Project Report: Exploring the Application of Machine Learning to Stock Investment

Hsiu-Hui (Shelly) Tseng
Affiliation: University of Washington - Applied mathematics
Address: 3635 Woodland Park Ave N, Seattle, WA, 98103
email: hhtseng@uw.edu

Abstract

In this project, unsupervised learning and supervised learning are applied to daily stock price of S&P 500 stock datasets from August 2006 to August 2013. By using the scientific-based methods, the main objective of this study is to decrease the investment risk and maximize the stock investment profit based on the machine learning techniques. First, principal component analysis (PCA), an unsupervised learning tool, is applied, which can help extract the major trends determining the variability of stocks and identify specific stocks that represent the whole 500 leading stocks groups. The results show that the largest variability is dominated by the negative co-variability between financial and high-technology stocks. In addition, it is found that the S&P 500 market moving trend could be well explained by a periodic cycle of major PCA components (80% of the total variance) during stable market period, and the abnormal cycle during 2007-2008 financial crisis could be captured by this analysis as well. Based on the analysis of PCA, a risk index of stock investment is proposed and discussed. Second objective of this study is to predict future daily moving trend of individual stocks by using supervised deep learning. By using deep learning, nine selected stocks daily stock trend (up or down, binary question) are trained and predicted. Our results suggest that test accuracy is varied among stocks, and could reach 60% in average, which implies that deep learning could be an useful reference and tool for stock trading strategy.

1 Introduction

Stock investment market is an extremely popular and exciting option for investor because of its high gains and its accompanying high risk embedded in the market. We are living in an era of big data, and stock market is certainty no exception. There are two major challenging in the field of stock investment strategy - risk management and trend prediction. The risk in stock market is strongly related with the stock price variability, namely that the market with high variability of stock price is often considered as an unstable market. It is important to noted that the goal of risk management is not to predict the absolute trend of the market, but is to reduce the risk based on the relationship among stocks derived from historical data. For example, if A stock price is strongly negative correlated with B stock price, then buying A and B stocks together could help reduce the investment risk even though the absolute trend is unknown or is challenged to predict (i.e., loss could be compensated by gain). Therefore, principal component analysis, a unsupervised learning technique, which can extract the major trends determining the variability among the stocks is an useful tool for the purpose of risk management and is applied in our study. On the other hand, trend prediction is a much more aggressive and ambitious objective of stock investment strategy, the goal of which is to predict the

Submitted to CSE 546 machine learning class project milestone on exploring the application of machine learning to stock investment (2017).

absolute trend of stock price (i.e., up or down), and thus maximize the investment profit. Thus, supervised deep learning is a suitable tool for the goal of trend prediction, and is used in our study. This report is structured in 4 sections. Section 1 is introduction. Section 2 describes the methods and data used in this study. Section 3 presents the results. Section 4 gives the summary.

2 Methodology

2.1 Data collection and data cleaning

Daily stock price of 7 years (from August 2006 to August 2013) of S&P 500 stocks from QuantQuote is used here, which included split/divided adjustments, symbol change tracking, and fairly straightforward format make this the easiest to use as a research dataset. Note that only the stocks that have full records of stock prices from August 2006 to August 2013 are selected. Note that part of the S&P 500 stocks are included after August 2006. Therefore, total of 467 out of 500 stocks are selected here.

2.2 Principal component analysis

Stocks market has a large number of variables, which results in huge challenge for identifying correlations between the stocks. For our purpose of identifying the covariance among the stocks and proposing an useful index of risk management, the dimension reduction technique, principal component analysis, is applied here. First, we select Apple, Google and AT&T as our first example to demonstrate the strength of PCA as an unsupervised exploratory tool for stock market. Then, for a bigger picture of S&P 500, only the stocks that have full records of stock prices from August 2006 to August 2013 are selected. Principal component is derived from the covariance matrix of stock prices, and PCA1 is defined as the principal component with largest eigenvalue of the covariance matrix, and so on. The percentage of variance of the stock price explained by the PCA could be derived by its eigenvalue divided by the sum of eigenvalues.

2.3 Deep learning

In our study, the supervised deep learning package provided by TensorFlow is used to explore the prediction accuracy of daily stock trend by deep learning technique. The training data ranges from Aug 2006 to Feb 2012 (80%), and the test data ranges from Mar 2012 to Aug 2013 (20%). Input features are all S&P 500 stock prices and daily stock change ratio at time T=t and t-1 (i.e., previous two days), and output is the stock price at time T=t+1. Our model includes 4 hidden layers with 1024, 512, 256 and 128 neurons at 1st, 2nd, 3rd and 4th layers, respectively.

3 Result

3.1 Risk management of stock investment by PCA

In this section, three selected stocks price of Apple, Google and AT&T are analyzed with principal component analysis first. Figure 1 shows the demeaned trend of the stock prices. There is an interesting question can be asked here. How much percentage of co-variability between Google, Apple and AT&T is positive? and how much is negative? Figure 2 shows that the major variability of three selected stocks is dominated by the positive correlation between Apple and Google, which could explain 88% of the variance. Interestingly, there is 11% of the variance is explained by the negative correlation between Google and Apple stock prices. Then, we apply the same analysis technique to the whole S&P 500 stocks, which are summarized in Figure 3. The result shows that the fraction of variance of S&P 500 stock price explained by the first principal component (PCA1) was 60% and the second principal component (PCA2) was 20%. PCA1 indicates that the major trend of daily stock price of S&P 500 is primarily controlled by the negative correlation between high technology stocks and financial stocks. On the other hand, PCA2 indicates the positive correlation between high technology stocks and financial stocks. With the PCA direction (i.e., eigenvector of the covariance matrix), the projection of S&P 500 stocks price could be derived, which is concluded in the upper panel of Fig. 4. Now, the historical stock market price could be re-constructed by the PCAs and their projection (called PC afterward). In addition, the change rate of the market (i.e. stock price) could be informed from the moving rate of stock price projection

on PCAs (see second panel of Fig. 4). A critical idea here is that the stocks prices changes are resulted from the PC change along each PCAs direction. In other words, if there is no changes on PC (projection), then stock market is constant. With these information, we propose that the return due to the change of PC1 (the projection of stock prices on PCA1 direction) could be written as:

$$IR_i = \frac{PCA_1}{S_{price}} P_{portfolio} R_{PCA_1} = Risk_i R_{PCA_1} \quad \text{..... (1)}$$

where IR_i is the investment return (d\$/dt) $\in R^1$, PCA_1 is the eigenvector of PCA1 (d\$/dPC) $\in R^{467}$ (i.e., 467 stocks are selected), S_{price} is the stock prices (\$) $\in R^{467}$, $P_{portfolio}$ is the portfolio for each stock investment combination (\$) $\in R^{467}$, R_{PCA_1} is the PC change rate (dPC/dt) $\in R^1$ (which is the moving rate of projection of stock price on PCA1), and $Risk_i$ is the investment risk due to the market variability caused by PCA1. Note that same formula could be applied to PCA2, PCA3 and so on. Therefore, based on the information of PCA, a simple risk index as a function of investment portfolio, stock price and PCA eigenvector is derived. The equation 1 above indicates that the risk due to certain PC (i.e., projection) is worth considered only when the change rate of projection R_{PCA_1} is large enough (second panel of Fig. 4).

From Eq. (1), it is known that R_{PCA_i} is the key parameter determining the sign of the return (i.e., gain or loss). The second panel of Fig. 4 suggests that there might be some periodic cycle/relationship among R_{PCA_i}, and might be useful for the investment strategy. Figure 5 shows that PCA scatter plot (color bar is the time) of the PC projection moving rate of stock prices from 2007 to 2013 (left panel) and from 2010 to 2013 (right panel) (i.e., PCA analysis on the bottom panel of Fig. 4), which showing that during the stable market after 2010, there is period cycle in the market controlled by PCA1 and PCA2 of PC moving rate of stock prices, and during the financial crisis of 2007 to 2008, there is abnormal cycle of R_{PCA_i}. The result here suggests that the metric R_{PCA_i} could potentially be a useful tool to determine the stock investment strategy and predict the future market trend. Figure 6 shows the PCA analysis on the stock trade volume historical data, which indicate that after the financial crisis of 2007 to 2008, the stock market is completely transformed from a financial-leading market to the market that is leaded by high technology companies (see bottom panels of Fig. 6).

3.2 Application of deep learning to trend prediction

In the application of deep learning to trend prediction, the input features are all selected 467 stock prices and daily stock change ratio at time T=t and t-1 (i.e., previous two days), and output (the goal) is the stock price at time T=t+1 (i.e. tomorrow's stock price). There are 9 stocks that are selected here for the prediction. Figure 7 shows the accuracy of training set and the testing set. As one can imagine, training accuracy could reach 95% for the selected stocks. However, the testing set accuracy could only reach 60% in average for the selected stocks. Figure 8 summarizes the accuracy for the test set, indicating that the testing accuracy of predicting up or down binary questions is varied among stocks, and could reach 60% in average. Given that the random guessing accuracy could reach 50% in average, the results from deep learning can beat random guessing by 10%, which could be a useful reference for the development of stock trading strategy.

4 Summary

There are two part of study here - risk management and trend prediction for stock markets. First, risk management is explored with unsupervised learning technique, principal component analysis. Risk can be considered as the variability of stock market, which can help or hurt profit (high variability, high risk). By using PCA, our study propose an useful index $Risk_i$, which can be written as a function of portfolio, stock prices and PCA direction. In addition, temporal variation of R_{PCA_1} (PC variation) in stable market after 2010 shows stable period cycle, which could be used a useful reference for stock investment strategy. Second is the daily trend prediction based on two days historical data by using deep learning. Nine selected stocks daily stock trend (i.e., up or down, binary question) are trained and predicted. Our results suggest that test accuracy is varied among stocks, and could reach 60% in average, which implies that predicting daily up and down trend by using deep learning is better than random guessing (50% accuracy, up or down), and is potentially useful for practice.

5 References

[1] JP Morgan (2017): Big Data and AI Strategies: Machine Learning and Alternative Data Approach to Investing.

[2] Hargreaves, C. A., & Mani, C. K. (2015): The Selection of Winning Stocks Using Principal Component Analysis. American Journal of Marketing Research, 1(3), 183-188.

[3] Yang, L. (2015): An Application of Principal Component Analysis to Stock Portfolio Management.

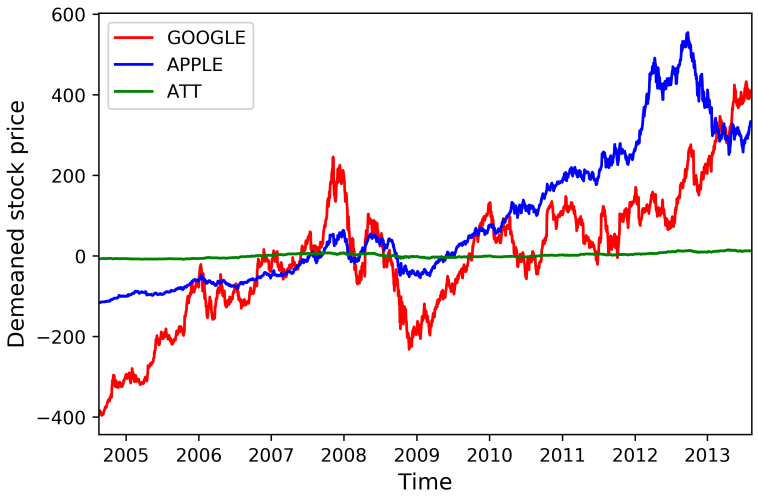

Figure 1: The trend of demeaned stock price of Google, Apple and AT&T.

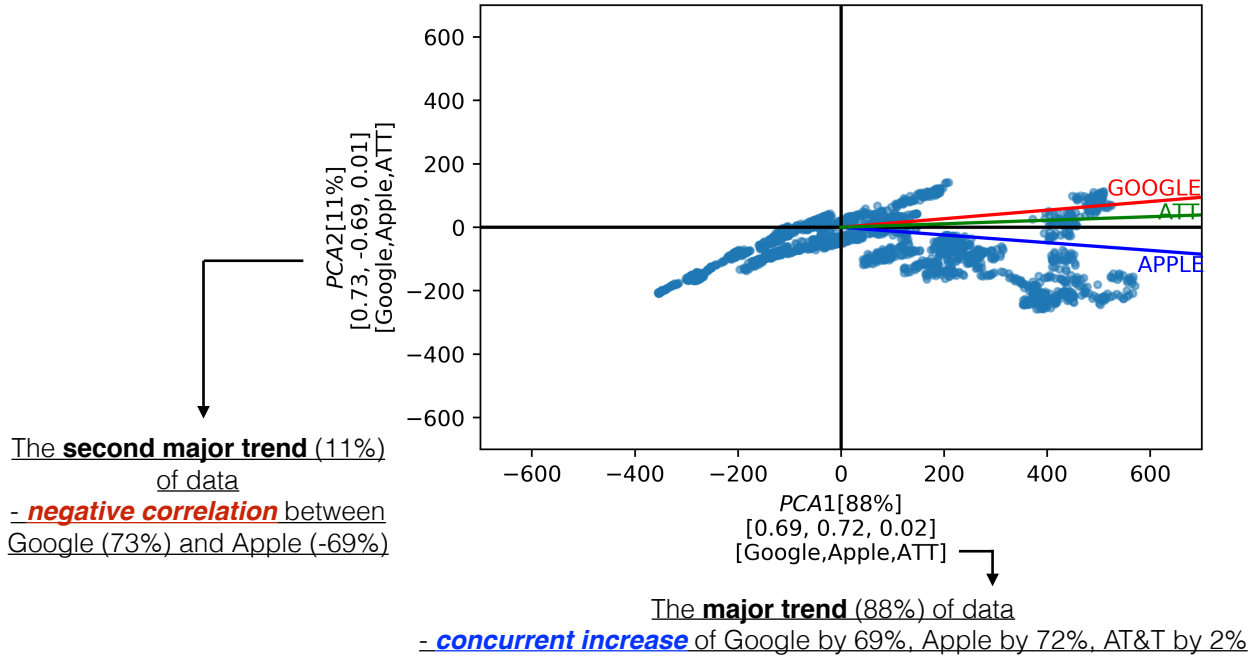

Figure 2: PCA1 and PCA2 of demeaned Google, Apple and AT&T stock price from August 2006 to August 2013.

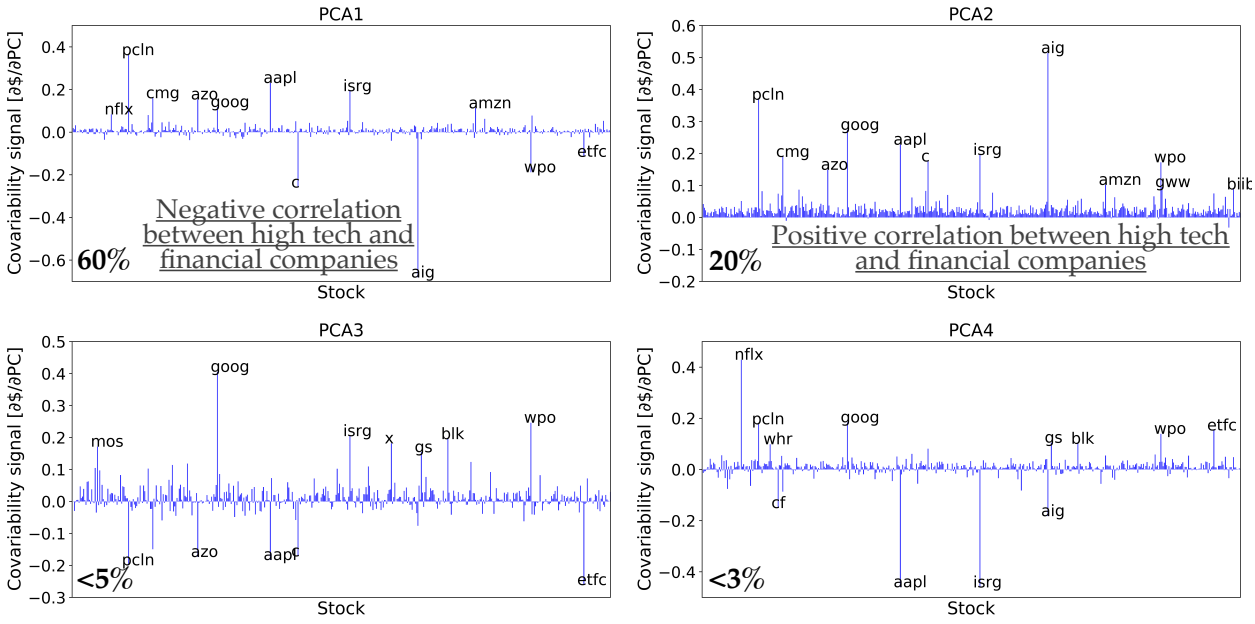

Figure 3: PCA1 to 4 of all available stocks from S&P 500 stock from August 2006 to August 2013.

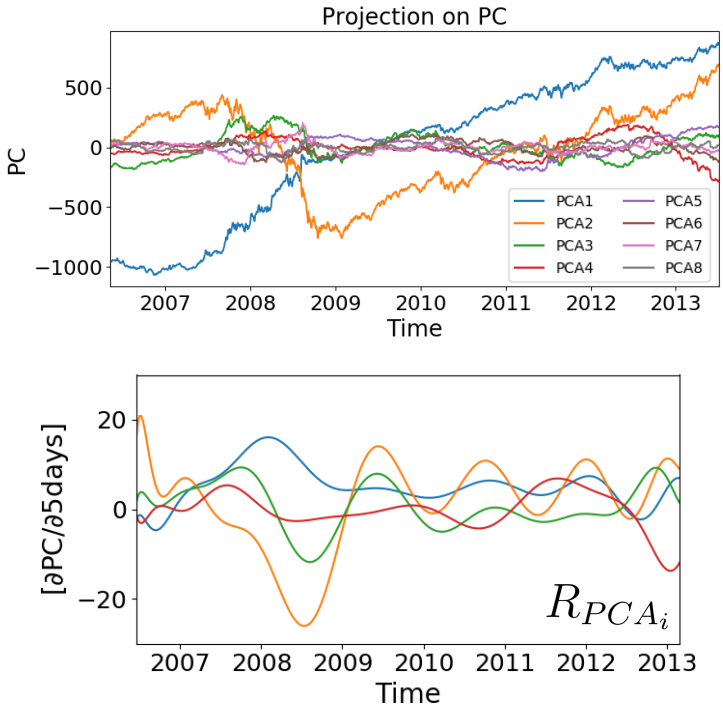

Figure 4: Projection of S&P500 stocks price on the PCAs (top panel) and their moving rate (gradience of upper panel) (bottom panel).

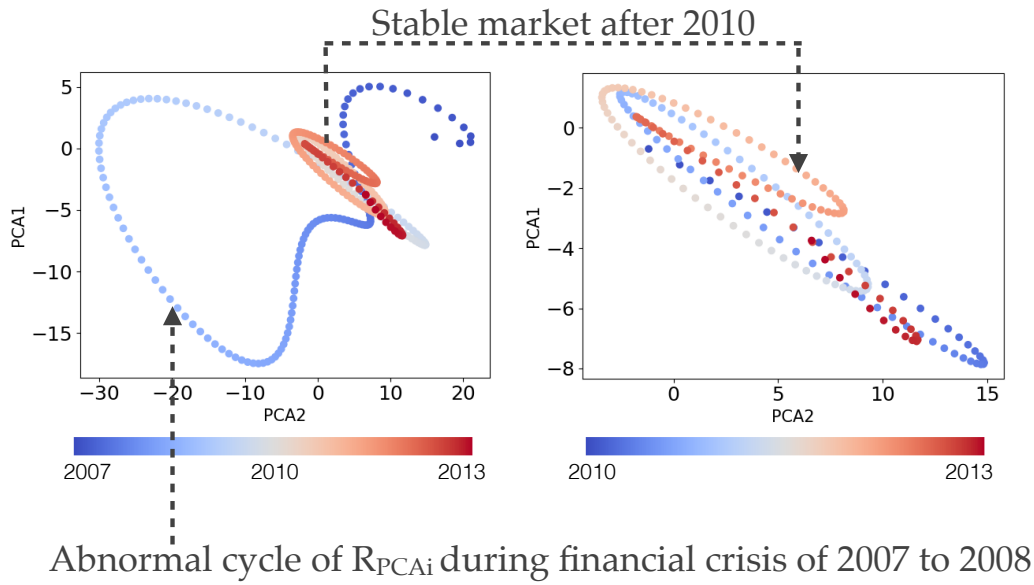

Figure 5: PCA of the PC projection moving rate of stock prices from 2007 to 2013 (left panel) and from 2010 to 2013 (right panel) (i.e., PCA analysis on the second panel of Fig. 4). PCA1 can explain 60% of the variance, PCA2 can explain 20% of the variance.

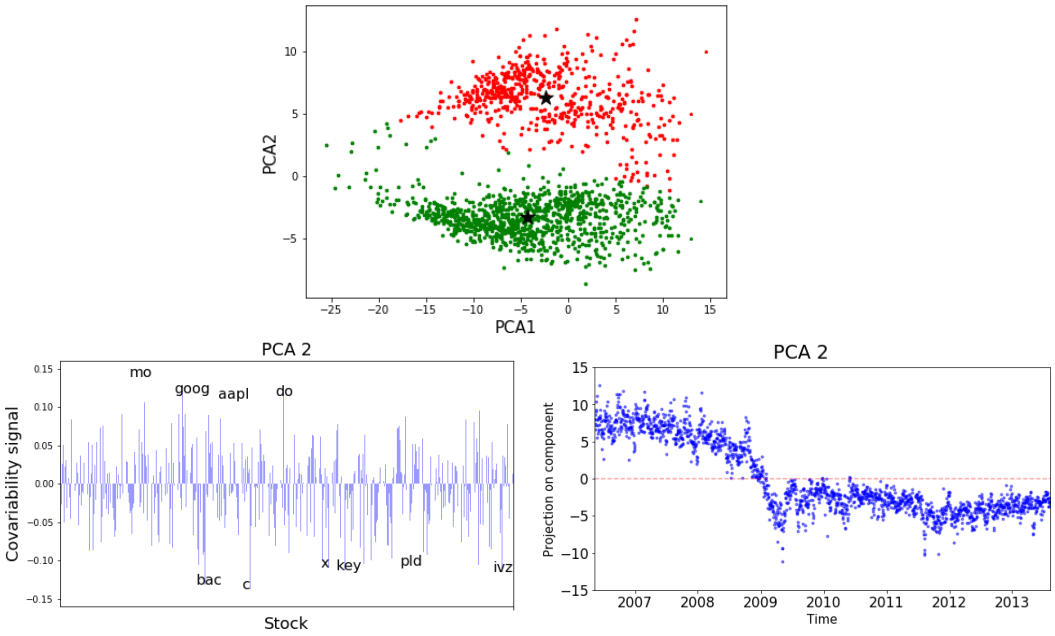

Figure 6: PCA of stock trade volume history. Scatter plot of PCAs (top panel), eigenvector of PCA2 (bottom left panel) and time series of PCA2 projection (bottom right panel).

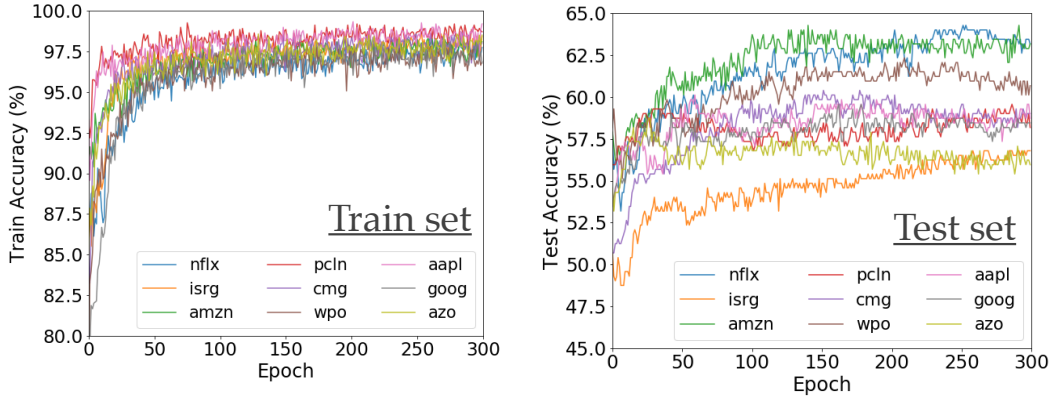

Figure 7: Training accuracy and testing accuracy of daily price trend prediction (up or down compared to yesterday's price) of deep learning for 9 selected stocks.

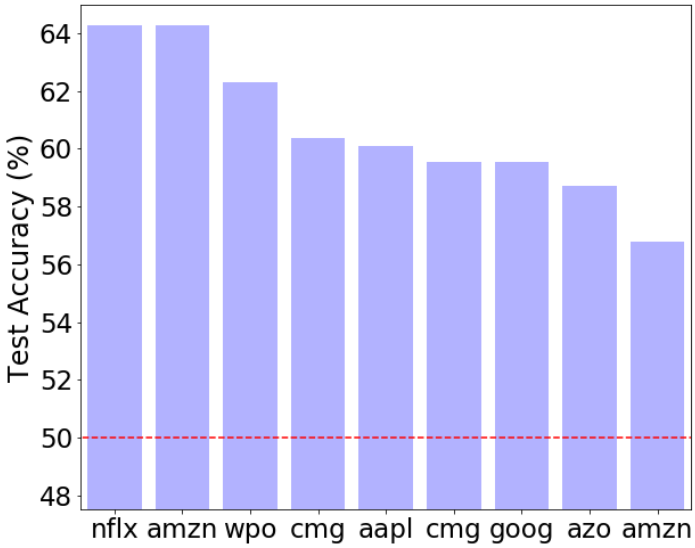

Figure 8: Test accuracy of daily price trend prediction (up or down compared to yesterday's price) of deep learning for 9 selected stocks.