BIS-557 Computational Statistics Final Project

Instructor: Professor Michael Kane

Unemployment Rate Prediction by Economic Indexes through Deep Neural Network

Zichen Zhao (zichen.zhao@yale.edu)

1. Introduction

In March of 2020, COVID-19 pandemic started to outbreak in the U.S. All industries have been affected dramatically due to the high transmission rate and population movement in the spring. The downgrade of stock market and upgrade of unemployment rate are two sensible evidence to show the serious impact from COVID-19. In my point of view, unemployment rate and stock market should be highly correlated because they are all direct outcome of the country's economy change. Also, there should be a lag between the changes of stock market and unemployment rate (stock market changes before the movement of unemployment rate). This thought has been confirmed after reviewing *Quantifying Macroeconomic Expectations in Stock Markets using Google Trends (Johannes Bock, 2018)* and *The Stock Market's Reaction to Unemployment News: Why Bad News Is Usually Good for Stocks (H. Boyd et al, 2005)*. Thus, I decided to test whether unemployment rate can be predicted by economic indexes through neural network.

2. Data Description

There are six market indexes data and one unemployment data used for this project, and all of them were extracted from FRED economic data. The earliest data can be traced back to 05/16/1949 and the latest data is updated every day now. However, the earliest date all data have in common is 01/01/1986. Then, I took 01/01/1986 as my start point and 11/01/2020 as my end point to collect those data. The table below addresses seven variable names and their brief description.

Variable Name	Description
NASDAQ100	NASDAQ 100 Index
NASDAQCOM	NASDAQ Composite Index
WILL5000PR	Wilshire 5000 Price Index
NIKKEI225	Nikkei Stock Average, Nikkei 225
WLEMUINDXD	Equity Market-related Economic Uncertainty Index
VXOCLS	CBOE S&P 100 Volatility Index: VXO
UNRATE	Unemployment Rate

(Table.1)

The NASDAQ 100 Index contains 100 of the largest domestic and international non-financial securities listed on The NASDAQ Stock Market based on market capitalization.

The NASDAQ Composite Index describes market capitalization weighted index with more than 3000 common equities listed on the NASDAQ Stock Market. The types of securities in the index include American depositary receipts (ADRs), common stocks, real estate investment trusts (REITs), and tracking stocks.

The Wilshire 5000 Price Index describes the market-capitalization-weighted index of the market value of all the U.S. trading stocks.

The Nikkei Stock Average (Nikkei 225) is the major stock market index comprising of 225 highly liquid stocks of the Tokyo Stock Exchange. I think unemployment rate is mainly impacted by the U.S. economy, however, globalization plays some factor in the U.S. employment. The U.S. and Japan have strong and mutually advantageous economic relationship. That's why NIKKEI225 index is included for limited impacted from globalization.

The Equity Market-related Economic Uncertainty Index is an policy related index developed by Baker et al to address economic policy uncertainty, which is related to stock price volatility, investment and employment.

The CBOE S&P 100 Volatility Index estimates the expected 30-day volatility of the S&P 100 stock, originally provided by Chicago Board Options Exchange.

The Unemployment Rate represents the number of unemployed as a percentage of the labor force. The labor force is defined as people, living in 1 of the 50 states or the District of Columbia, 16 years of age and older.

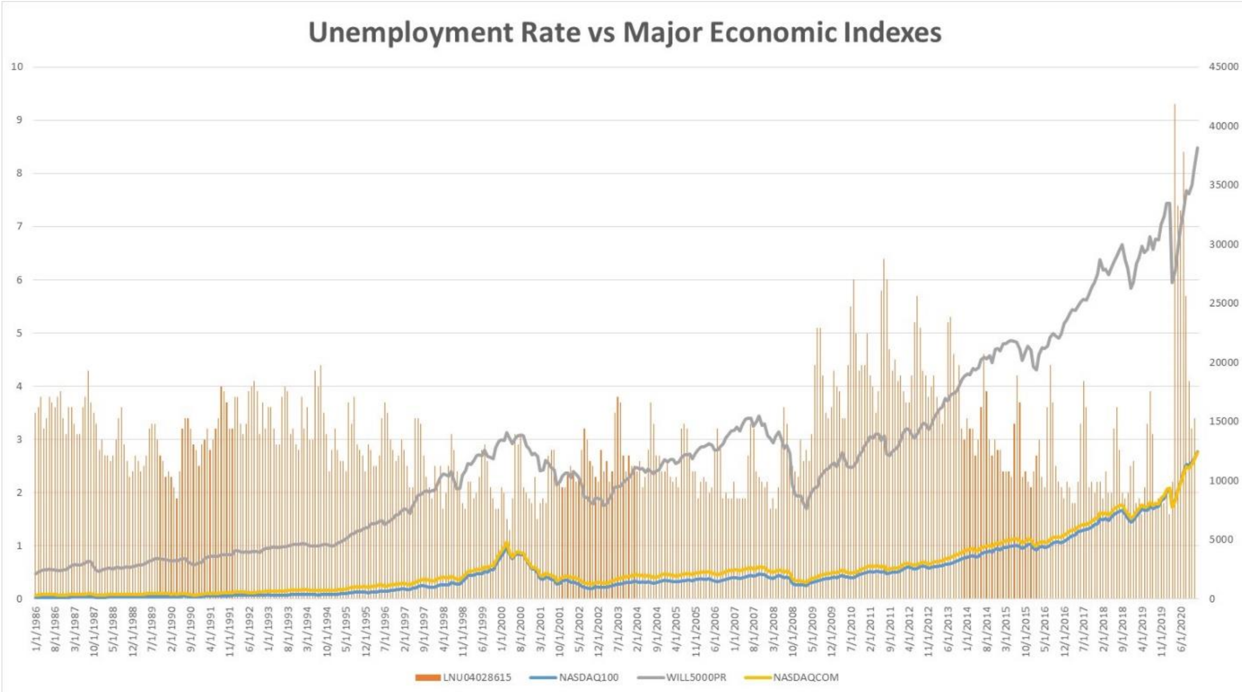
The economic indexes data listed above are providing an idea that what kinds of data should be considered for this project, and there are always other related data should be included in order to provide better accuracy.

3. Data Cleaning

Data cleaning is the focus of my study, although it is very tedious, I believe it is meaningful and important to construct statistical model effectively. The dataset is downloaded from source website and saved as csv file. I used Excel Pivot table and calculation functions to compile the daily indexes data into monthly average, and then eliminated a minimum amount of data to have a square shaped dataset, which means every observation has a dependent variable and valid independent variables.

3.1 Economic Indexes Modification

Original economic indexes data are represented in a daily format, yet the unemployment rate is reported monthly. Therefore, I calculated the monthly average of all economic indexes as general performances in that month to make connection with the unemployment rate. The below figure displays a linear trend of unemployment rate versus some major economic indexes over time (LNU04028615 is the unemployment rate).



(Figure.1)

It is reasonable to have some roughly increasing trendlines for the economic indexes overtime, as a result of technology development and globalization. However, a decline of economy will generally lead to a consequence of unemployment. Based on my observation, most increases in unemployment rate were incurred after a downgrade of economic indexes. For example, unemployment rate had an increasing trend when NASDAQ100, WILL5000PR and NASDAQCOM started to decline around the middle of 2000. Also, another decrease of these three economic indexes and an increase of unemployment rate occurred in the end of 2008 due to global financial crisis. The same pattern is observed in the beginning of 2020 due to COVID-19. All these information suggest economic indexes could be used for unemployment rate prediction.

3.2 Unemployment Rate Modification

As unemployment rate is displayed numerically and varies by each month, I changed it into indicator of unemployment rate change (in 0,1 format) to fit the model. For this indicator, 0 means the unemployment rate of one month has decreased compared to previous month and 1 means the unemployment rate of one month has increased or not changed compared to previous month. The emergence of a sequence of 1 in consecutive months is indicating a growing trend of unemployment rate. The emergence of a sequence of 0 in consecutive months is indicating a declining trend of unemployment rate.

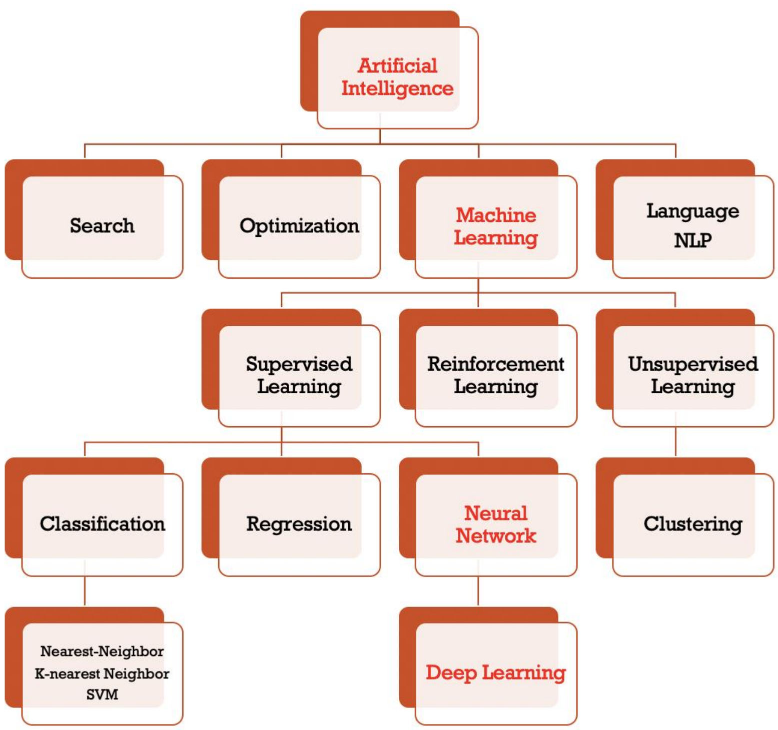
3.3 Testing and Training Separation

The data is randomly separated into two groups, one is training and another is testing. I have tried multiple separations such as 20%-80%, 30%-70%, 40%-60% and 50%-50%. The model does show difference between different training and testing data separations, which indicates the model is not completely converged. However, the model is able to reach to more than 80% accuracy which I believe is useful in terms of explaining the relationship between unemployment rate change and major economic indexes, at least in certain degree.

4. Model Description

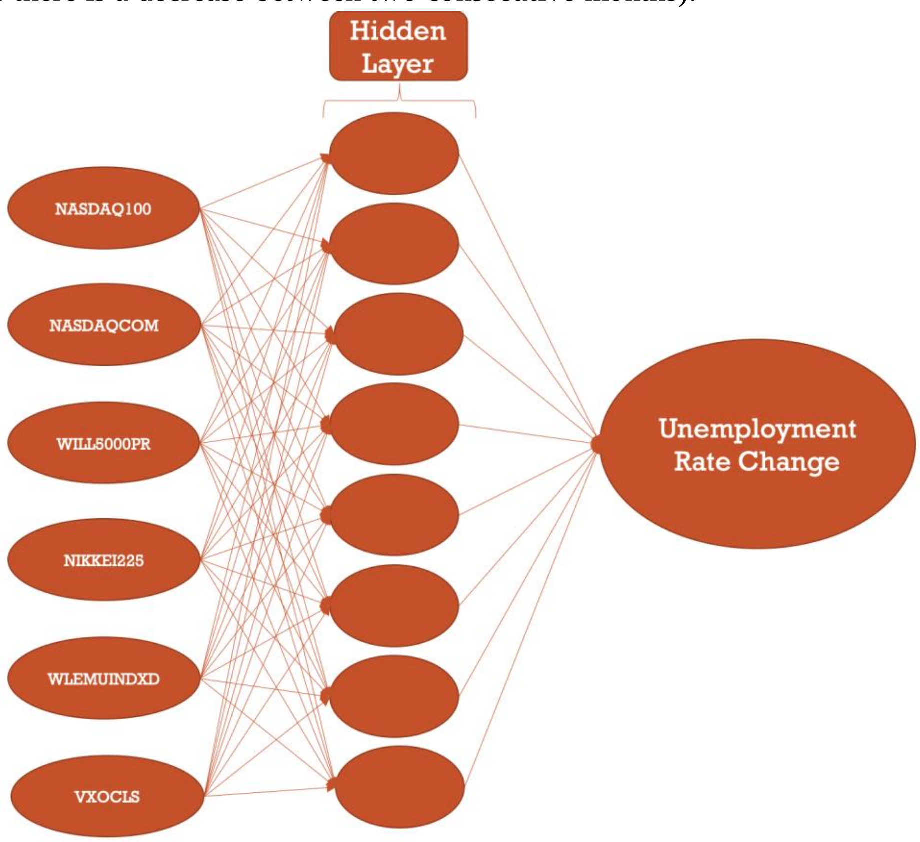
My underlying subject is modeling economical impact on unemployment rate change, and the dataset I am using well suits for a supervised learning, because target metric (unemployment rate) is available. Therefore neural network models fit to my research very well.

This course enhanced my understanding in computational statistics. The follow graphs show my understanding of the different AI and computational statistics techniques and their relationships beyond what course material presented directly.



(Figure. 2)

I have constructed multiple neural network models, and finalized my model with the following parameters. There are eight neurons each represents an economic index. I only include 1 hidden layer to enhance model flexibility. I did test multiple layers settings however, they showed suboptimal accuracy, I believe overfitting is the main reason for it. Therefore, I choose a model with only 1 hidden layer and 8 neurons within the layer. There is only one neuron for the output which is a binary data indicating unemployment rate change (1 represents there is an increase between two consecutive months, 0 represents there is a decrease between two consecutive months).



(Figure.3)

In the hidden layers, rectified linear unit (ReLu) activation function is used and for the final layer (output), sigmoid activation function is selected. Other activation functions such as linear activation function, step function, are also tested. ReLu and Sigmoid are selected because they produce better model accuracy.

The neural network model is fundamental, most of my research focuses on data modification, transformation so that neural network model can be used to produce meaningful results.

5. Results

I have tested multiple testing and training dataset separation scenarios. Smaller testing size which corresponds to larger training size had the best accuracy level of 91.67% as presented in Figure 3. The efficiency of convergence is high as the accuracy improves quite fast.

The accuracy definition is a standard metrics in TensorFlow package, however, there are many other definitions of accuracy measure statistically can could be considered and compared. I did not focus on that either. I believe a consistent accuracy comparison of neural network models is sufficient for model validation.
Test size of 0.2:

```
 32/333 [=>............................] - ETA: 0s - loss: 94.8375 - accuracy: 0.8438
333/333 [==============================] - 0s 56us/sample - loss: 140.9268 - accuracy: 0.8408
Epoch 17/20

 32/333 [=>............................] - ETA: 0s - loss: 92.1191 - accuracy: 0.8125
333/333 [==============================] - 0s 63us/sample - loss: 93.7349 - accuracy: 0.8468
Epoch 18/20

 32/333 [=>............................] - ETA: 0s - loss: 142.5508 - accuracy: 0.7812
333/333 [==============================] - 0s 72us/sample - loss: 55.9844 - accuracy: 0.8769
Epoch 19/20

 32/333 [=>............................] - ETA: 0s - loss: 19.3418 - accuracy: 0.9688
333/333 [==============================] - 0s 51us/sample - loss: 26.9999 - accuracy: 0.9159
Epoch 20/20

 32/333 [=>............................] - ETA: 0s - loss: 13.5067 - accuracy: 0.9375
333/333 [==============================] - 0s 57us/sample - loss: 9.9066 - accuracy: 0.9369
84/1 - 0s - loss: 2.6238 - accuracy: 0.9167

Process finished with exit code 0
```

(Figure. 4)

However, as the testing size increases, the accuracy levels drops significantly from testing size 0.5 of 88.52% to testing size 0.6 of 58.53%. The accuracy of the model is also depends on the initial random weight selection, however, it does converge to a relatively stable results especiually as the training dataset size increases.

Test size of 0.5:

```
 32/208 [===>..........................] - ETA: 0s - loss: 295.5945 - accuracy: 0.8125
208/208 [==============================] - 0s 57us/sample - loss: 244.8703 - accuracy: 0.8221
Epoch 17/20

 32/208 [===>..........................] - ETA: 0s - loss: 266.1645 - accuracy: 0.8125
208/208 [==============================] - 0s 60us/sample - loss: 189.6674 - accuracy: 0.8365
Epoch 18/20

 32/208 [===>..........................] - ETA: 0s - loss: 119.2233 - accuracy: 0.9062
208/208 [==============================] - 0s 67us/sample - loss: 136.6411 - accuracy: 0.8606
Epoch 19/20

 32/208 [===>..........................] - ETA: 0s - loss: 98.3144 - accuracy: 0.9062
208/208 [==============================] - 0s 72us/sample - loss: 100.0628 - accuracy: 0.8798
Epoch 20/20

 32/208 [===>..........................] - ETA: 0s - loss: 148.7449 - accuracy: 0.8750
208/208 [==============================] - 0s 77us/sample - loss: 74.5955 - accuracy: 0.9231
209/1 - 0s - loss: 70.2971 - accuracy: 0.8852

Process finished with exit code 0
```

(Figure. 5)

Test size of 0.6:

```
 32/166 [====>.........................] - ETA: 0s - loss: 2969.8450 - accuracy: 0.2188
166/166 [==============================] - 0s 66us/sample - loss: 2614.4915 - accuracy: 0.3313
Epoch 17/20

 32/166 [====>.........................] - ETA: 0s - loss: 2868.0781 - accuracy: 0.5625
166/166 [==============================] - 0s 78us/sample - loss: 2364.0307 - accuracy: 0.6024
Epoch 18/20

 32/166 [====>.........................] - ETA: 0s - loss: 1455.1835 - accuracy: 0.7188
166/166 [==============================] - 0s 66us/sample - loss: 2179.2804 - accuracy: 0.6265
Epoch 19/20

 32/166 [====>.........................] - ETA: 0s - loss: 2827.8977 - accuracy: 0.5938
166/166 [==============================] - 0s 72us/sample - loss: 2032.3065 - accuracy: 0.6386
Epoch 20/20

 32/166 [====>.........................] - ETA: 0s - loss: 2012.5742 - accuracy: 0.6562
166/166 [==============================] - 0s 84us/sample - loss: 1878.7186 - accuracy: 0.6386
251/1 - 0s - loss: 1653.1060 - accuracy: 0.6853

Process finished with exit code 0
```

(Figure. 6)

6. Conclusion

There are multiple potential issues or limitations within the data and model I am working with. The neural network model I am proposing here is already extremely simplified, however, it still includes 6 (inputs)* 8 (neurons in hidden layer) + 8 (neurons in hidden layer) *1 (output) = 48 + 8 = 56 weights. These settings could be easily expanded into a sophisticated model by adding in more hidden layers or convert into a deep learning model. However. I did not pursue that direction in order to avoid over fitting. My dataset is relatively small, which more properly should be considered as social statistical analysis instead of a big data analysis.

The model does show difference between different training and testing data separations, which indicates the model is not completely converged. However, the model is able to reach to more than 80% accuracy which I believe is useful in terms of explaining the relationship between unemployment rate change and major economic indexes, at least in certain degree.

As the model is providing a decent prediction results and accuracy, it could be used to explain the relationship between employment rate change and major economic indexes. However, the model is still limited because of the small data size. The reason of small size is that unemployment rate is reported in a monthly basis. Even though the starting point of dataset is in the beginning of 1986, the size is still small. The starting weights for the model are randomly chosen, thus small dataset have a hard time to get converged model result. Also, there are more correlated variables, which need to be explored, could be included in the model. For example, try to search more findings on globalization drivers and more policy related data. A possible way to expand the data could be calculating a half month unemployment rate by taking the average of previous month rate and current month rate, or a quarter month unemployment rate using quantiles between previous month rate and current month rate.

7. Reference

[1] John H. Boyd (2004). The Stock Market's Reaction to Unemployment News: Why Bad News Is Usually Good for Stocks.
[2] Johannes Bock (2008). Quantifying macroeconomic expectations in stock markets using Google Trends.
[3] FRED ECONOMIC DATA. Retrieved from https://fred.stlouisfed.org/