1976, 1981-1984, and 2009-2012, which are all recession years. I am curious what underlying variables, besides the fact that the U.S. is in a recession, drive the changes in unemployment rate.

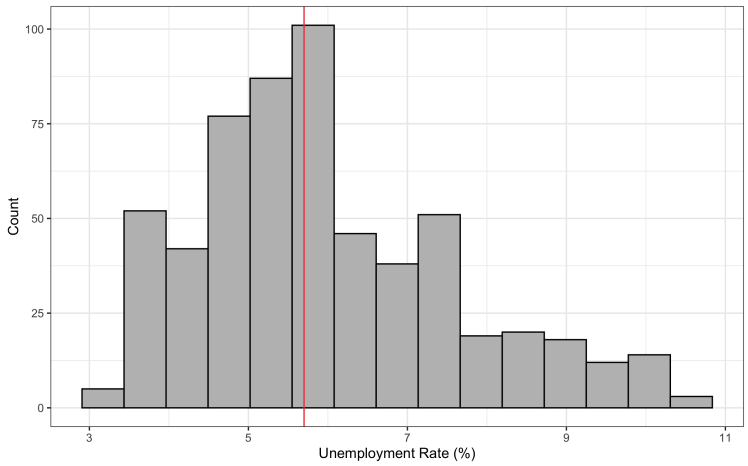

Figure 3: Histogram of Unemployment Rate

Next, I looked at the mean unemployment rate from 1960-2020 and from 2010-2020 in a chart ((Figure 5) to see if recent unemployment rates are substantially higher or lower than the overall mean. I found that recent unemployment rates are relatively consistent with unemployment rates across the entire time frame.

Finally, I plotted unemployment rate by date for all years (Figure 6) to visualize how the unemployment rate has changed over time.

0.0.2 Inflation

I would expect for inflation to rise as unemployment rates falls due to the Phillips Curve (Figure 2). However, as the unemployment rate falls, inflation does not appear to rise. There is not apparent relationship between these two variables.

year	UNRATE
1983	9.6
1982	9.5
2010	9.5
2009	9.4
2011	8.9
1975	8.5
2012	8.1
1976	7.7
1981	7.6
1984	7.5

Figure 4: Years with the Highest Average Monthly Unemployment

Time	Unemployment Rate
1960-2020	5.98
2010-2020	6.17

Figure 5: Mean Unemployment Rate by Time Period

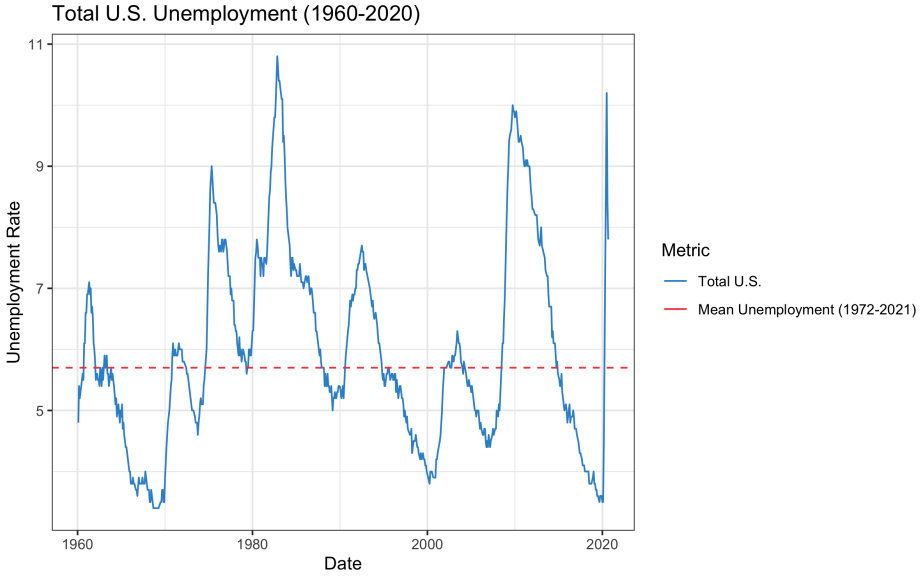

Figure 6: Total U.S. Unemployment by Date

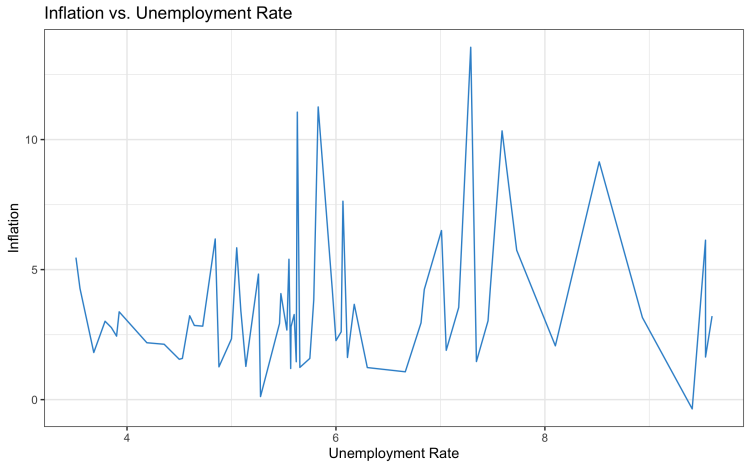

Figure 7: LATER

0.0.3 Federal Funds

Next, I checked to see if the federal funds rate and unemployment rate had a negative relationship in the last 20 years. I choose to only look at the last 20 years because line plots with all 585 observations can be hard to interpret. (Figure 8) implies a direct, negative linear relationship between the federal funds rate and the unemployment rate. Thus, (Figure 8) implies when the unemployment rate is high, the federal funds rate will be low and vice versa.

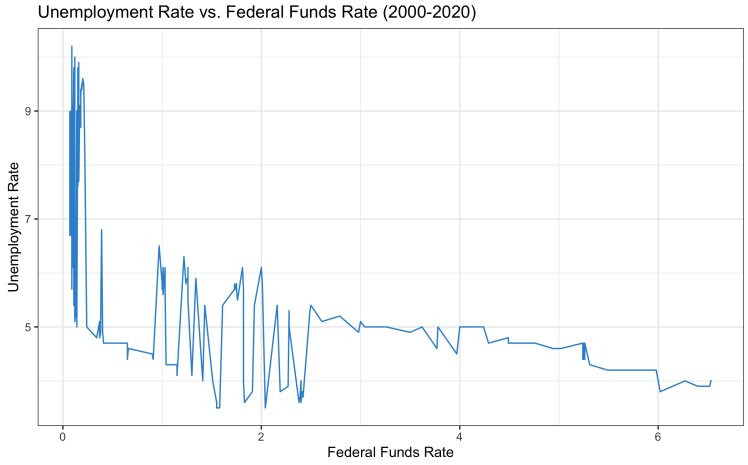

Figure 8: Unemploy by Federal Funds Rate

0.0.4 Feature Variation

As mentioned above, many of the orginal features had high covariation since FRED reports adjusted and unadjusted metrics. Now, no features in the data set have above a covaritation of 0.9 with each other.

While there are far too many features to create a correlation plot with every feature, I randomly subsampled 30 features to create a correlation matrix. (Figure 8) shows that majority of the features are not highly correlated, but a few are. This is not a cause for concern because regression techniques (e.g. lasso, ridge) will adjust for this multicolinearlity.

Additionally, in (Figure 10), I made a correlation plot, which shows the correlation plot between unemployment rate and all variables. The plot does not indicate any variables that are strongly positively / negatively

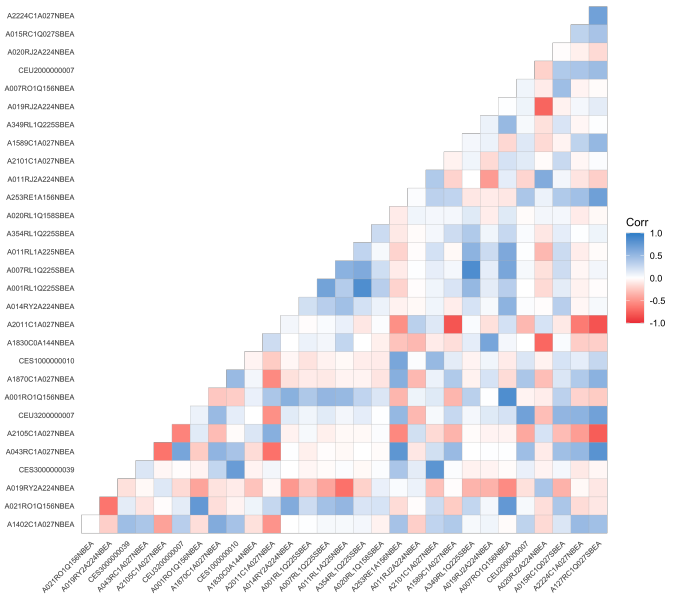

Figure 9: LATER

Variable Name	Correlation
A024RL1A225NBEA	-0.57
A024RL1Q225SBEA	-0.53
CES1000000006	0.49
CES1000000010	0.48
AWHMAN	-0.41

correlated with the plot. (Table **??**) indicates that Real Consumption of Fixed Capital: Private Annual (A024RL1A225NBEA) and Real Consumption of Fixed Capital: Private Quarterly (A024RL1Q225SBEA) are negatively correlation with unemployment rate. meaning real consumption of fixed capital decreases, unemployment rate increases. Also, (Table **??**)

as Production and Nonsupervisory Employees, Mining and Logging (CES1000000006) and Women Employees, Mining and Logging (CES1000000010) increases, so does employment rate, which potential suggests people turn to these jobs when they cannot find jobs elsewhere. Finally, verage Weekly Hours of Production and Nonsupervisory Employees, Manufacturing (AWHMAN) is negatively correlated with unemployment rate, suggesting as manufacturing employees work more hours, unemployment rate declines.

0.0.5 Number of Employees by Industry

Next, I wanted to see what industries affect or do not affect the overall unemployment rate. Specifically, I looked at the logging, shipping/boating, information, and federal. Due to recent increase in number of inoformation jobs, I also plotted information vs. unemployment for 2000-2020. (Figure 11) suggests that as number of employees decline in any industry, the unemployment rate rises. However, the number of federal employees stays relatively constant despite unemployment rate.

I will also add that the number of employees per industry is potentially a bad predictor because population size is constantly increasing and demand for workers across industries consistently shifts.

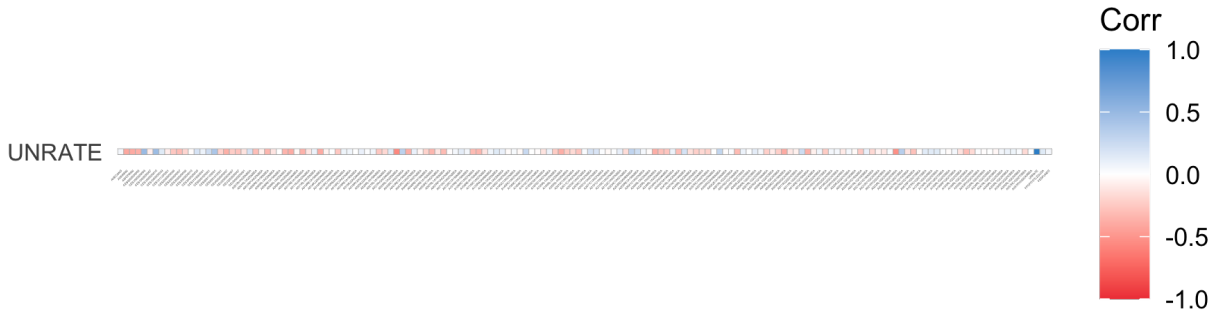

Figure 10: LATER

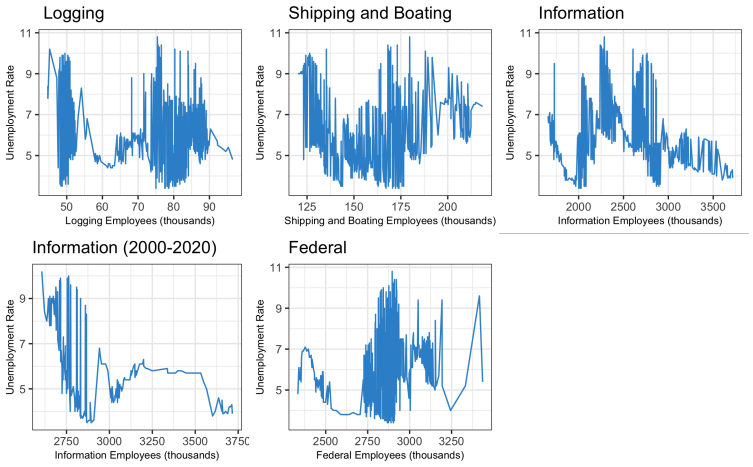

Figure 11: Unemployment Rate by Number of Employees by Industry

0.0.6 Average Weekly Hours of Production by Industry

I would predict that as the average number of hours of production increases, unemployment rate decreases. I looked at the manufacturing, mining/logging, and construction industries. (Figure 12) shows that my prediction holds true. Out of selected variables, the manufacturing industry has the clearest correlation between an increase in average weekly hours of production and a decrease in unemployment rate.

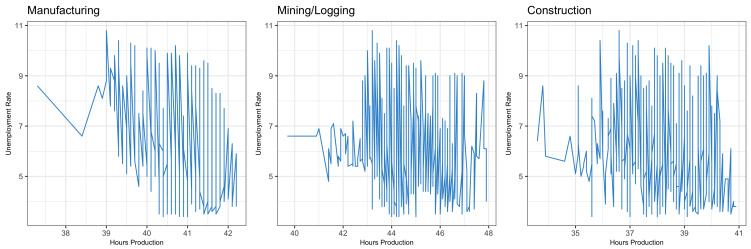

Figure 12: Unemployment Rate by Average Weekly Hours of Production by Industry

0.0.7 Factors from GDP report

I also examined features that were present in the GDP report. I choose to plot four variables: social benefits to persons, net government saving, real consumption of fixed capital, and real disposable personal income in (Figure 13). Unemployment rate increases when government social benefits increase, which likely occurs because unemployment is a qualifying factor for many of these benefits. Also, as real consumption of fixed capital and real disposable personal income increases, unemployment rate generally falls. This makes sense because corporations are willing to invest more in fixed assets (e.g. buildings) when the economy is doing well. Additionally, disposable perosnal income increases when someone is employed and when the economy is generallhy doing well.

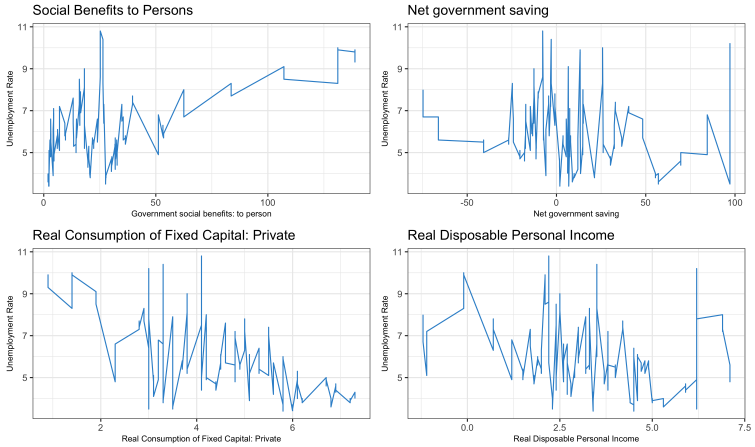

Figure 13: Unemployment Rate by Variables from GDP Report

0.0.8 Federal Aid

0.0.9 Farm Output

Reponse and women

Modeling

Modeling Class 1: Regression Methods

Linear Regression

Training

Tuning

Interpretation

Ridge Regression

Training

Tuning

Interpretation

Lasso Regression

Training

Tuning

Interpretation

Modeling Class 2: Tree-based Methods

Random Forest

Training

Tuning

Interpretation

Boosting

Training

Tuning

Interpretation

Conclusions

Method comparison

- Which method performed the best and why? Look to textbook and notes for this one

Takeaways

_ What things should they monitor in the report as signals for unemployment?

Limitations

- What were limitations of the data and/or the analysis, adn what obstacles did these limitations present? _ Different time frames
- Do not know if trends have changed over time/ dynamic of the work place probably has - Factors by gender
- Long load times, and took appraoch to getting rid of variables that are same, but might have eliminated crucial variables -> computer not large enough to handle data set

Follow-ups

What additional data collection or analysis would you recommend as a follow-up to your project? - Use black/ African American unemployment rate - Use data from other releases - See how predictors change across time - Issue with how unemployment rate is measured - Look at other countries data - What about unemployment by gender?

Follow-ups

Fredr https://cran.r-project.org/web/packages/fredr/vignettes/fredr.html

0.1 External PDF file is included below