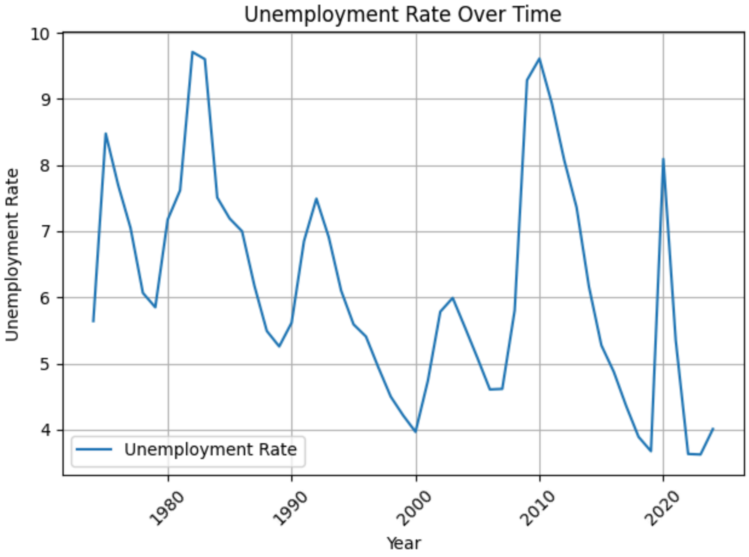
# Predicting the Average Annual Unemployment Rate

By Christina Salemme

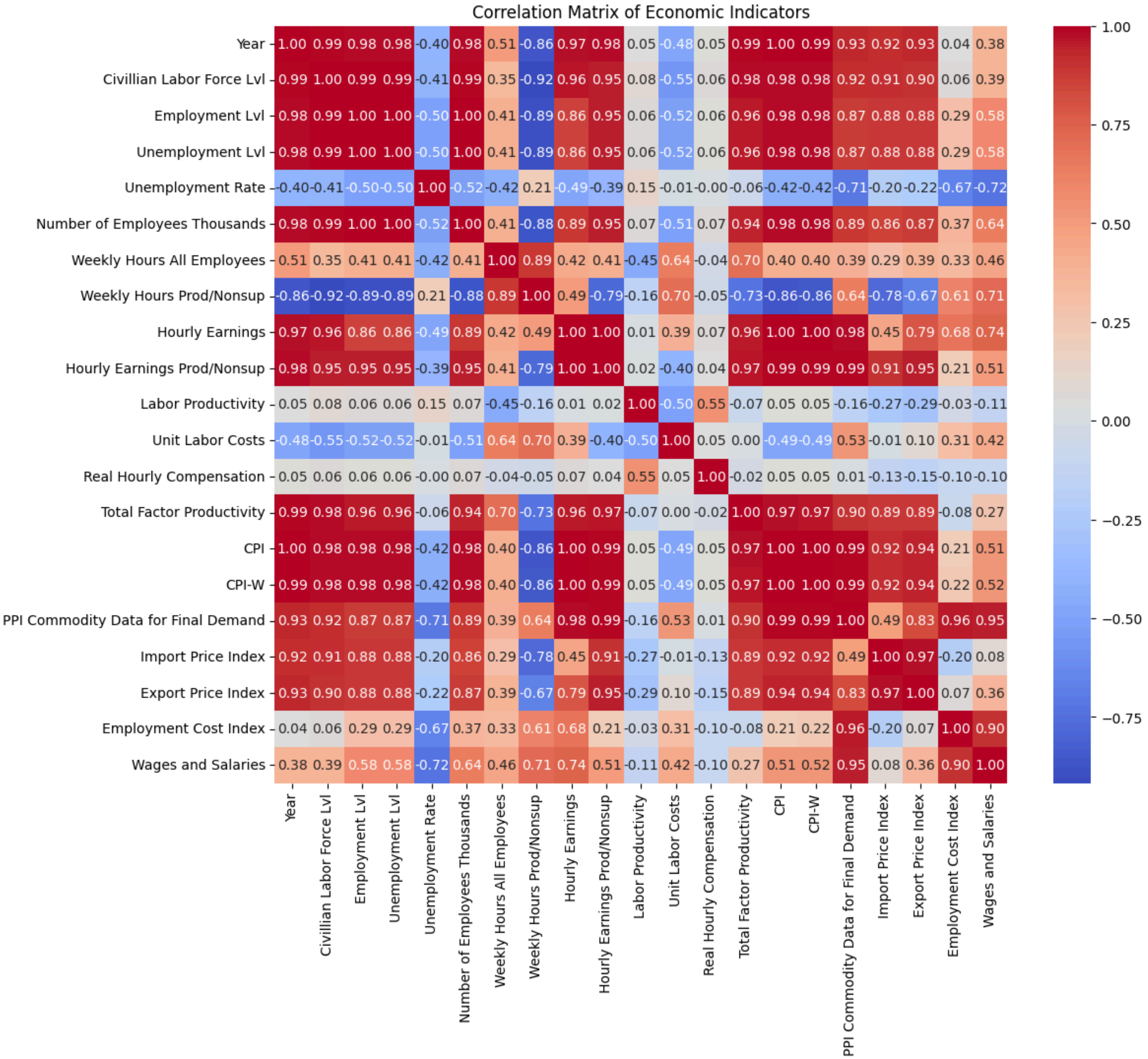
The George Washington University

## Background and Insight on the BLS Data (Q1)

The U.S. Bureau of Labor Statistics, otherwise known as the BLS is a government procured entity that collects, analyses, and publishes data on the U.S. economy. The BLS seeks to measure labor market activity, working conditions, price changes, and productivity, subsequently allowing users across a variety of fields to access data and use it at their discretion. For the sake of this report, we are utilizing BLS' top picks for the past fifty years (1974 - 2024). Upon downloading and compiling the data, I averaged all selected variables so they would be easily measured by year. Certain datatables were organized quarterly instead of monthly which caused some inherent discrepancies. Nonetheless, I aimed to visualize any and all patterns involving these economic determinants. To the right, I plotted  the average unemployment rate over the fifty year span, examining the endless waves of spikes and sharp declines brought on by the business cycle, often coupled with political disruptions, supply chain issues, and major global events. For instance, the major spike in average unemployment rate in 2009-10 was due to the 'Great Recession', which was primarily caused by the collapse of the housing bubble. This recession led to widespread job losses across various industries, ultimately leading to low consumer and investment spending (prices were also exceedingly high because of the recession at the time).  Below, the correlation matrix overlooks the relationships between the downloaded indicators.  Although extremely overwhelming to the untrained eye, the heat map matrix visualizes the important correlations between variables. For example, the Consumer Price Index, with the exception of food and energy, had



an extremely high correlation with both types of hourly earnings, showing a directly proportional relationship to one another.



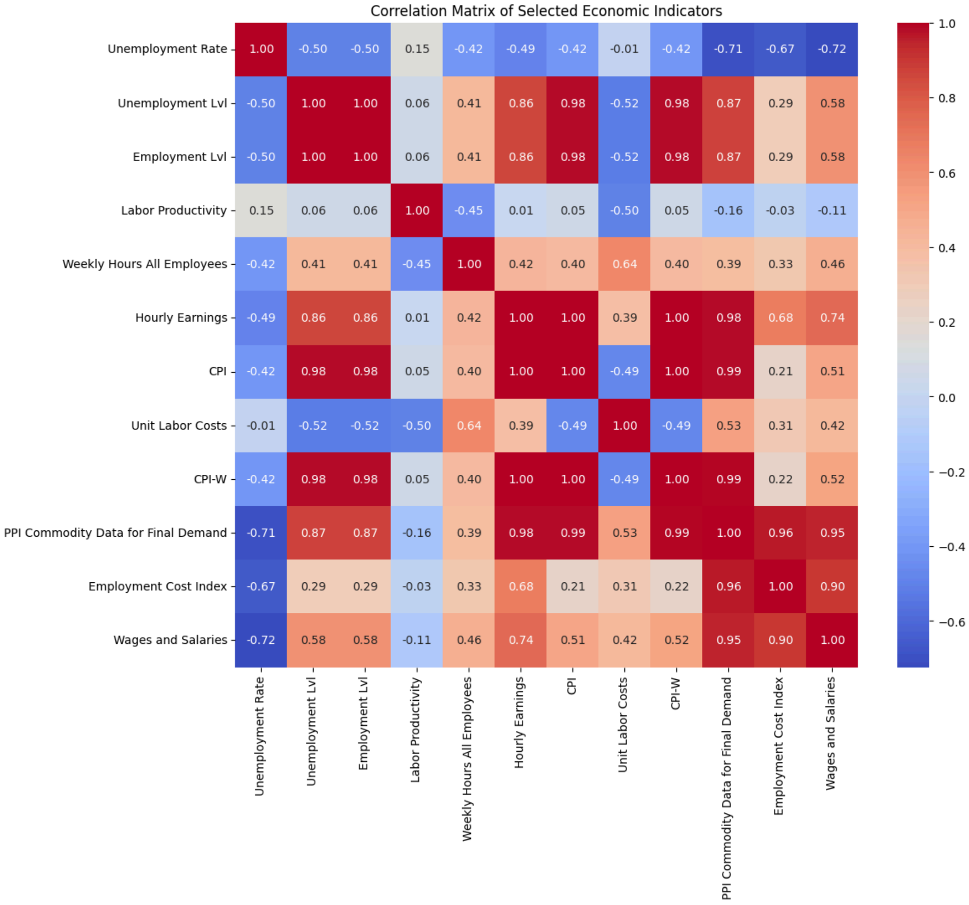
## The Path to Defining my Predictive Analytics Problem (Q2 and part of Q3)

While I explored various other trends with the BLS Top Picks variables, the average unemployment rate stood out to me. I spend a lot of time researching the labor market because of my background in economics. When it came time to choose my predictive analytics problem, I wanted to investigate the following: **Is it possible to successfully predict a baseline for the annual unemployment rate using economic indicators such as employment levels, labor productivity, average hourly earnings, and price indices as provided by historical BLS data?** Note that I am aiming to predict the unemployment rate annually not on a month-by-month or quarterly basis as the original datatables for specific indicators vary tremendously, which would limit the amount of variables available to compare.

Based on my question, the response variable is the average **Unemployment Rate** or the percentage of unemployed people in the labor force annually. order to determine the appropriate predictor variables I decided to do some external research regarding the most effective indicators and landed on this collective list: **Employment Level, Labor Productivity, Average Weekly Hours of Employees, Average Hourly Earnings, Consumer Price Index (CPI), Unit Labor Costs, CPI-W,PPI Commodity Data for Final Demand, Employment Cost Index, and Wages and Salaries.** I selected these indicators for predictors not only because my prior knowledge and research pointed to these out of everything would be instrumental in properly predicting the annual unemployment rate.

Before I begin, I would like to elaborate on some facts regarding my selected regressor predictor variables remember prior to running any models. Typically, employment levels and average weekly hours show a strong inverse correlation with the unemployment rate. This is largely due to labor demand, cost-cutting measures, and the business cycle impact. The fact that when the economy is strong and job demand is high, more people are employed, and existing employees may work longer hours. On the other hand, during a weak economy with high unemployment, employers are more likely to reduce hours worked by current employees before resorting to layoffs, leading to a decrease in both employment levels and average weekly hours. Additionally, CPI and labor productivity exhibit lagged effects, suggesting potential delayed impacts on unemployment. Unemployment traditionally follows growth with a delay, so naturally consumer price indices and labor productivity will fall behind. Lastly, sudden increases in unit labor costs tend to precede spikes in unemployment. As costs to create a product increases, businesses are more likely to decrease their supply of labor, thus leading to higher unemployment.

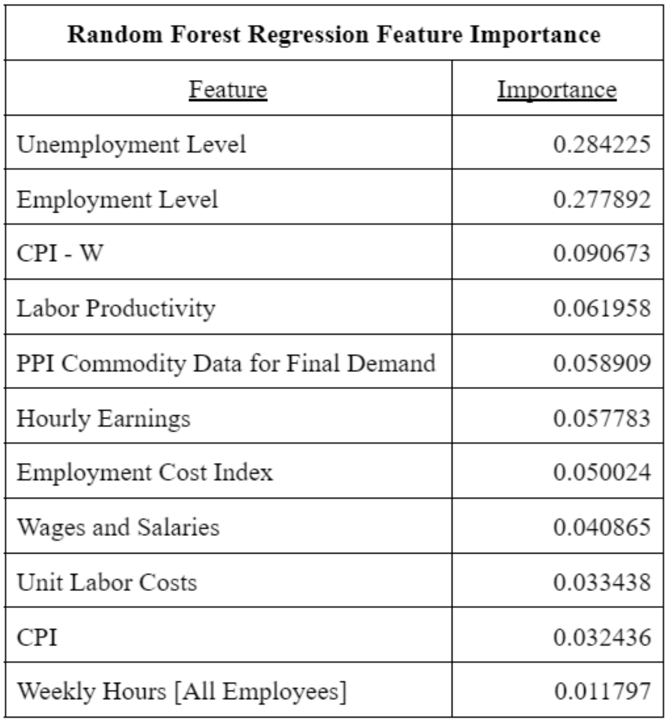
Alas, there are plenty of other factors that are outside the BLS data frame that are equally as instrumental in predicting the unemployment rate including GDP growth, inflation rate, labor force participation, job openings, initial jobless claims, consumer and business confidence, technological changes, government policies, industry trends, demographic factors like age distribution, and global economic conditions. Essentially, any variable that affects the overall health of the economy will also affect unemployment as it is a vital indicator of a country's economic status.



To note, the correlation matrix of all the variables in the BLS_Data_Frame shows that there is lower correlation between the unemployment rate and certain selected predictor variables. However, it does not mean that there is not a relationship between the two. Recall that the correlation matrix only shows a linear relationship between variables, meaning they can miss non-linear patterns, and importantly, correlation does not imply causation. Just as two variables are highly correlated, doesn't explain the whole story, potentially leading to incorrect interpretations of the data. This is why it is important to run various types of models to ensure accuracy and to examine the bigger picture. To help visualize the chosen predictive analytics problem, a secondary correlation plot shown above is a compressed version of the variables being explored.

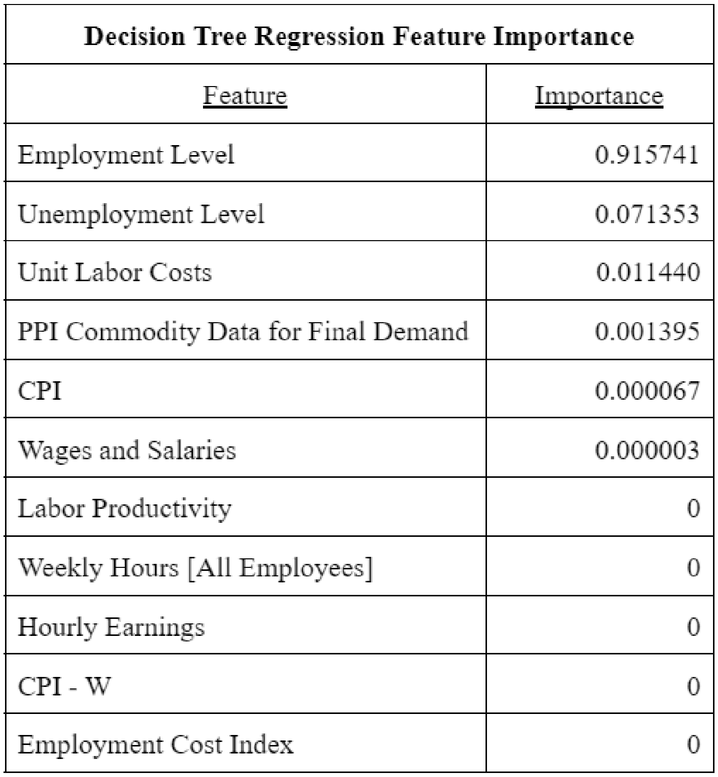
Before I can get into the specific models that I chose for my predictive analytics problem, I must detail how I decided to run a regression problem and select the predictor variables. As I previously stated, I completed external research to narrow my scope, but once selecting particular variables I wanted to see which had the utmost importance, leading me to running exploratory regression problems. At first, I ran a small OLS , a hypothesis testing model that statistically evaluates relationships between predictors and each response variable. This is used prior to training data in order to determine variables' statistical significance via p-values. A low p-value indicates a predictor is significantly associated with the response variable. Unfortunately, no p-values were under 95% confidence interval ($p < 0.05$), which would usually indicate that there is insufficient evidence to prove a relationship between the regressors and the independent variable. In this case, it also means that the relationship is non-linear, and the selected model does not capture the true relationship we are trying to predict.

With that being said I moved onto feature importance methods to select the most relevant features for predicting future unemployment rates.The "feature importance" refers to a score assigned to each feature in the dataset, indicating how much each one contributes to the model's prediction. This also helps determine feature selection and potentially model interpretation when examining mean squared error. For one, a random forest regression (RFR), a machine learning technique that builds multiple decision trees to model data trends and make predictions. Based on the feature importance, the RFR model highlights Unemployment Level (0.284771), Employment Level (importance: 0.278338), CPI-W (0.090673), Labor Productivity (0.062326), PPI Commodity Data for Final Demand (0.058909) as those which might serve as robust response variables to predicting annual unemployment rate. Next, I ran a decision tree regression, which selects the 'best' feature for splitting at each node based on information gain. The table below indicates that Employment Level (0.915741) serves as the most important variable with Unemployment Level, Labor Costs and PPI Commodity Data for Final Demand trailing behind. The rest of the variables resemble little to no importance at all, resembling only a few features actually affect the unemployment rate.Recall that a decision tree visually represents a series of potential decisions and their corresponding outcomes, essentially mapping out different paths to reach a final conclusion, allowing you to analyze and compare various choices based on specific conditions to identify



**Random Forest Regression Feature Importance**

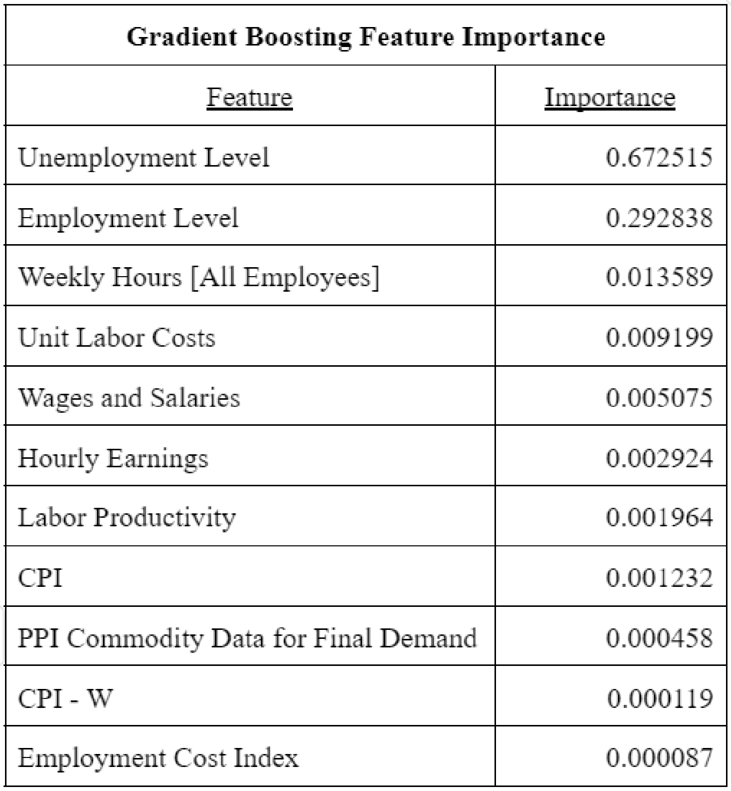
| Feature | Importance |
|---|---|
| Unemployment Level | 0.284225 |
| Employment Level | 0.277892 |
| CPI - W | 0.090673 |
| Labor Productivity | 0.061958 |
| PPI Commodity Data for Final Demand | 0.058909 |
| Hourly Earnings | 0.057783 |
| Employment Cost Index | 0.050024 |
| Wages and Salaries | 0.040865 |
| Unit Labor Costs | 0.033438 |
| CPI | 0.032436 |
| Weekly Hours [All Employees] | 0.011797 |



**Decision Tree Regression Feature Importance**

| Feature | Importance |
|---|---|
| Employment Level | 0.915741 |
| Unemployment Level | 0.071353 |
| Unit Labor Costs | 0.011440 |
| PPI Commodity Data for Final Demand | 0.001395 |
| CPI | 0.000067 |
| Wages and Salaries | 0.000003 |
| Labor Productivity | 0 |
| Weekly Hours [All Employees] | 0 |
| Hourly Earnings | 0 |
| CPI - W | 0 |
| Employment Cost Index | 0 |

the best course of action in a given situation; it essentially tells you which decisions to make based on different factors and what the likely results of those decisions will be [See ipynb for full diagram as formatting it is too small]. Then, I moved onto gradient boosting, a supervised machine learning algorithm combining multiple weak learners (like decision trees) to create a stronger predictive model for continuous target values (regression problems) by iteratively minimizing errors made by the previous models in the sequence. The feature importance for gradienting boosting highlights the unemployment and employment levels as well as weekly hours of all employees, unit labor costs, wages and salaries, and labor productivity, with the rest showing little purpose. Lastly, the support vector regression predicts the unemployment rate using a margin-based approach. SVR maps input features to a higher-dimensional space using kernels while also fitting a hyperplane within a margin of tolerance ϵ to predict continuous values. Even though SVR cannot run a feature importance, the top five utilizing the Recursive Feature Elimination was Labor Productivity, Hourly Earnings, CPI, CPI-W and Wages and Salaries, indicating quite a difference from the rest.



**Gradient Boosting Feature Importance**

| Feature | Importance |
|---|---:|
| Unemployment Level | 0.672515 |
| Employment Level | 0.292838 |
| Weekly Hours [All Employees] | 0.013589 |
| Unit Labor Costs | 0.009199 |
| Wages and Salaries | 0.005075 |
| Hourly Earnings | 0.002924 |
| Labor Productivity | 0.001964 |
| CPI | 0.001232 |
| PPI Commodity Data for Final Demand | 0.000458 |
| CPI - W | 0.000119 |
| Employment Cost Index | 0.000087 |

All in all, each exploratory regression model helped me narrow down my decision for selecting my features. Random Forest emphasizes Unemployment Level and Employment Level similarly to Gradient Boosting, suggesting that these two features are crucial for explaining the Unemployment Rate.

Random Forest emphasizes Unemployment Level and Employment Level similarly to Gradient Boosting, suggesting that these two features are crucial for explaining the Unemployment Rate.

Random Forest gives significant importance to CPI-W, which isn't as prominent in Gradient Boosting but is considered in the SVR with RFE. Labor Productivity and Hourly Earnings are also identified as important in Random Forest, though with lower importance than CPI-W and Employment Level. Alas, Unemployment Level and Employment Level are consistently among the most important features across all four models (Gradient Boosting, Random Forest, Decision Tree, and SVR), confirming that these variables are likely to be strong predictors for the Unemployment Rate. Features such as CPI-W, Labor Productivity, and Hourly Earnings are identified as significant by Random Forest and SVR, indicating that economic conditions and compensation factors could also provide valuable information for predicting unemployment. Ultimately, I decided to retrain my model using the selected key features and compare their performances without the excess:

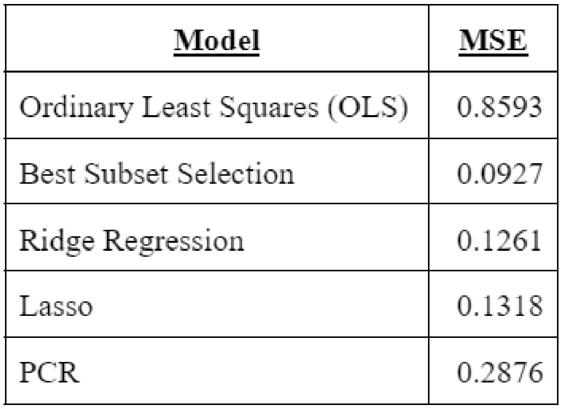
```
selected_features = ['Unemployment Lvl', 'Employment Lvl', 'Employment Cost Index', 'CPI','CPI-W', 'Wages and Salaries', 'Labor Productivity', 'Hourly Earnings']
```

With the second retraining, my results all pointed to these features significant in predicting unemployment rate. Thus, I could officially declare my predictive analytics problem: **Forecasting the average annual unemployment rate utilizing selected features, Unemployment Lvl, Employment Lvl, Employment Cost Index, CPI, CPI-W, Wages and Salaries, Labor Productivity, Hourly Earnings to predict the outcome, which could aid economists, especially those at the Federal Reserve to help construct policies that would impact the future state of the economy.**

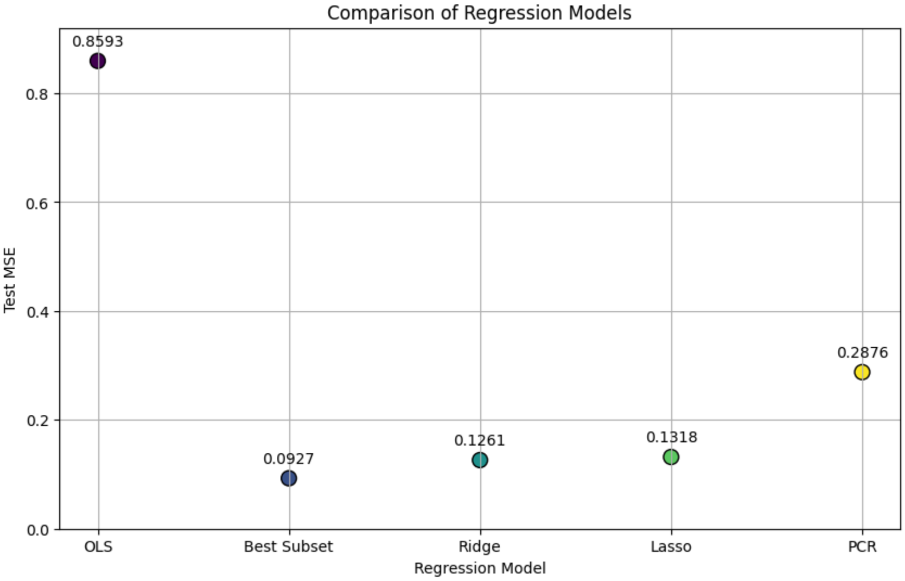
**Justification and Conclusion of Model Selection (Second Part of Q3 and Q4 )**

For predicting annual average unemployment rates, I evaluated several models, including OLS, Best Subset Selection, Ridge Model, Lasso Model and PCR. I selected the Best Subset Selection because it captures the optimal combination of predictor variables from a larger set by evaluating every possible subset of variables and choosing the one that best predicts the outcome based on mean squared error. The predictors were selected using feature importance scores from Random Forest, Gradient Boosting, Decision Trees and RFE from SBR, ensuring inclusion of only the most impactful variables.

To evaluate model performance, I measured MSE for regression tasks. This ensured models generalized well to unseen data which for predicting average annual unemployment rates, is vital. Below I am presenting the results graphically and tabular.



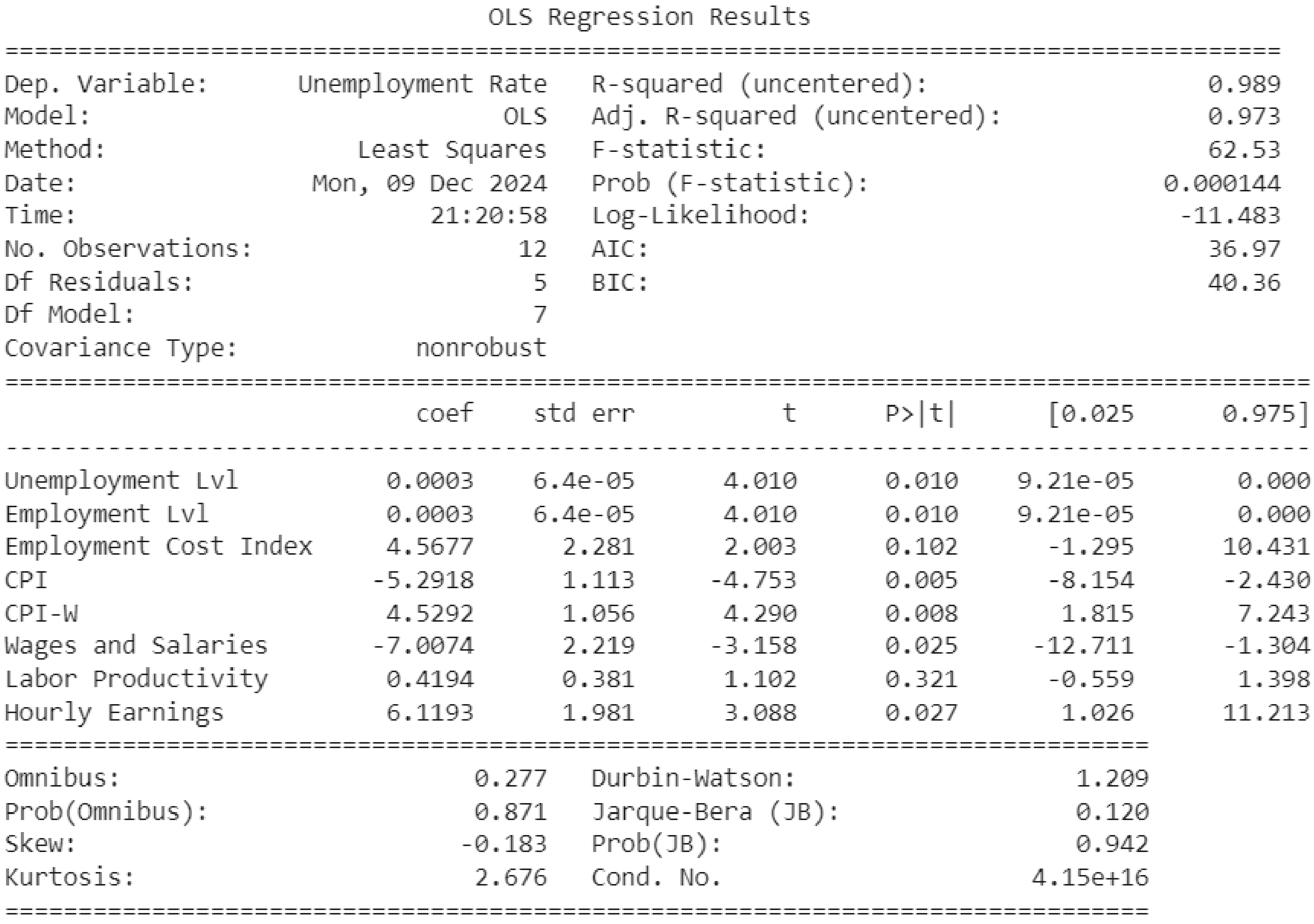
| Model | MSE |
|---|---|
| Ordinary Least Squares (OLS) | 0.8593 |
| Best Subset Selection | 0.0927 |
| Ridge Regression | 0.1261 |
| Lasso | 0.1318 |
| PCR | 0.2876 |



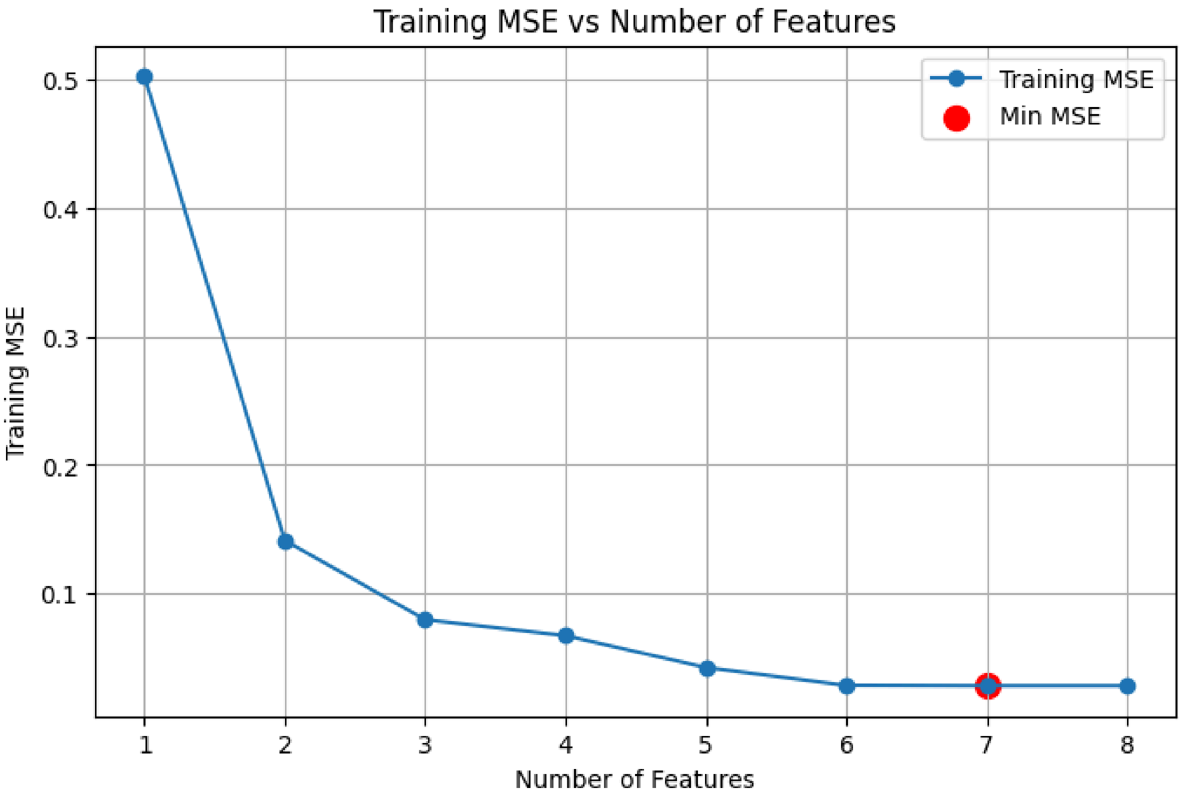
The final model selection of the Best Subset was based on both quantitative performance and interpretability of predictors. It performed the best, with the lowest MSE of 0.0927, indicating this model is the most reliable method for this dataset in terms of predicting the average annual unemployment rate. The Best Subset model is extremely flexible and handles multicollinearity well by allowing a subset of independent variables that are not highly correlated with each other (as we saw in the correlation matrix previously). This effectively reduces the impact of multicollinearity by eliminating redundant information from highly correlated variables. The Ridge Regression was not far behind with an MSE of 0.1261. Ridge regressions also handle multicollinearity better, thus if the MSE was a bit lower, it could have been selected. The Lasso followed with an MSE of 0.1318, less than 0.01 difference with Ridge Regression. The Lasso is capable of shrinking coefficients to zero, but implies that it might have removed less variables, useful but didn't outperform the previous two models. PCR was the second to last, with an MSE of 0.2876. PCR reduces dimensionality and then extracts useful information despite, but perhaps not enough in this case. Last and unsurprisingly, the OLS was the highest MSE; in all honesty, I used it as a baseline model because it generally helps rule out cases of multicollinearity and nonlinear relationships. Alas, the MSE was 0.8593 and compared to the rest of the models is too high to be used to predict annual unemployment rate. Enclosed in the page below are some of the results from these models to get a better picture as to how they all played out in the ipython notebook.

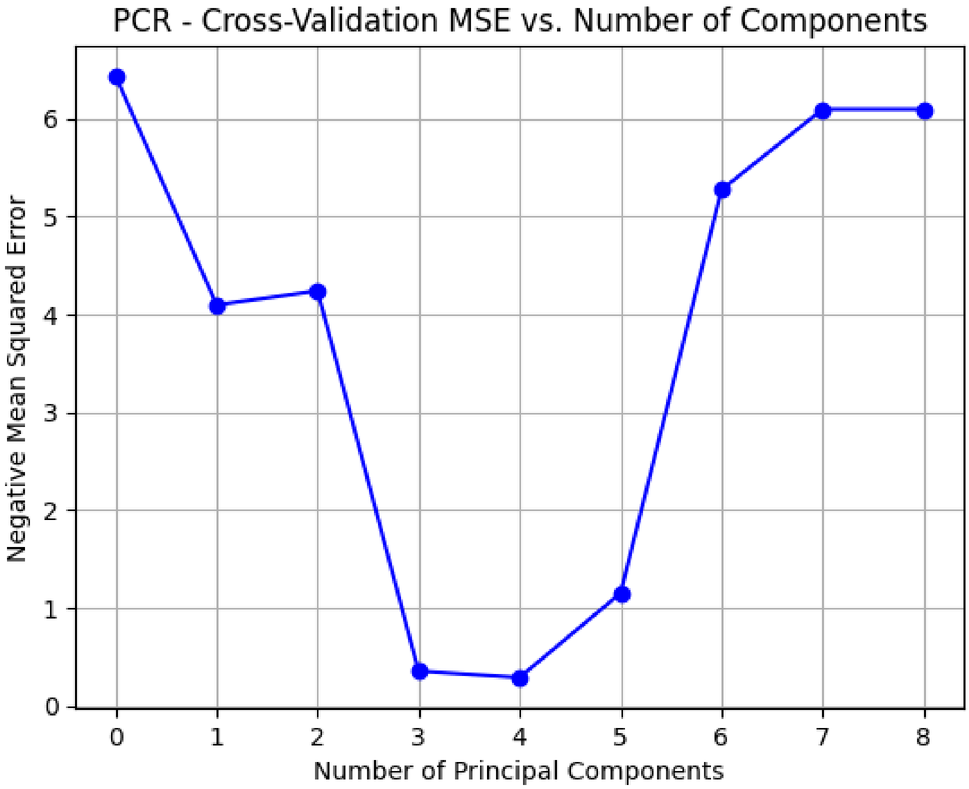
To evaluate what could be done in the future to improve these methods is to draw on classification models. I did explore some regressions as seen in the ipython notebook, but I found it a tad redundant for what exactly I was trying to predict. It is also important to note that certain features that I did include such as Employment Cost Index and Wages and Salaries are limited based on years, thus proving some data inconsistency. Alternatively,I imputed values for the earlier periods based on assumptions or trends, but this needs to be handled with precaution. Ultimately having true data that spans 1974-2024 would help capture the cyclical nature of the economy, especially the rise and fall of unemployment during recessions and recoveries.

Additionally, I would seek to find variables that are a bit more common for predicting unemployment such as inflation rate, labor force participation, job openings (JOLTs survey is incredibly informative). In conclusion, there always could be more explored when predicting unemployment rate and with some limitations, this model may have resulted in oversight.



```
                                 OLS Regression Results
=======================================================================================
Dep. Variable:      Unemployment Rate   R-squared (uncentered):                   0.989
Model:                            OLS   Adj. R-squared (uncentered):              0.973
Method:                 Least Squares   F-statistic:                              62.53
Date:                Mon, 09 Dec 2024   Prob (F-statistic):                    0.000144
Time:                        21:20:58   Log-Likelihood:                         -11.483
No. Observations:                  12   AIC:                                      36.97
Df Residuals:                       5   BIC:                                      40.36
Df Model:                           7
Covariance Type:            nonrobust
=========================================================================================
                            coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------
Unemployment Lvl          0.0003    6.4e-05      4.010      0.010    9.21e-05       0.000
Employment Lvl            0.0003    6.4e-05      4.010      0.010    9.21e-05       0.000
Employment Cost Index     4.5677      2.281      2.003      0.102      -1.295      10.431
CPI                      -5.2918      1.113     -4.753      0.005      -8.154      -2.430
CPI-W                     4.5292      1.056      4.290      0.008       1.815       7.243
Wages and Salaries       -7.0074      2.219     -3.158      0.025     -12.711      -1.304
Labor Productivity        0.4194      0.381      1.102      0.321      -0.559       1.398
Hourly Earnings           6.1193      1.981      3.088      0.027       1.026      11.213
==============================================================================
Omnibus:                        0.277   Durbin-Watson:                   1.209
Prob(Omnibus):                  0.871   Jarque-Bera (JB):                0.120
Skew:                          -0.183   Prob(JB):                        0.942
Kurtosis:                       2.676   Cond. No.                     4.15e+16
==============================================================================
```


Training MSE vs Number of Features
Training MSE
Min MSE
0.5
0.4
0.3
0.2
0.1
Training MSE
1
2
3
4
5
6
7
8
Number of Features


PCR - Cross-Validation MSE vs. Number of Components
6
5
4
3
2
1
0
Negative Mean Squared Error
0
1
2
3
4
5
6
7
8
Number of Principal Components