Machine Learning Engineer Nanodegree

Capstone Project

Yudai Furukawa
May 7th, 2017

I. Definition

Project Overview

This project is about stock investing, and I am focusing on price prediction of a stock market index. A stock index is an aggregate value produced by combining several stocks, and it helps investors to measure and compare values of the stock markets such as in the US and Japan.
The Dow Jones Industrial Average (DJIA), Nasdaq Composite index and the S&P Composite are examples of stock indices.
As a wealth of information such as price, earnings, dividends, and CPI are available, I am going to use those information to do the prediction.
A dataset of S&P Composite published by Yale Department of Economics will be used in this project. For more infomation, please refer the link below:
https://www.quandl.com/data/YALE-Yale-Department-of-Economics (https://www.quandl.com/data/YALE-Yale-Department-of-Economics)

Problem Statement

For this project, the task is to build a stock index price predictor. A 12 month forward price change of S&P composite will be predicted by using regression. The project is going to be a supervised learning.

Following steps will be taken to make the predictor.
1. Figuring out necessary inputs to predict a 12 month forward price change by using

correlation between each feature and 12 month forward price changes.
* For example, earning growth and PER will be major inputs as those are considered leading indicators to predict a stock price.
2. Figuring out the best regression by trying a multiple regressions.
* Coefficient of determination will be used to measure performances of regressions.

Metrics

The coefficient of determination (the r2 score) will be used for scoring the result of the prediction. The r2 score provides a measure of how well the regression line represents the data. However, it has a weakness as the score could be greatly affected by unusual data points. Since the problem of this project is regression, the r2 score will be sufficient for this project as long as outliers are omitted.

II. Analysis

Data Exploration

A dataset of S&P Composite published by Yale Department of Economics will be used in this project.
The dataset (named snp in this project) is monthly time series of S&P Composite Price, Dividend, Earnigns, CPI, Long Interest Rate, Real S&P Composite Price, Real Dicidend, Real Earnings, and Cyclically Adjusted PE Ration since 1831-1-31 up to date.

For more infomation, please refer the link below:
https://www.quandl.com/data/YALE-Yale-Department-of-Economics (https://www.quandl.com/data/YALE-Yale-Department-of-Economics)

As of 2017-04-08, the basic statistics of snp the dataset is following.

Statistics	S&P Composite	Dividend	Earnings	CPI	Long Interes Rate
count	1756.000000	1755.000000	1749.000000	1756.000000	1756.000
mean	242.537415	5.344903	12.046968	56.433670	4.584025

Statistics	S&P Composite	Dividend	Earnings	CPI	Long Interes Rate
std	478.579184	9.010165	22.474642	69.298402	2.290630
min	2.730000	0.180000	0.160000	6.279613	1.500000
25%	7.680000	NaN	NaN	10.100000	NaN
50%	16.005000	NaN	NaN	18.100000	3.870000
75%	115.550000	NaN	NaN	84.200000	5.240000
max	2357.000000	46.380000	105.960000	244.176000	15.32000

As you can see some of the cells are filled by NaN as some data are missing in the dataset. Also, because when dealing with economical data, inflation has to be carefully taken into account as CPI tends to grow overtime and values of price and earnings tend to have smaller values in the past. Therefore, only real values, Long Interest Rate, and Cyclicall Adjusted PE Ratio in the previous table can be taken seriously in statistical analysis without any modification.

Exploratory Visualization

Figure 1: Time Series for all factors

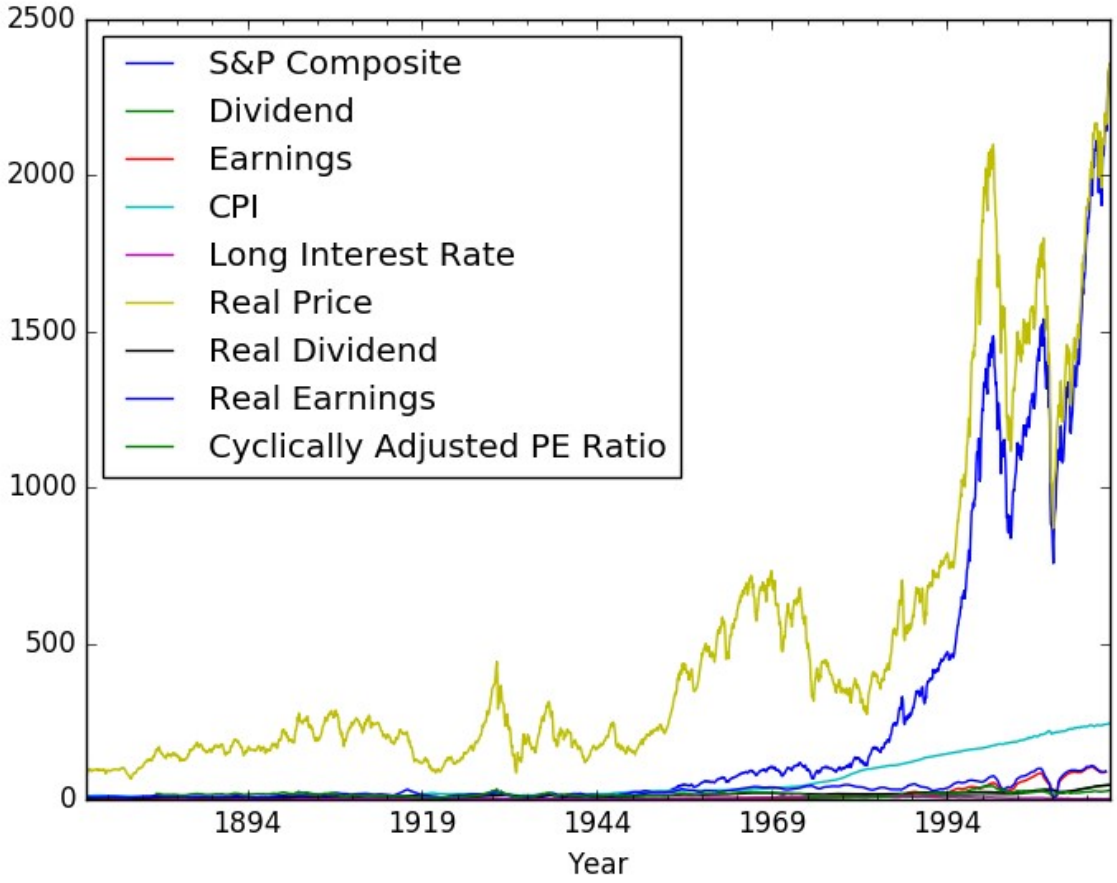


Figure 2: Time Series for Factors Except “Real Price” and “S&P Composite”

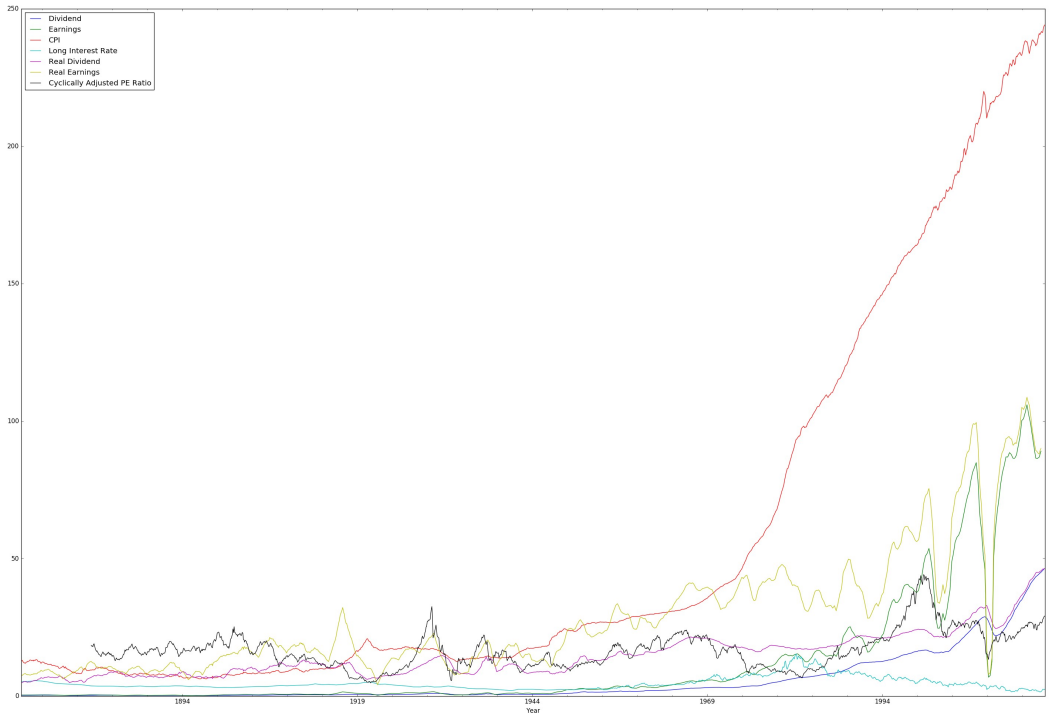


Figure 1 is a time series plot of the raw dataset. Figure 2 is also time series plot omitting “Real Price” and “S&P Composite” in order to visualize other features clearly. Time series plot was chosen as the data itself is a time series. For the both figures, the horizontal axis displays years and the vertical axis displays figures for each feature where the unit is different for each feature.

In figure 1, you can clearly see both “Real Price” and “S&P Composite” have been increasing in general as time goes. Therefore, it would be better nomarizing data when doing analysis. Also, in figure 2, you can see the same trends besides for “Cyclically adjusted PE Ratio” and “Long Interest Rate”. Therefore, for the same reasons, it would be better nomarizing data when doing analysis besides those two features.

Algorithms and Techniques

In this project, following regressions will be used in predicting the 12 month forward price change in S&P Composite.

1. Linear Regression
2. KNeibors
3. SVR

I chose the linear regression as it is the most commonly used regression in the financial industry. However, the weakness of the linear regression such as sensitivity to outliers and assumption of data independency should be carefully treated. I chose KNeibors as it is robust to noisy training data and the dataset is noisy as a lot of other features that might affect the price are missing. I chose SVR as it can do both linear and non-linear regressions and it is less likely to overfit.

Benchmark

r2 score of 0.5 will be used as a benchmark. 0.5 is a reasonable benchmark as equity return is often said unpredictable

III. Methodology

Data Preprocessing

As discussed, the original dataset needs to be nomalized in order to predict % change of

S&P Composite price with a return horizon of 1 year (named snp_changes). I first removed the columns for real values as this project aims to predict change of nominal S&P Composite price change and CPI in the dataset can be used to calculate the real value when needed. After this procedure, instead of using raw dataset, 12 months changes for each feature and the original figures of “Cyclically adjusted PE Ration” and “Long Interest Rate” will be used. The target feature will be 12 month forward changes in S&P Composite price.

The following modifications were made on the dataset.

1. Removed features with real values as the project is focusing on predicting % change of nominal S&P Composite price.
2. Generated a dataset called snp_changes which shows 1 year change of each feature in order to see the relationship of changes of 1 yeare S&P Composite price and other features.
3. Added a target feature “y” which is 12 months forward return of S&P Composite.
4. Added following features that seem to be have an effect on the prediction.
    - The real value of PE Ratio as the ratio is considered as a good indicator on predicting return in general.
5. Removed outliers
    - Outliers are omitted to avoid having a misleading r2 score - it has a weakness as the score could be greatly affected by unusual data points

Implementation

Cross validation is used when fitting the algorithms to the dataset. Cross validation is used in order to estimate how accurately a predictive model will perform in practice.

Refinement

At the beginning I used algorithms with default settings to determine which algorithm best serves the purpose of this project. After determining the algorithm, I used grid search in order to optimize the algorithms in this project.

IV. Results

Model Evaluation and Validation

r2 score of each algorithm with default settings are shown below.

1. Linear regression: 0.91077659608990325
2. KNeibors: 0.76730206091093156
3. SVR: 0.85812976342580705

As linear regression has the highest r2 score, I concluded linear regression best serves the purpose of this project. After using grid search on linear regression algorithm, the r2 score changed to 0.91077659608990336 which is almost the same to the score with default settings. Parameters after grid seach are followings {'copy_X': True, 'normalize': True, 'fit_intercept': True} which shows the parameter 'normalize' changed from False to True. It is reasonable that the linear regression is the best estimator as this is the most commonly used model in finance. The prediction for some samples also shows that the model is working well.

Comparison of Samples' y and Predicted y

Index	Acctual y	Predicted y
164	-0.043763676148796504	-0.05016142
333	0.26212319790301453	0.27635389
1000	0.13199406768235122	0.1581155

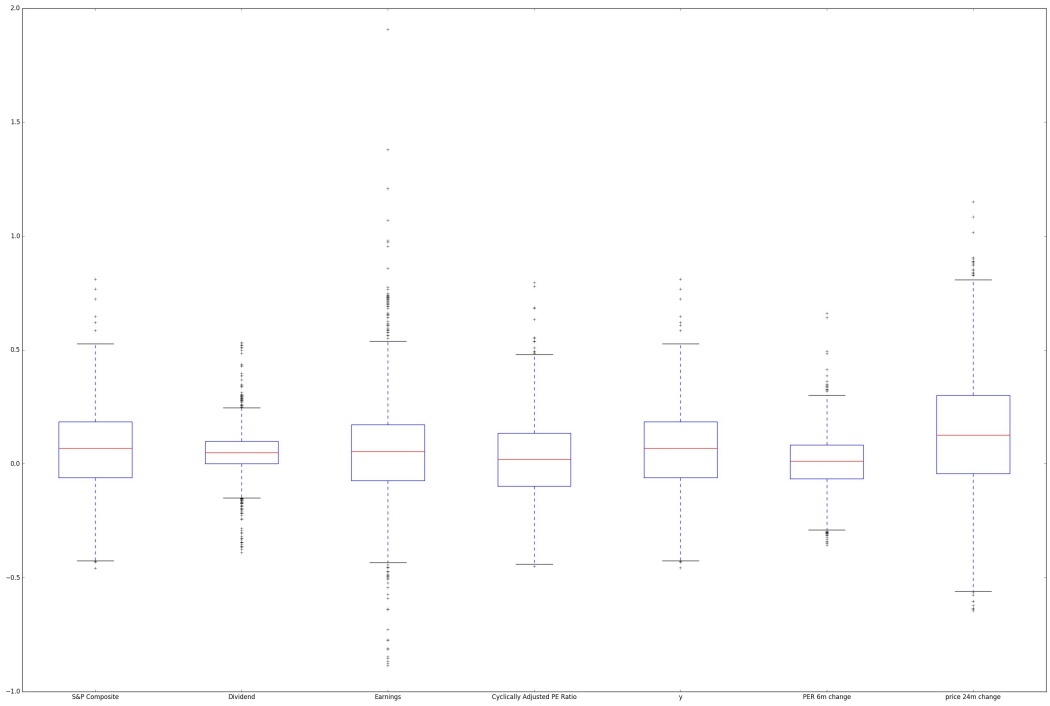
Justification

r2 score of 0.91 with the selected linear regression is definitely higher than benchmark which is r2 score of 0.5. The result is reasonable as linear regresison is historically working well in economics.This result is significant as the project shows existing factors can be good predictors of the future S&P Composite return although it is often said that the it is nearly impossible to predict the future return of S&P Composite.

V. Conclusion

Free-Form Visualization

Figure 3: % change in each factor (24 month % change for 'price 24m change'. 12 months % change for other factors)



This graph is a box plot of the dataset. As you can see mean values of all the features are abobe 0. Especially as the mean value of y is above 0, you can make money on average when you invest in S&P Composite.

In this section, you will need to provide some form of visualization that emphasizes an important quality about the project. It is much more free-form, but should reasonably support a significant result or characteristic about the problem that you want to discuss. Questions to ask yourself when writing this section:
- *Have you visualized a relevant or important quality about the problem, dataset, input data, or results?*
- *Is the visualization thoroughly analyzed and discussed?*
- *If a plot is provided, are the axes, title, and datum clearly defined?*

Reflection

I started with definition the goal, searching datasets I can use, analyzing the characteristics of the dataset, and processing the data in order to have it ready for regression. After having the data ready, I decided how to measure the result, chose regressions that would be useful with taking strength and weakness for each regression in mind, and figured out which regression will best serve the purpose.

I found it interesting that I could define the problem as any of supurvised, unsupervised, and reinforcement learning although the data being used is the same. I found it difficult to find data that I need for the project as most of datasets are not free.

The final result fit my expectations for the problem as the linear regression is the most commonly used regression in the financial industry, and I think

Improvement

I think I could have more factors such as fandamental data (book value, working capital etc.) and economic data (GDP, NFP, etc.) in the dataset in order to make a better stock price predictor. With more factors in the dataset, I expect to have a better regression with higher r2 score.