**Final Project**

FTEC 6321 Advanced Statistical Methods

Yu Chieh Hung (Phyllis)

1. **Problem Identification**

With the pandemic, people nowadays are more likely to stay at home no matter working, attending classes or relaxing. One thing I do a lot than last year is watching Netflix. I usually went out to shopping, having dinner in the restaurant and watching movies in the theater; however, I cannot do anything now. Therefore, I watched dramas more often on Netflix, Netflix is now the app I use every day. Furthermore, I want to dig more into the dataset of Netflix, so I get 21 financial data from the financial statements and compute 5 ratios, price-to-earnings ratio (P/E ratio), EV / EBITDA, debt-to-equity ratio (D/E ratio), annual return on equity (ROE) and price-to-book ratio (PB ratio), to analyze. I want to know whether the stock price of Netflix has the correlation with the unemployment rate, P/E ratio, EV / EBITDA, D/E ratio, annual ROE and PB ratio. How can these six features affect the log return price of Netflix. Thus, we can predict the future stock return with 5 models, Multiple Regression model, Polynomial Regression model, Support Vector Regression model, Decision Tree model and Random Forest model.

Since the pandemic keeps ruining our lives, many people were fired from the company or lose their jobs. Thus, with those who doesn’t have jobs, will they watch Netflix more often or less often. For me, both situations make sense, they might watch Netflix more often to release their stresses or stop subscribing Netflix to concentrate on looking for a job. Either taking Netflix as an entertainment or interference, Netflix’s stock price will change. On the other hand, the changes on its financial statements will have influence on the stock price too. The only different is how much difference and how correlated with these features.

1. **Data Gathering**

I found the US Unemployment dataset on Kaggle provided by U.S. BUREAU OF LABOR STATISTICS and I got the financial dataset on WRDS within the authorization from UTD

McDermott Library. Links and the datasets are attached in the Reference Part at the end of this document.

First of all, because the Netflix stock had data from 2011, I then decided to analyze the data from 2002 to 2019. Second, because the data of US Unemployment rate is monthly and the financial data is annually, so I take average of 12 months unemployment rate to annual unemployment rate. There is no missing value or any nan value to deal with. Third, I calculate the stock’s log return and remove the first blank row, so there are 17 rows with annually values in 17 years from 2012 to 2019. Forth, I compute five independent variables from 21 variables. I show the results of 6 features and log returns in the next concept.

1. **Feature Engineering**

I choose annual unemployment rate for my first novel feature; this is what I want to figure to and what I am curious about. I selected 21 variables and decided to take five features from these values.

The price-to-earnings ratio (P/E ratio) is the ratio for valuing a company that measures its current share price relative to its per-share earnings ([EPS](https://www.investopedia.com/terms/e/eps.asp)). The P/E ratio is also sometimes known as the price multiple or the earnings multiple. P/E ratios are used by investors and analysts to determine the relative value of a company's shares in an apples-to-apples comparison. It can compare a company against its own historical record or to compare aggregate markets against one another or over time. I believe it will have some impact on the stock price.

The EVEBITDA is Enterprise Value divided by Earning before Interest, Taxes, Depreciation & Amortization. EV/EBITDA is a ratio that compares a company’s [Enterprise Value](https://corporatefinanceinstitute.com/resources/knowledge/valuation/what-is-enterprise-value-ev/) (EV) to its Earnings Before Interest, Taxes, Depreciation & Amortization ([EBITDA](https://corporatefinanceinstitute.com/resources/templates/valuation-templates/what-is-ebitda/)).  The ratio of EV/EBITDA can compare the entire value of a business with the amount of [EBITDA](https://corporatefinanceinstitute.com/resources/templates/valuation-templates/what-is-ebitda/) it earns on an annual basis.  This ratio tells investors how many times EBITDA they have to pay, were they to acquire the entire business. Therefore, it can be a key factor whether investors will invest NFLX or not, and it will surely reflect on the stock return.

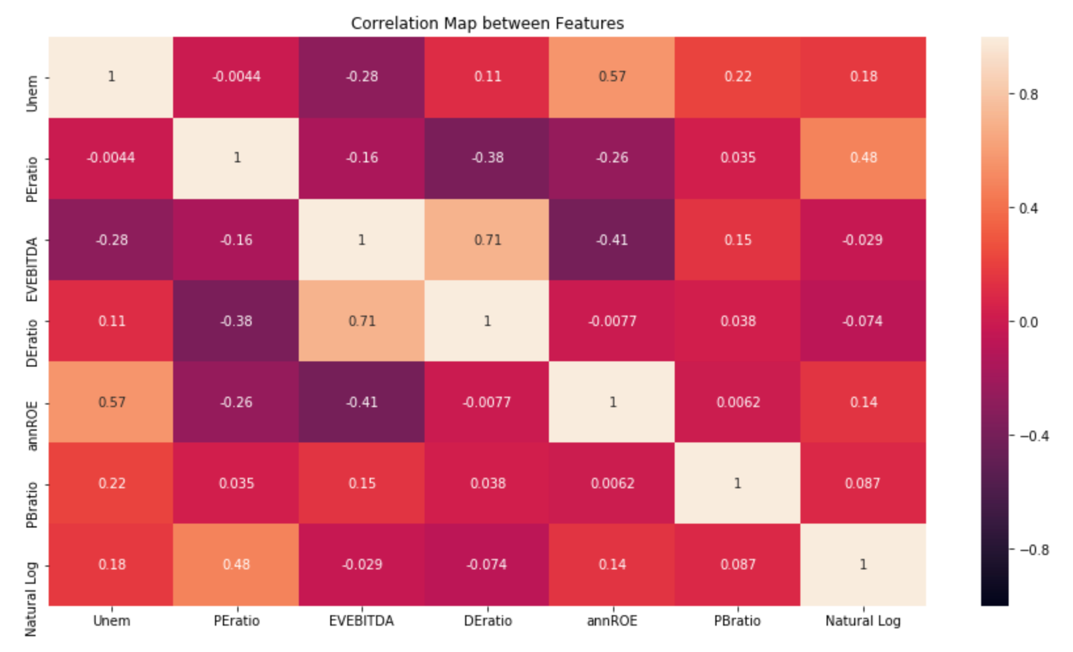
The debt-to-equity (D/E) ratio is calculated by dividing a company’s total liabilities by its shareholder equity. These numbers are available on the balance sheet of a company’s financial statements. The D/E ratio can evaluate a company’s financial leverage. The D/E ratio is an important metric used in corporate finance. It is a measurement of the degree to which a company is financing its operations through debt versus whole-owned funds. More specifically, it reflects the ability of shareholder equity to cover all outstanding debts in the event of a business downturn.

The annual return on equity (ROE) is a measurement of a company’s annual return (net income) divided by the value of its total shareholder’s equity. It is a target that can bring together the income statement and the balance sheet. The number represents the total return on equity capital and shoes the firm’s ability to turn equity investment to profits. Thus, I consider it as one of my features that can affect the log return of Netflix’s price.

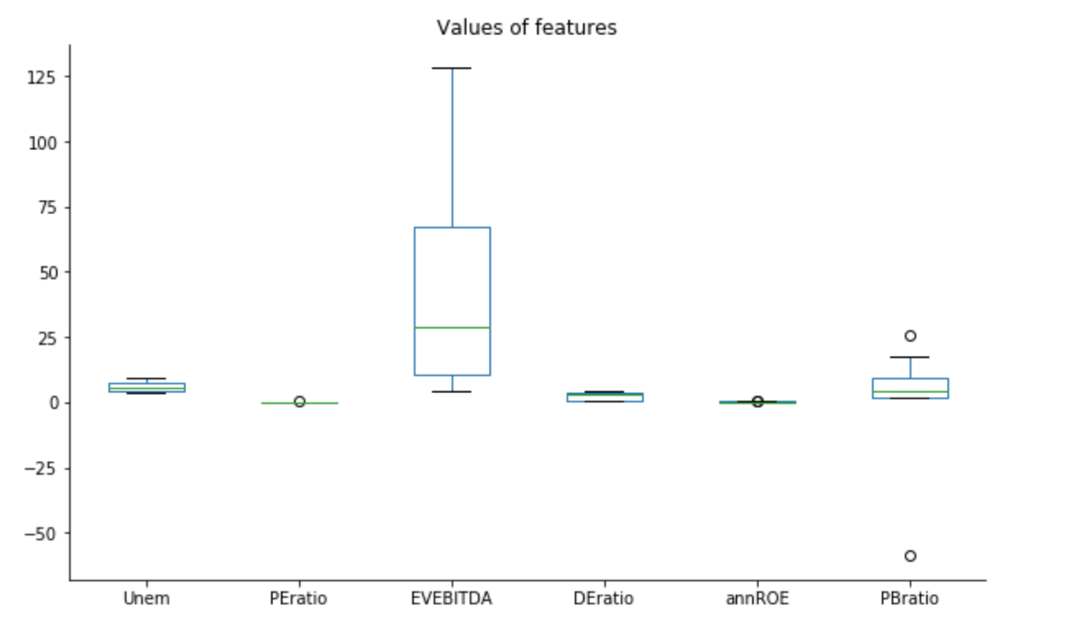
The price-to-book ratio (PB ratio) reflects the value that market participants attach to a company's equity relative to the book value of its equity. A stock's market value is a forward-looking metric that reflects a company’s future cash flows. The book value of equity is an accounting measure based on the historic cost principle and reflects past issuances of equity, augmented by any profits or losses, and reduced by dividends and share buybacks.

1. **Data Analysis and Insight Generation**

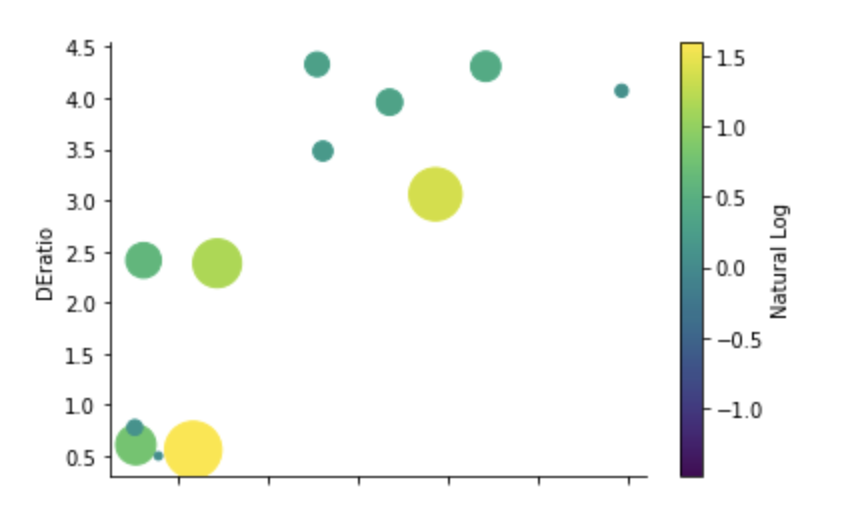
Features are not normally symmetric distributed; however, it depends on the few dataset I have. According to the correlation matrix, the P/E ratio has the highest correlation 0.48 with the log return but the unemployment rate has only 0.18 correlation. On the other hand, EVEBITDA and D/E ratio have the negative correlation with the log return.



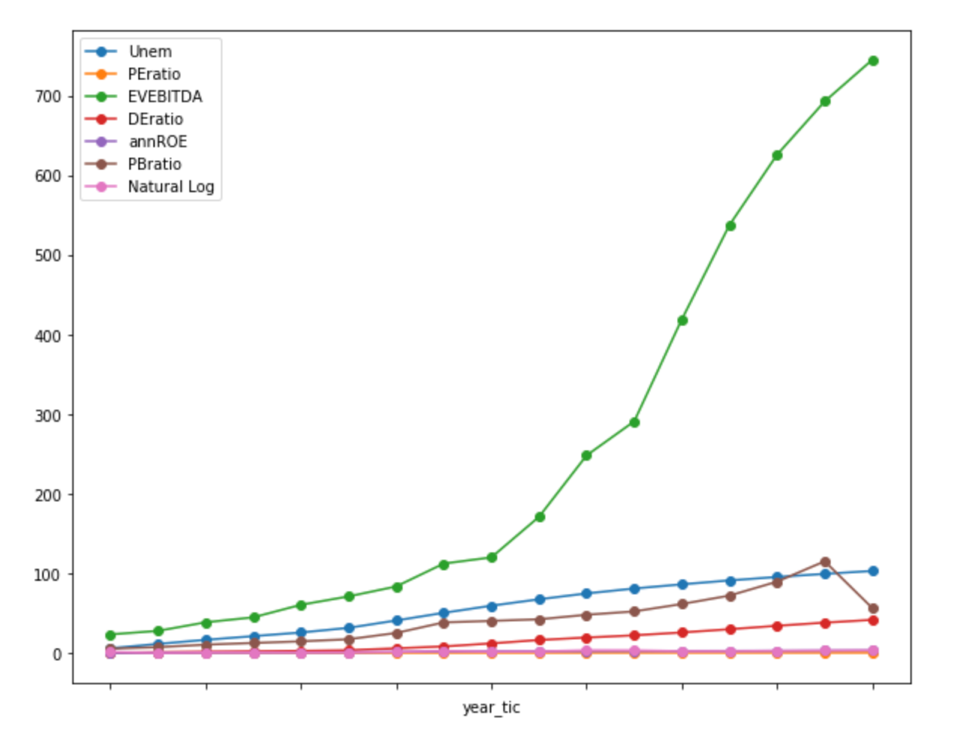
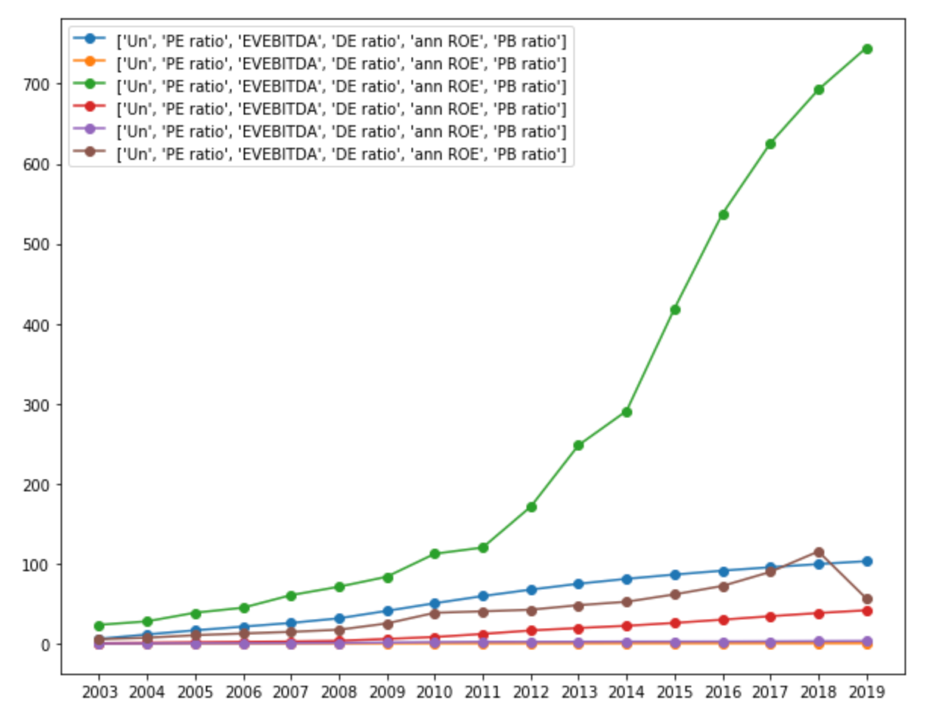
As we can see from the box plot, the range of EVEBITDA is really large as others have range from -1 to 25. P/B ratio have an extreme low value.



Let’s take a look at the scatter plot of EVEBITDA and DE ratio with the dot size is log return values. We can see that while EVEBITDA and DE ratio are low, the stock return is high, but when both values are high, stock return has no relationship with it.



Let’s see how these values from 2003 to 2019. We can see that the EVEBITDA keep going up and never falls but PB ratio felt from 2018 to 2019.



1. **Train and tune the model**

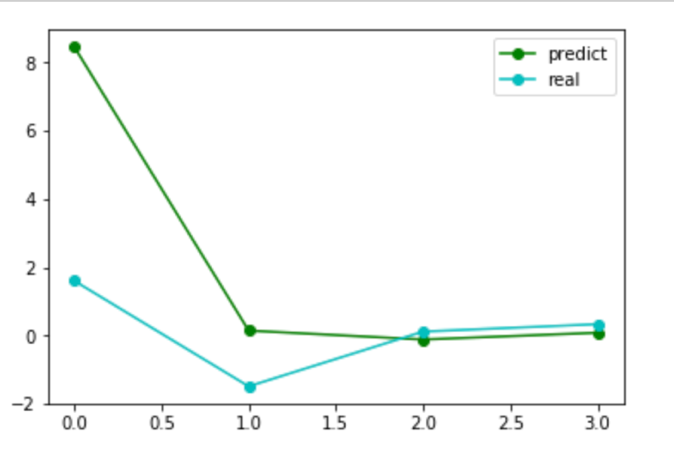
Next, I am going to predict the model with dependent variable Y with 6 independent variables Xs: X1=unemployment rate, X2=price-to-earnings ratio, X3=EVEBITDA, X4= debt-to-equity ratio, X5=annual return on equity, X6=price-to-book ratio. This is what one row represents.

I split the data to 80% training set and 20% test set. Training set is for developing models and feature sets and test set is for estimating the final results from the training set.

1. **Evaluate Performance**
2. Multiple Regression model:

With this model, I have the R-squared for negative and rooted mean squared error 3.53,the R-squared cannot be considered and the mean error is too large. Seeing the result of the prediction, the differences are low except the first value.

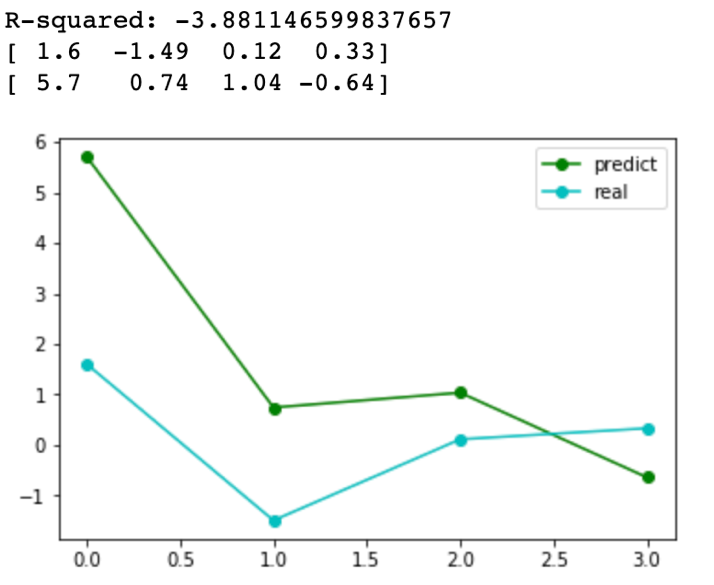
一張含有 桌 的圖片

自動產生的描述 

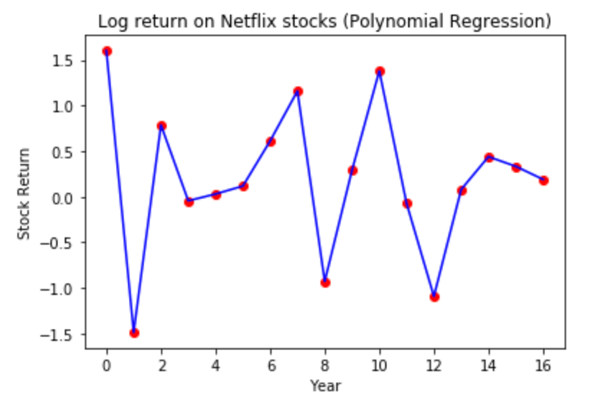
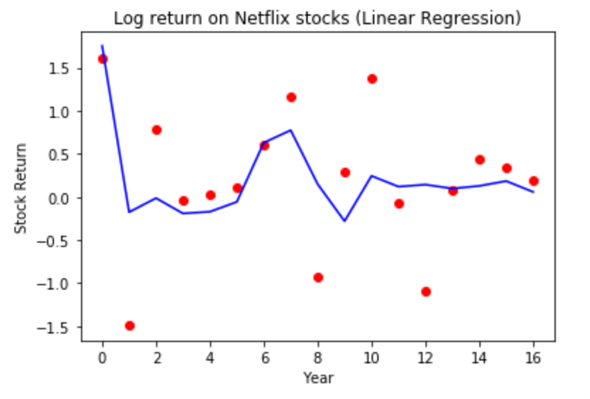
1. Polynomial Regression model

With this mode, the R squared has the lowest value in degree 2, but -3.88 is still a negative value, and the rooted mean squared error is 2.43. So this model cannot be used to my dataset. We can see that it doesn’t even have the right trend compared to the real test data.

一張含有 桌 的圖片

自動產生的描述

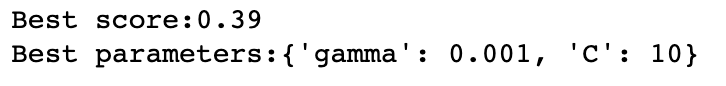
Visualizing the whole dataset by multiple regression model and polynomial regression model.



1. Support Vector Regression (SVR)

The SVR model is well known in classification problems. Also, it gives the flexibility to define how much error is acceptable in our model and will find an appropriate line to fit the data. Thus, I think I can give my data a try for using Support Vector Regression model.

First, I tried different gamma and penalty values(C), and I found out that when gamma is 0.001 and C is 10, the model has the best performance. Second, I tried 3 different kernel to see which one optimize the R squared value. If kernel is linear, R squared is 96%; -9% for kernel is polynomial and 63% for kernel is rbf. Therefore, the highest R squared is 96.9% and rooted mean squared error is 0.13 with linear kernel.

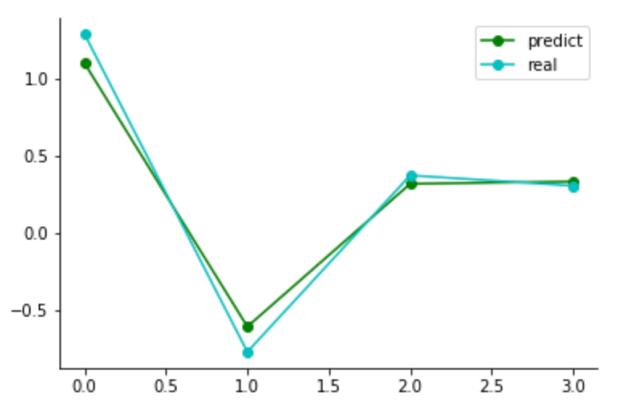


一張含有 文字, 植物, 鳥, 水鳥 的圖片

自動產生的描述一張含有 文字 的圖片

自動產生的描述

一張含有 桌 的圖片

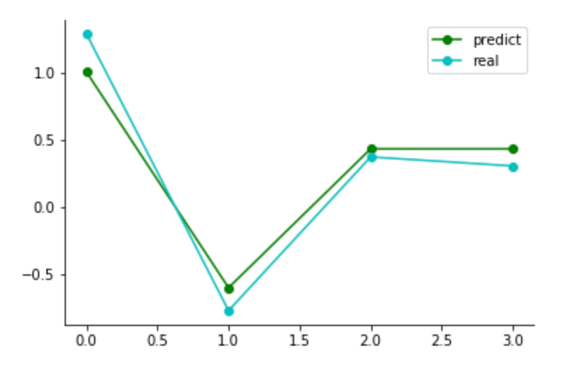
自動產生的描述

1. Decision Tree model

Decision Tree regression model is a model with tree structure. It breaks down a dataset into a smaller subset while an associated decision tree is incrementally developed at the same time. The final result is a tree with decision nodes and leaf nodes, so how you set the parameters will affect the performance of your model. Thus, I found out that when the maximum depth is 2, the minimum leaf is 1 and the minimum split is 2, the decision tree model will optimized the result with the R squared of 93.94% and the rooted mean squared error is 0.18.



一張含有 桌 的圖片

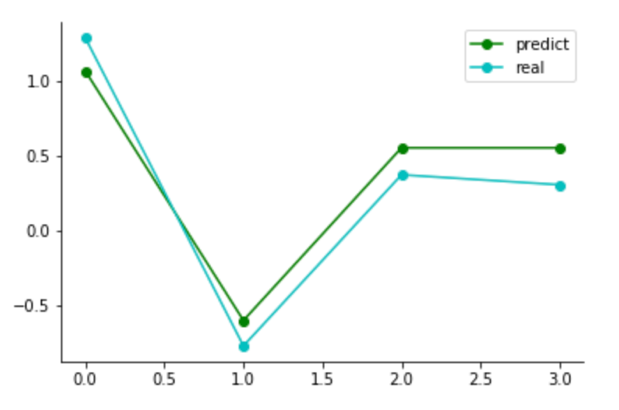
自動產生的描述

1. Random Forest model

The random forest model is an ensemble model which outperform many decision trees. The output is the mean prediction of the individual trees. As I tried many different parameters, the optimized performance is when the maximum depth is 2, the minimum leaf is 1, the minimum split is 2, the estimator is 1 and the random state is 1. This gives the best R squared of 91.93% and the rooted mean squared error is 0.21.



一張含有 桌 的圖片

自動產生的描述

1. **Use the model**

As the results of 5 models above, we consider SVR, Decision Tree and Random Forest models. For tree models, considering the performances of mean error is better than R-squared because it measures how well the estimator measures a parameter and the optimized tree will choose the smallest MSE. Therefore, Support Vector Regression (SVR) is the best appropriate model for the dataset predicting the future Netflix stock return with unemployment rate, P/E ratio, EV / EBITDA, D/E ratio, annual ROE and PB ratio.