**3. Solution of the Requirement One (America)**

**3.1 Analysis of the Requirement**

For Requirement One, we combine the speeches of two candidates and analyze them by extracting key words from the speeches for the four dimensions of Consumption, Investment, Government Purchase, and Net Export. At the same time, the LSTM algorithm in machine learning will be introduced to predict and analyze the GDP trends of different leaders.

**3.2 Data Explanation**

**3.2.1 Consumption**

In macroeconomics, Consumption refers to the total expenditure of consumer goods consumed by one person or one country in a certain period. In the process of analyzing consumption, CPI and stock trading volume of consumer sector are selected for analysis.

* CPI: The Consumer Price Index (CPI) is an index that measures the purchase prices of a selected basket of consumer goods, and the weight of each commodity is determined based on its share of urban consumer spending between 1982 and 1984. It is the index of price change that reflects the product and service price statistics that are related to people's life, and it is usually used as an important index to observe the level of inflation.
* Consumer stocks: the change of the consumer stocks can be from a certain consumer degree of positive reaction. We selected the five relatively representative consumer stocks in different areas, which are Coca-Cola, Procter & gamble, PepsiCo, Wal-Mart and Philip Morris. We used python to crawl the daily stock index from 2015 to 2020, in order to analyze consumer stocks situation changes.

**3.2.1.1 PCA**

We use principal component analysis (PCA) to integrate each evaluation index into an overall index, that is, to prevent Multicollinearity.

Principal component analysis (PCA) takes correlated data from multiple indicators (e.g., X1, X2… XN) and using spatial downscaling of high dimensional variables, where the components reflect most of the original variable information that is not duplicated. Mathematically, the original N indicators are usually linearly combined to form a new composite indicator. The most widely used method is to use the variance of the first composite (F1) selected, the larger the variance, the more information it contains. In the analysis of Consumption-related data, we selected principal components with a cumulative contribution greater than 80%.

For the U.S. consumption data, we have six variables, the five consumer sector stock indices and the residential consumption index, with a total of 1,478 pieces of data.

**1. Standardize the data, calculate the mean value and standard deviation of the column and conduct z-score standardization.**





So the standardized data is:



Therefore, the normalized matrix is



**2. Calculate the eigenvalue and eigenvector of R**



In which r refers to 

**3. Calculate the contribution rate and cumulative contribution rate of principal components**

Contribution Rate: 

Cumulative Contribution: 

The following table describes the eigenvalues of the correlation coefficient matrix, the corresponding eigenvectors and contributions.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Eigenvalues | a1 | a2 | a3 | a4 | a5 |
| KO | 0.421 | 0.795 | -0.238 | 0.333 | -0.156 |
| PG | 0.453 | -0.072 | 0.673 | -0.181 | -0.552 |
| PEP | 0.467 | 0.145 | 0.094 | -0.559 | 0.663 |
| WMT | 0.453 | -0.428 | 0.116 | 0.693 | 0.343 |
| CPI | 0.440 | -0.399 | -0.685 | -0.252 | -0.338 |
| Eigenvalues Value | 4.343 | 0.332 | 0.202 | 0.082 | 0.0413 |
| Contribution Rate | 0.869 | 0.067 | 0.040 | 0.016 | 0.008 |
| Cumulative Contribution | 0.869 | 0.935 | 0.975 | 0.992 | 1.000 |

It can be seen from the table that the cumulative contribution rate of the first principal component reaches 86.9%, so only the first two principal components are effective to summarize the original variables.



We use python to calculate the reduced dimensional data, which is the U.S. Consumption Index. (Data is attached to the supporting material)

**3.2.2 Investment**

In the United States, investment is mainly reflected in the stock market, and there are three indices that can reflect the trends and tendencies of investment, namely the NASDAQ, Dow Jones and S&P 500. The Dow Jones has been established for the longest time and is relatively more stable.

**3.2.3 Government Purchase**

The Government to buy (Government Purchases) refers to purchase goods and services from all levels of Government spending, it is a part of Government spending. In order to analyze the changes in government purchases in recent years, instead of analyzing the government purchase data, we selected the federal government financial budget from the Open data of the U.S. Bureau of Commerce to simulate the government purchase.

**3.2.4 Net Export**

NX which refers to the amount of net exports, can be reflect the Balance of trade(the Balance of imports and exports measured in local currency in a certain period of time).

The Balance of trade, also known as net exports, is the difference between a country's imports and exports measured in its own currency over a period of time and forms part of the Current account. The trade account exports more than the import is called "surplus", "surplus" or "surplus"; Otherwise, it is called "deficit" or "deficit" or "excess". Our open data set of the US Bureau of Commerce intercepted the data from 2015 to November 26, 2020 for analysis, and used it to show the change of the net export quota.

**3.2.5 COVID-19**

**3.2.6 Daily GDP**

GDP (Gross Domestic Product) is a good indicator to reflect the economic operation, so we take it as the dependent variable to simulate and forecast. We used the daily closing price of NASDAQ index to simulate the daily GDP according to the following formula:



In which  represents the closing number, and represents the monthly GDP.

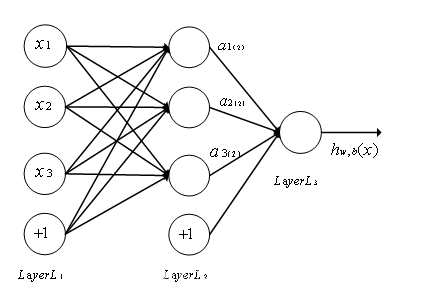
|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **C** | **I** | **G** | **NX** | **new\_cases** | **new\_deaths** | **GDP** |
| **C** | 1.00 | 0.47 | 0.93 | -0.15 | 0.37 | 0.38 | 0.43 |
| **I** | 0.47 | 1.00 | 0.51 | -0.30 | 0.14 | 0.26 | 0.18 |
| **G** | 0.93 | 0.51 | 1.00 | -0.16 | 0.28 | 0.33 | 0.46 |
| **NX** | -0.15 | -0.30 | -0.16 | 1.00 | -0.22 | -0.35 | 0.10 |
| **new\_cases** | 0.37 | 0.14 | 0.28 | -0.22 | 1.00 | 0.68 | -0.29 |
| **new\_deaths** | 0.38 | 0.26 | 0.33 | -0.35 | 0.68 | 1.00 | -0.46 |
| **GDP** | 0.43 | 0.18 | 0.46 | 0.10 | -0.29 | -0.46 | 1.00 |

**3.5 LSTM**

**3.5.1 Choosing of the Model**

We used algorithm of Stacking integration learning, GBXT algorithm, LGB algorithm, and LSTM algorithm to fit and test predict Formula 1, and we find THE LSTM algorithm has the best testing result.

LSTM neural network model is a special cyclic neural network model (RNN). One of the key points of RNN is that it can connect the previous information to the current task by memorizing the previous information and applying it to the calculation of the current output, so it is a very effective method to predict time series data. LSTM mainly solves problems such as GRADIENT disappearance and gradient explosion of RNN, and has its own memory and can make accurate prediction. As a very powerful prediction tool, LSTM is specially designed to solve long-term problems.



**3.5.2 Testing of the Model**

**1.** **The transformation of the raw data**

First, under the influence of different dimensions, we normalized the raw data by Max-min: according to the following formula



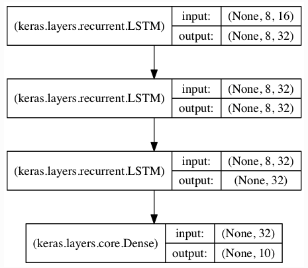
We set the first 80% of the data as a training set, which is a total of 1,200 pieces, and the rest as a test set to evaluate the accuracy of the model.

**2. Selection of Eigenvalues**

According to the initial definition, we had six variables, and we divided them into two categories for simulation testing. The first includes consumption, investment, government purchase and net export (NX), which are the elements of the four-sector cycle in macroeconomics. The second category added the average number of deaths per day and the average number of cases per day to the epidemic data based on the first category. Compare which data classification is more appropriate.

**3. LSTM Neural Network Training**

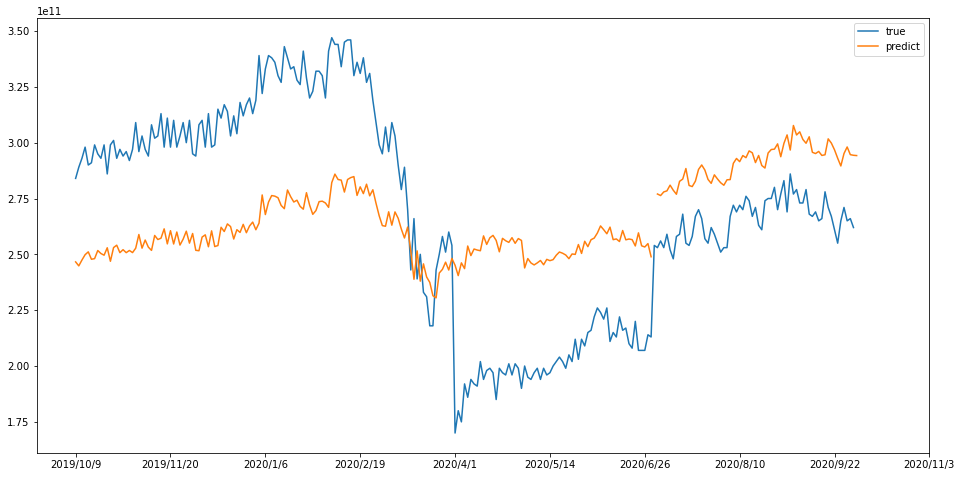
We used Python Keras package for training. With two hidden layers, the epoch trained three times, the batch\_size (number of steps per training) was 8, the time step was 1, the loss function was mean square error, the activation function of LSTM module was tanh, and Adam optimizer was used to optimize and output the result.



**4. The Result of Training**

Two different training models can be obtained for the two feature choices. The data are inverse normalized to make an image. When training for the first time, the loss is about 0.036, and as the number of training sessions increases, the loss shows a decreasing trend, and when training more than three times, the loss value is stable at about 0.006. In order to prevent the occurrence of over-fitting of the model, the team set the training period as three cycles.

For the existing data, our fit of the second category accuracy is as high as 86%, so we decided to use the LSTM as the main model to predict the future GDP.

****

**3.5.3 Predict the Data**

In the analysis of the indicators for predicting the U.S., we use different methods for different indicators. First, we use the AHP model to analyze the impact of different candidates' political platforms on various aspects. For the covid-19 data, we use the SIRD epidemic model to predict different infection and cure rates for different candidate attitudes, and AdaBoost combined with the 60-day moving average for the other variables.

**3.5.3.1 Text Analysis**

* **TP\_IDF**

Term frequency (TF) refers to the number of times a given word appears in the document, while inverse document frequency (IDF) is a measure of the general importance of the word.

In order to present the key words in the debates and speeches of Trump and Biden, the following steps are taken：

First, searching materials such as policy guidelines, speeches and campaign debates from both Trump’s and Biden’s campaign websites. Using python package ‘jieba’ to separate the words of two candidates.

Secondly, creating a thesaurus based on the separated words, combined with common political and economic terms.

Third, we counted the term frequency (TF) and inverse document frequency (IDF) of the two corpus separately, and obtained the TF\_IDF weight ranking, which will be used as a reference for selecting the AHP criterion layer in the following steps.

C:\Users\LENOVO\AppData\Local\Temp\ksohtml23988\wps11.jpg

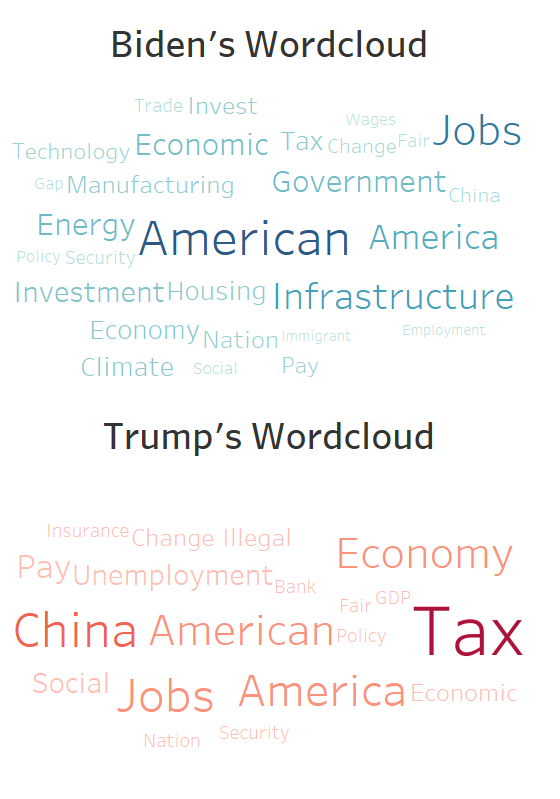
C:\Users\LENOVO\AppData\Local\Temp\ksohtml23988\wps12.jpg

C:\Users\LENOVO\AppData\Local\Temp\ksohtml23988\wps13.jpg

D: Total number of words in the document

Dw: the number of a specific word in a text

We collected a large number of Biden and Trump speeches and debate transcripts and counted their word frequency, as shown in the following word cloud



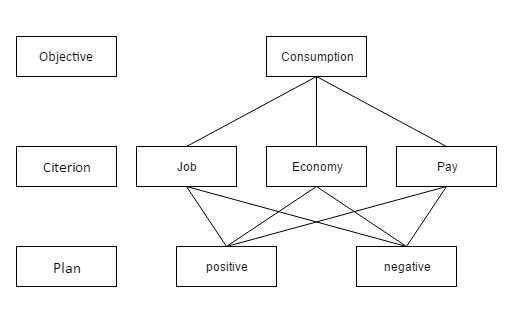
* **AHP**

For each variable (C, G, I, NX), three representative words were selected to analyze the tendencies of the two candidates in the four directions using AHP.

The main characteristic of AHP is to transform the human judgment to the comparison of the importance between several factors by establishing the hierarchical structure, thus transforming the qualitative judgment which is difficult to quantify into the comparison of the operational importance. We will take Trump's Consumption related attitude as an example to explain the establishment and solution of AHP model.

1. Analyze the relationship between the factors in the system and establish the hierarchy of the system.

The hierarchical model is divided into three levels. The top level is the target level, that is, the selected variables of the four departments of economics selected by us; the middle level is the criterion level with selected the vocabularies; and the bottom level named program level contains the positive and negative degree of vocabularies.



We judge relative importance according to the following table:

|  |  |
| --- | --- |
|  | Meaning |
| 1 | The Same Important |
| 3 | Slightly More Important |
| 5 | Significantly More Important |
| 2、4 | The mean of the Two Importance |
| Reciprocal | If the scale of A and B is 3, then B and A are 1/3 of the scale. |

We construct judgment matrix O-C, and let the three indexes of C layer compare in pairs

|  |  |  |  |
| --- | --- | --- | --- |
| O | Job | Economy | Pay |
| Job | 1 | 3 | 5 |
| Economy | 1/3 | 1 | 1 |
| Pay | 1/5 | 1 | 1 |

We calculate the maximum eigenvalue, and calculate Ci and CR with the following formula





The CI value of the matrix is 0.0146, and the RI value of the matrix is 0.58. Therefore, the CR value of the matrix is 0.0252 less than 0.1 and pass the consistency test

Therefore, we use the eigenvalue method to calculate its eigenvectors, and get the normalized weight vectors as follows

The corresponding feature vector is: [0.9385 0.2639 0.2226]



After normalization, the weight vector is: [0.6586 0.1852 0.1562]



Repeat the above steps to calculate the eigenvalues and eigenvectors of each C-P matrix to obtain the following consistency matrix, all of which pass consistency test.

|  |  |  |
| --- | --- | --- |
| C1 | positive | negative |
| positive | 1 | 5 |
| negative | 1/5 | 1 |

|  |  |  |
| --- | --- | --- |
| C2 | positive | negative |
| positive | 1 | 3 |
| negative | 1/3 | 1 |

|  |  |  |
| --- | --- | --- |
| C3 | positive | negative |
| positive | 1 | 3 |
| negative | 1/3 | 1 |

The table below shows the statistical score results of final Consumption related two different attitudes

|  |  |  |  |
| --- | --- | --- | --- |
|  | The Weight Indicator | Positive | Negative |
| Job | 0.6586 | 0.8333 | 0.1667 |
| Economy | 0.1852 | 0.75 | 0.25 |
| Pay | 0.1562 | 0.75 | 0.25 |

Therefore, the PC score is 0.80486138, and the NC score is 0.19513862

Repeating the above steps, we separately calculated the positive and negative ratings for each of the two categories of leaders, as shown in the table below

|  |  |  |
| --- | --- | --- |
|  | Trump | Biden |
| PC | 0.80486138 | 0.79580667 |
| NC | 0.19513862 | 0.20419333 |
| PI | 0.8030621 | 0.75872151 |
| NI | 0.1969379 | 0.24127849 |
| PG | 0.8030621 | 0.80392009 |
| NG | 0.1969379 | 0.19607991 |
| PNX | 0.79843895 | 0.8030621 |
| NNX | 0.20156105 | 0.1969379 |

For each category of positive trends, we normalize it using the following formula, which is used to predict economic trends in each category of weighted calculations.



|  |  |  |
| --- | --- | --- |
|  | Trump | Biden |
| PC | 1.005656832 | 0.994343168 |
| PI | 1.028390995 | 0.971609005 |
| PG | 0.999466086 | 1.000533914 |
| PNX | 0.997113239 | 1.002886761 |

The final formula of GDP direction is as follows:



**3.5.2.2 SIRD**

Unlike the epidemic in China, the COVID-19 epidemic in the United States is not yet under control and will have long-term impacts. Therefore, we will consider the impact of the epidemic on the economy and add it as an independent variable.

Basic assumptions as follows

|  |  |
| --- | --- |
| Symbol | Explaination |
|  | Percentage of current susceptible population |
|  | Percentage of the population currently infected |
|  | Percentage of population currently cured |
|  | Percentage of the population currently dying |
|  | Number of effective contacts |
|  | Daily Contact Rate |
|  | Daily Cure Rate |
|  | Daily Dead Rate |

Carriers and incubation periods are not considered so the following relationships can be listed for all groups.

|  |  |
| --- | --- |
|  | （4） |

Therefore, the daily contact rate has the following expression (i.e., the basic regeneration rate)



According to the above relationship we can list the difference equation for each group.

For the infected group

|  |  |
| --- | --- |
|  | （5） |

For the safe group

|  |  |
| --- | --- |
|  | （6） |

For the recovery group:

|  |  |
| --- | --- |
|  | （7） |

For the dead group:

|  |  |
| --- | --- |
|  | （8） |

Since the value is of a small period of time , the difference equations are converted into differential equations.

|  |  |
| --- | --- |
|  | （9） |

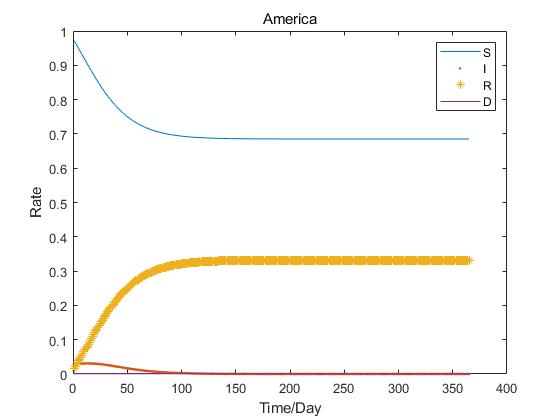
|  |  |
| --- | --- |
|  | （10） |

|  |  |
| --- | --- |
|  | （11） |

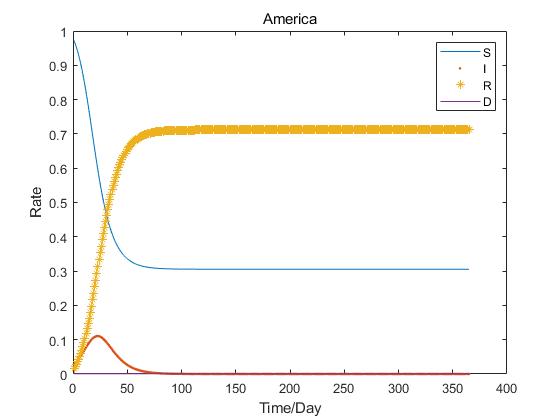
|  |  |
| --- | --- |
|  | （12） |

With the analysis of epidemic in the United States and the census, and with the population of 303 million, we can calculate the current infection rate, recovery rate and death rate to forecast the evolution of the epidemic. We calculated the average infection rate from the beginning of the epidemic to the present as the benchmark, then added the positive and negative factors of the policy, and used MATLAB to predict the development of the epidemic in the next first year.

Biden's policy includes the management of the COVID-19 outbreak, which will increase health care spending on the epidemic. Therefore, it has a positive impact on the control of the epidemic situation, the following graph is the forecast of the development of the covid-19 under the Biden’s policy



According to the news and his policy, Trump has a negative attitude towards the epidemic, adopts herd immunization, and intends to abolish universal health insurance, so fewer medical resources will be used for the increase of the epidemic infection rate and the reduction of the cure rate, thus the graph below shows the forecast trend of the epidemic development under Trump's leadership.

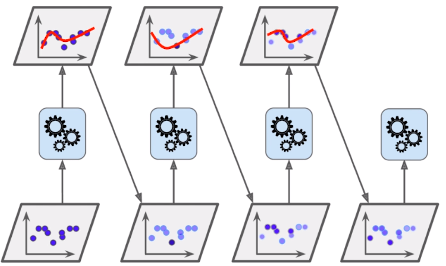


**3.5.2.3 Ada boost and 60 average line (C)**

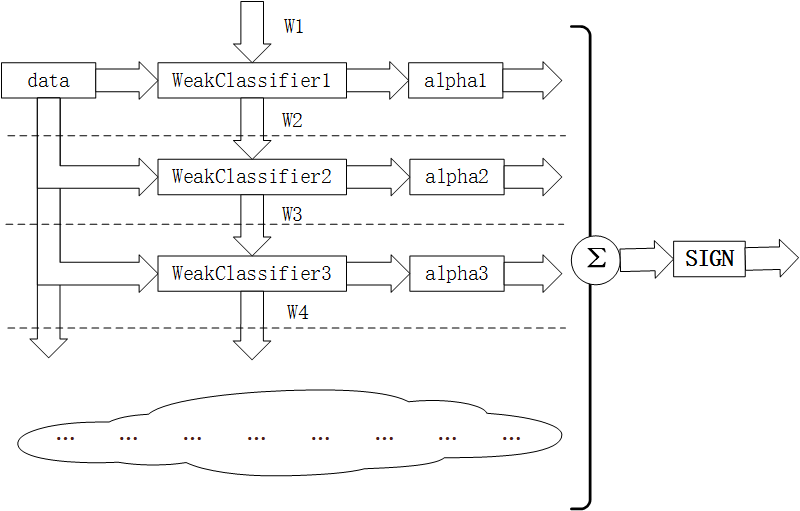
**We use Ada boosts, companying with**

* **Adaboost**

The AdaBoost algorithm is a boosting method that takes several weak classifiers and combines them into a strong classifier. In the AdaBoost algorithm, the weights of the weak classifiers are adjusted after each training round, and the weight of the misclassified points in the previous training round will increase.



Since the relationship between several classifiers in Adaboost is such that the N classifier is more likely to split the data that the N-1 classifier does not, the data that was previously split will be split at the same time. So in Adaboost, each weak classifier has its own most important point of interest, and each weak classifier is only interested in a part of the whole data set, so they must be combined together to be effective. The weights are based on the classification error rate of the weak classifiers. The general rule is that the lower the error rate of a weak classifier, the higher the weight.



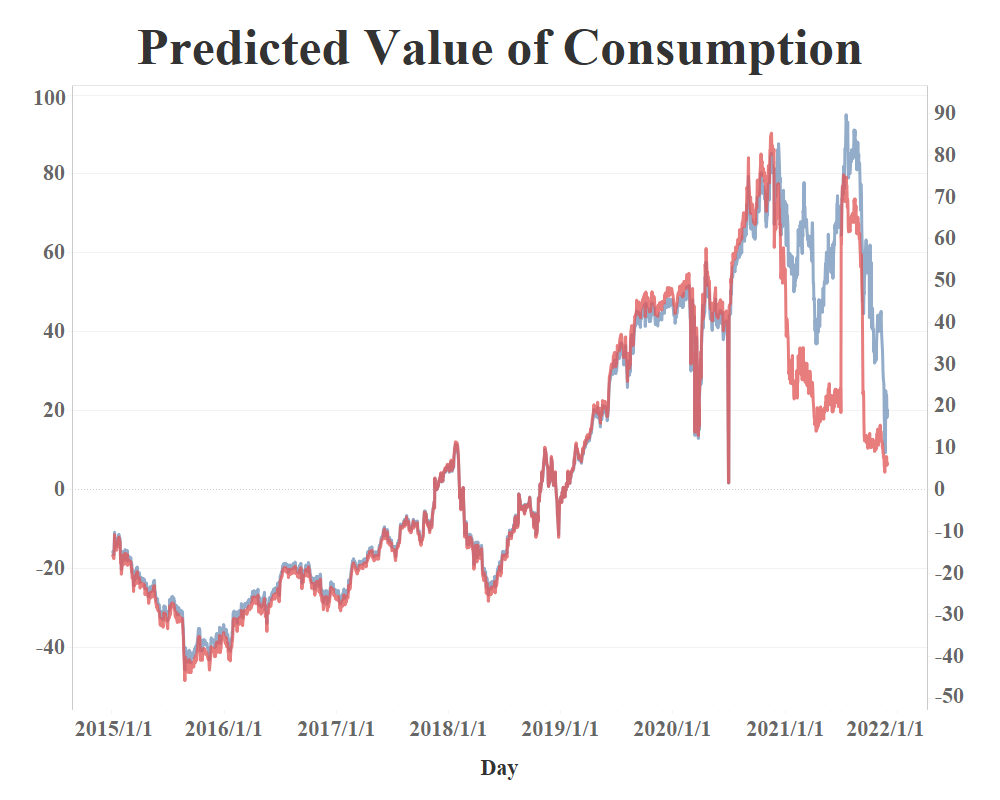
* 60-Day Moving Average

The moving average is the average closing price of a stock for the previous fixed days, and its significance lies in the fact that it reflects the long-term or short-term changing trend . 60-day moving average is generally a medium- to long-term trend, and the 60-day moving average is the average closing price of the last two months, which is important for the later trend of a stock.

Therefore, we use the 60-day moving average in combination with AdaBoost to make rolling forecasts that better reflect the impact of economic factors in the near future.

Because of this prediction of economic development after the election, we use various data from before and after the 2015 election as weights, while adding thresholds to slow its convergence.

Taking the U.S. consumption data as an example, the following graph is predicted to meet the trend well.



**3.5.2 The Result**

After predicting each input data, we train the LSTM model by weighting it with political preferences, and then we obtain the following graph of GDP trends under different candidates.

