Capstone Proposal

Nobuyoshi Shimmen
April 10th, 2017

Domain Background

There are mainly two ways to evaluate stock market: technical analysis and fundamental analysis. Technical analysis assumes that a stock price reflects all the information that decides the price available to public and predicts how a stock moves by statically looking at the history of the price movements. Fundamental analysis, on the other hand, tries to see through intrinsic value of a company by analyzing corporate financial statements, like balance sheet and income statement. They are used to see if a stock is undervalued or not. Technical analysis is said to be effective for predicting short-term trading trends because there is less uncertain elements than in the long term future. Fundamental analysis show its value in long term stock investment because you need time to wait before a company's true value is recognized in the market.

Hedge funds and institutional investors have been focusing on technical analysis because of the nature of the business they are in. In short period of time, they have to leverage their customers' money and make some profit for the customers. Also they do this for themselves because their salaries are greatly affected by how much profit they make. So they have been putting effort and money into applying machine learning to technical analysis for a long time.

Retail investors prefer fundamental analysis to technical analysis, because they simply cannot afford so much transaction cost for buying and selling stocks that technical analysis requires to make capital gains in short period time. Also, fundamental analysis has been proven effective and recommended by many famous investors like Warren Buffet. Interestingly and sadly, however, even though fundamental analysis have been seen as practical and helpful to individual investors, little research hasn't been done for applying machine learning to fundamental analysis.

Therefore, in this project, my focus is on fundamental analysis and I would like to see how machine learning can predict winning stocks in the long run by using fundamentals of companies. One of the academic research relevant for this project is done by Michael Dickens(2015)[1]. He shows us fundamental analysis using support vector machines could overperform simple fundamental analysis. His findings opened up new possibilities; machine learning applied fundamental analysis have a greater chance of picking winning stocks than conventional one.

Problem Statement

Fundamental analysis is suited for individual investors by nature but almost no one has tried to apply machine learning to this. My goal is to make a model that can pick winning stocks in the long run and outperform the stock market. What I mean by the long run is one year from the time prediction occurs. The model predicts prices of a stock in one year based on fundamentals of companies.

Datasets and Inputs

I need two kinds of datasets to make a model; one is fundamentals and the other is historical stock prices. These datasets are obtained from kaggle (https://www.kaggle.com/dgawlik/nyse) and Yahoo! Finance.

The features are corporate financial statements of 2012-2015, extracted from fundamentals.csv on the kaggle page. The examples of the features are Earnings Before Interest and Tax, Gross Profit, Net Income, Total Liabilities, and so on. I will also some indicators mentioned in a book called stock investing for dummies[2] as features. For instance, PE ratio (price to earning ratio) is mentioned in the book and thought of as an important indicator representing how overvalued/undervalued a stock is, but is not directly available from financial statements. So you have to calculate that type of indicators by using their fundamentals. In the case of PE ratio, you divide the price of a stock by its earning per share.

The labels are extracted from the prices-split-adjusted.csv on the kaggle page and some of them are from Yahoo!Finance. I will use the stock price of December 31 of a following year as a label. For example, when I use the fundamental data of a company of 2012 as features, a label is the average price of the stock on December 31 of 2013.

Solution Statement

The Solution is simple. I will make a model that makes a prediction for the price of a stock in next year. I make three models and each model uses a different algorithms. Here' a list of algorithms to implement with the reasons I chose them.

1. Linear Regression (Implement through TensorFlow)
    a. It's the most common and basic algorithm when it comes to regression problem
    b. It can be combined with neural network

2. Decision Tree (Implement through sklearn)
    a. It is easy to understand and to interpret how a result is decided
    b. It is robust to change and outliers
    c. It performs well with small amount of data points

3. Neural network (Implement through TensorFlow)
    a. Even though the logic behind the result it produces is almost impossible to figure out, people say it works well

I will use along the following techniques with these algorithms

1. Cross Validation

   It is a technique to see how well a model predict unseen data by splitting an original sample dataset into a training set to train the model, and a test set to evaluate

2. Grid Search

   It is a technique to find parameters that perform best for a model. It can work with Cross Validation

Benchmark Model

The benchmark is the average growth rate of return in stock market in general. This is because fundamental analysis is not a method but a concept of using fundamentals for evaluating stocks. The method of fundamental analysis varies depending on the person you ask. Therefore, there is no one best way to do that and I will use the average returns of stock investment. I will base it on Standard & Poor's 500 Index, which has been seen as a reflection of the performance of American stocks in general. The data is obtained from this website (http://people.stern.nyu.edu/adamodar/New_Home_Page/datafile/histretSP.html). The owner of this website keeps track of the annual growth rate of Standard & Poor's 500 Index.

Evaluation Metrics

In this project, there are two types of evaluations. One is used for evaluating the performance of a model, so that I can choose the best one. I use coefficient of determination, which is often called R2. This metric is useful in regression analysis, because it describes how good a model is at making predictions. The maximum score of it is 1 and the minimum score of it is 0 and the greater the r2 score is, the better the model is fitting.

$$R^2 \equiv 1 - \frac{SS_{\rm res}}{SS_{\rm tot}}.$$

Image from wikipedia

The other metric is for comparing the performance of a model to benchmark. I use compound annual growth rate (CAGR) for this. Many stock investors and analysts use GAGR to evaluate the performance of investments. The definition is as follows.

$$\text{CAGR} = \left(\frac{\text{Ending Value}}{\text{Beginning Value}}\right)^{\left(\frac{1}{\text{\# of years}}\right)} - 1$$

Image from INVESTOPEDIA

I will have the best model predict prices of all the stocks in one year. Then I will calculate the CAGR of 30 stocks that would get the highest growth in price, using the sum of original price of stocks as the beginning value and the sum of the next year price* as the ending value. Finally I will compare the result to the CAGR of return on the market.

*These are not what are predicted by the model, but are the actual prices of the following year.

Project Design

This project will be implemented in a Jupyter Notebook so that anyone interested in this problem can reuse my work. I will proceed with this project as follows.

1. Setting up infrastructure
    a. Create a git remote repository and a Jupyter Notebook in my laptop
    b. Install numpy, pandas, matplotlib, scikit-learn and TensorFlow through Anaconda.
    c. Set up a Capstone report template.
2. Project Overview
    a. Summarize what I wrote in Domain Background
3. Problem Statement & Metrics
    a. They will be the same as what I wrote in corresponding sections of this paper
4. Data Exploration
    a. Process the dataset into Pandas Dataframe
    b. Get descriptive statistics of labels for one year
        i. Min value, Max value, Mean value, Median value and Standard deviation of values
    c. Get descriptive statistics of features for one year
        i. Check for the same things as labels
        ii. See if there is any feature that has highly-skewed distribution
    d. Describe the characteristics of each feature from a perspective of investors(common sense)
        i. I might use them later to justify the result of this project
        ii. What feature is said to make the price of a stock goes up?
5. Exploratory Visualization
    a. Visualize highly-skewed features based on what I find on 4-c.
6. Algorithms and Techniques
    a. This will be the same as what I wrote in Solution Statement of this paper
7. Benchmark
    a. This will be the same as what I wrote in Benchmark of this paper
8. Data Preprocessing

   a. Transforming some skewed features
      i. Some features like Research & Development highly skewed, so apply a logarithmic transformation to them
   b. Remove rows and columns that have missing values or substitute something for the values
   c. Normalize features by using sklearn.preprocessing.MinMaxScaler

9. Implementation
   a. Use scikit-learn to construct Decision Tree models and TensorFlow to construct Linear Regression and Neural Network.
   b. Train and test each model
      i. Split dataset with sklearn.model_selection.TimeSeriesSplit
      ii. Use sklearn.grid_search.GridSearchCV to find the best performing parameters except for Neural Network model
   c. Choose the best performing to proceed with
10. Refinement
   a. It depends on the model performing best at previous step
   b. If Linear Regression / Neural Network model is best, I will combine it with neural network / linear regression
   c. If Decision Tree model works best, I will apply Ensemble method
11. Model Evaluation and Validation
   a. Have the model make price predictions with a group of split dataset from 9-b-i
   b. Calculate the CAGR of 30 stocks that would get the highest growth in price, using the sum of original price of stocks as the beginning value and the sum of the next year price* as the ending value
   c. Compare the result to the CAGR of return on the market.
12. Justification
   a. I will justify my result using what I write in Data exploration
13. Free-Form Visualization
   a. I'm still thinking of what to do with this
14. Reflection
15. Improvement

Reference

[1] Dickens, Michael. "Examining Long-Term Trends in Company Fundamentals Data." Stanford (2015). Web.
[2] https://www.amazon.com/Investing-Dummies-Fifth-Eric-Tyson/dp/B0031569MO