EMSE 6992 Final Paper

Yiwei Wang

introduction

There are many factors influencing a nation’s economic. Among them, some have greater impact than other. This time, I choose 9 of them, and want find weather they have something to do with GDP.

Abstraction

In this research, I first apply the methods of data manipulation, aggregation, and visualization. Second, I apply statistical analysis. Third, a full machine learning pack was applied. However, since it is the first time, and lack of many initial knowledge, the whole process is in a mass.

Business case evaluation

The main purpose of this research is to have some basic understanding of big data life cycle, and use some basic methods in all the procedures.

Data Identification

For the US economic related data, I choose the dataset from the world bank.

For the stock market related data, I choose the dataset of S&P500, NASDAQ, and D&J.

Data Acquisition & Filtering

Getting the dataset from the World Bank and yahoo finance.

Since the original dataset is too large, and many terms did not start recording in a quite long time, under professor’s instruction, I shrink the large dataset into a ten by ten dataset. Among them, there are S&P500 closing price based on year-average from 2006 to 2015, and employment ratio, household expenditure, export value index, energy use, net portfolio investment, import value index, and population from the same period of time. The annual GDP is also included.

Data Extraction

The S&P500 number are calculated based on data got from Yahoo Finance, and is in perfect format.

Other nine features are from the data provided by the World Bank, and the World Bank have preprocessed all the data, and they are also in perfect format.

Data Validation & Cleansing

Since the dataset used in this research are all preprocessed by reliable organizations, I suppose I do not need pay attention on this step this time.

Data Aggregation & Representation

In this research, data are from two datasets. The World Bank data are annually based, and the stock market data from Yahoo Finance are daily based. Therefore, when aggregate S&P500 to the economic features, I choose closing price every trading day in a year, and get the average number to representing the stock market performance in that year.

Data Analysis

For the whole project, there are three sections.

In the section one, according to the sample provided, a simple analysis on the dataset is performed.

First, get all the number of the chosen dataset. Then, try to find some unique feature in each column. Since the chosen dataset is too small, every data is unique in this case. Then, try to find whether there exists some significant relationship between two features, and there is no such significant relationship. While plotting scatter plot between two features, it seems there exist a linear relation between them. Then, methods used in sample codes go beyond my knowledge at this time, and I do not understand what happened. Even though I got some graphs that may say something about the dataset, I cannot tell what their mean.

In the section two, according to the sample, the first step is doing linear regression on the dataset. Since the dataset is very small, the regression result means nothing at all. Then, a logistical regression was performed. I did not find the meaning of doing logistical regression on my purely number based dataset. After logistical regression, some method I do not know performed. Then, comes KNeighborhood classifier. However, from the graph, it seems nothing is revealed, and at present, I do not know how to interpret them. In the end of section two, something worrying happened. It is clear that I am running into some kind of error, however, I do not know how to deal with it.

In section three, the example is a complete procedure on how to perform machine learning. However, the example is based on data included in the sklearn module, and my whole effort is to apply my dataset to the example. But my effort seems meaningless. In application step, detect outlier in dataset, it runs with no graph with meaning with it, since I do not know how to make adjustment to the original codes that can make it suitable for me. The adjustment is the major problem I met through out section three. Since I do not understand what I am doing with those codes, all procedures come to error in different types.

Data Visualization

Since nothing valuable got from the analysis, the only visualization graphs are from the very first of the analysis, the relationship scatter plots, and linear regression plots.

Utilization of Analysis Results

Reference

The World Bank. (2017). United States Data. Available from the World Bank Web site: <http://api.worldbank.org/v2/en/country/USA?downloadformat=csv>

Yahoo Finance. (2017). Yahoo Finance historical data. Available form Yahoo Finance Web site:

S&P500:

<https://query1.finance.yahoo.com/v7/finance/download/%5EGSPC?period1=1509995479&period2=1512587479&interval=1d&events=history&crumb=jLkEr1TNo4i>

Dow Jones:

<https://query1.finance.yahoo.com/v7/finance/download/%5EDJI?period1=1509995543&period2=1512587543&interval=1d&events=history&crumb=jLkEr1TNo4i>

NASDAQ:

<https://query1.finance.yahoo.com/v7/finance/download/%5EIXIC?period1=1509995593&period2=1512587593&interval=1d&events=history&crumb=jLkEr1TNo4i>