Capstone Project

Jian (Kevin) Wang

Machine Learning Engineer Nanodegree

February 24, 2017

This is a short summary version updated on March 3, 2017

Predicting long-term stock gain using machine learning methods

Definition

Project Overview

This capstone project was inspired by my personal interest and passion in long term investing in the stock market, as well as Udacity's suggested problem on Investment and Trading Investment for the capstone project. The goal of this project is to build a classification model that predicts whether or not the stock price will rise at least 10% in one year for stocks in the consumer staples sector. Unlike other approaches focusing only on the stock historic pricing data for prediction of future price, herein I utilized both the stock historic pricing and fundamental data for predicting the long-term (one year) stock price trend.

With stock market index such as the S&P 500 index reaching new highs, the S&P 500 is expected to increase only 4% by the end of this year [1]. John Bogle, the founder of mutual fund giant Vanguard, recently predicted that US stocks will produce only 4% annualized return in the next decade [2]. Recently, Milosevic reported a machine learning based approach using various algorithms to evaluate whether or not a stock's price will increase by 10% in one year [3]. Dai and Zhang used various machine-learning models to forecast the price trend of a single stock using various fundamental data of the stock [4]. In addition, Madge and Bhatt reported using Support Vector Machines to predict the price direction of technology stocks [5].

As an individual investor passionate about learning stock market investing, I believe certain individual stocks may produce superior returns than the others based on stock's fundamental data. Given the abovementioned prediction of US stock market return, a 10% gain on individual stocks will be a superior investment return. In addition, long term gain (one year or more) will result in more favorable tax rates for individual investors. This motivated the work for this capstone project: predicting whether or not a stock will gain at least 10% in one year. For the capstone project, I limited the study of stocks to the consumer staples sector that I believe will have similar characteristics and stock price movements. I am also personally interested in the sector of consumer staples given the relatively less complexity of the business model, slow and steady growth, and good diversification advantage with its relatively low correlation with the overall stock market [6]. The

stock fundamental data are available from the SEC's SEC's EDGAR database [9] and stock's historic pricing data are available from Yahoo Finance [11]. The target variable to predict can be determined by whether or not the stock's price rises at least 10% in one year.

Problem Statement

I used supervised learning methods to predict whether or not a stock will rise 10% or more in one year, given the historical stock data, including EPS, P/E ratio, Dividend, Adjusted Closing Price, Market Cap, and Current Ratio (Current Assets / Current Liability), for stocks in the consumer staples sector. For each of the stocks, quarterly financial data from company 10-Q reports as well as the stock pricing data from the end of each quarters were collected and used to predict a binary class variable: 1 if the stock will rise at least 10% in one year and 0 otherwise. All input data are of numerical type and the target variable to predict is binary. The data set will be collected from 2009 until 2015 such that the target variable based on stock prices will be known for both the training and test sets. The performance of the classification model was evaluated by overall accuracy and other metric as discussed in the later section. The classification model was built using six supervised learning methods, including Decision Tree, Naïve Bayes, Logistic Regression, K Nearest Neighbor (KNN), Random Forest and Support Vector Machines (SVM).

The strategy to solve this problem is outlined in these overall steps:

1. Data acquisition: Acquire stock fundamental and historic pricing data
2. Data preprocessing: Explore the data and cleanse redundant and undesirable data. Transform the data by using scaling and normalization techniques.
3. Create training and testing set by shuffling and splitting the data.
4. Build classification models and evaluate model performance.
5. Identify the most important features for classification.
6. Evaluate performance using only the most important features.

The performance of the classification models were compared against a naïve predictor, which always predict the stock price will rise at least 10% in one year (target variable = 1 for all cases), using overall accuracy and other relevant metric. It was anticipated certain classification models would clearly outperform the naïve predictor based on the defined metrics.

Disclaimer

The methodologies, approaches, opinions and any other information presented in this project report are exclusively for the purpose of completing the Capstone Project for the Udacity Machine Learning Nanodegree. Any information presented in this report must not be regarded as advice on trading and investing strategies or on

the financial markets. Do not use any information presented in this project to make trading or investing decisions.

References:

1. http://www.marketwatch.com/story/sp-500-expected-to-rise-only-4-by-end-of-2017-2016-12-24

2. http://www.marketwatch.com/story/john-bogle-says-you-wont-make-much-money-from-stocks-2015-11-05

3. Milosevic. Equity forecast: Predicting long-term stock price movement using machine learning. 2016. https://arxiv.org/pdf/1603.00751.pdf

4. Dai & Zhang. Machine Learning in Stock Price Trend Forecasting. http://cs229.stanford.edu/proj2013/DaiZhang-MachineLearningInStockPriceTrendForecasting.pdf

5. Madge and Bhatt. Predicting Stock Price Direction using Support Vector Machines. 2015. https://www.cs.princeton.edu/sites/default/files/uploads/saahil_madge.pdf

6. Schmidt. A Guide To Investing In Consumer Staples. http://www.investopedia.com/articles/economics/08/consumer-staples.asp

7. Global Industry Classification Standard (GICS) https://www.msci.com/gics

8. https://en.wikipedia.org/wiki/List_of_S%26P_500_companies

9. SEC's EDGAR database https://www.sec.gov/edgar/searchedgar/companysearch.html

10. pystock-crawler 0.8.2 https://pypi.python.org/pypi/pystock-crawler

11. Yahoo Finance https://finance.yahoo.com

12. https://pypi.python.org/pypi/yahoo-finance/1.1.4

13.http://scikit-learn.org/stable/modules/generated/sklearn.metrics.precision_score.html

14.http://scikit-learn.org/stable/modules/generated/sklearn.metrics.recall_score.html

15. https://docs.scipy.org

16. https://en.wikipedia.org/wiki/Decision_tree_learning

17. Andrew W. Moore
https://pdfs.semanticscholar.org/4049/1dc18f32292d56508729e4d66738999a1542.pdf

18. https://www.quora.com/What-is-logistic-regression.

19. https://en.wikipedia.org/wiki/Logistic_regression

20. Stephen G. Powell, Kenneth R. Baker, Management Science. Chapter 6: Classification and Prediction Methods. PowerPoint slides from: faculty.tuck.dartmouth.edu/images/uploads/faculty/management-science/Ch06.ppt

21. Lalit Sachan 2015, http://www.edvancer.in/logistic-regression-vs-decision-trees-vs-svm-part2/

22. http://classroom.synonym.com/disadvantages-logistic-regression-8574447.html

23. https://en.wikipedia.org/wiki/Bayes%27_theorem

24. https://en.wikipedia.org/wiki/Naive_Bayes_classifier

25. https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm

26. http://www.cs.upc.edu/~bejar/apren/docum/trans/03d-algind-knn-eng.pdf

27. https://en.wikipedia.org/wiki/Support_vector_machine

28. Rutgers University CS 536: Machine Learning Littman (Wu, TA)
http://www.cs.rutgers.edu/~mlittman/courses/ml04/svm.pdf

29. http://www.svms.org/disadvantages.html

30. Laura Auria, Rouslan A. Moro, Berlin, August 2008, Support Vector Machines (SVM) as a Technique for Solvency Analysis
https://core.ac.uk/download/pdf/6302770.pdf

31. https://en.wikipedia.org/wiki/Random_forest

32. http://www.math.usu.edu/adele/RandomForests/Ovronnaz.pdf

33. http://amateurdatascientist.blogspot.com/2012/01/random-forest-algorithm.html

34. Caruana, Rich; Karampatziakis, Nikos; Yessenalina, Ainur (2008). "An empirical evaluation of supervised learning in high dimensions".Proceedings of the 25th International Conference on Machine Learning (ICML).

35. Segal, Mark R. (April 14 2004). Machine Learning Benchmarks and Random Forest Regression. Center for Bioinformatics & Molecular Biostatistics.