
Writing Sample: 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

1 In this project, unsupervised learning and supervised learning are applied to daily
2 stock price of S&P 500 stock datasets from August 2006 to August 2013. By using
3 the scientific-based methods, the main objective of this study is to decrease the
4 investment risk and maximize the stock investment profit based on the machine
5 learning techniques. First, principal component analysis (PCA), an unsupervised
6 learning tool, is applied, which can help extract the major trends determining
7 the variability of stocks and identify specific stocks that represent the whole 500
8 leading stocks groups. The results show that the largest variability is dominated
9 by the negative co-variability between financial and high-technology stocks. In
10 addition, it is found that the S&P 500 market moving trend could be well explained
11 by a periodic cycle of major PCA components (80% of the total variance) during
12 stable market period, and the abnormal cycle during 2007-2008 financial crisis
13 could be captured by this analysis as well. Based on the analysis of PCA, a risk
14 index of stock investment is proposed and discussed. Second objective of this study
15 is to predict future daily moving trend of individual stocks by using supervised
16 deep learning. By using deep learning, nine selected stocks daily stock trend (up
17 or down, binary question) are trained and predicted. Our results suggest that test
18 accuracy is varied among stocks, and could reach 65% in average, which implies
19 that deep learning could be an useful reference and tool for stock trading strategy.

20 1 Introduction

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

36 supervised deep learning is a suitable tool for the goal of trend prediction, and is used in our study.
37 This report is structured in 4 sections. Section 1 is introduction. Section 2 describes the methods and
38 data used in this study. Section 3 presents the results. Section 4 gives the summary.

39 2 Methodology

40 2.1 Data collection and data cleaning

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

46 2.2 Principal component analysis

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

57 2.3 Deep learning

58 In our study, the supervised deep learning package provided by TensorFlow is used to explore the
59 prediction accuracy of daily stock trend by deep learning technique. The training data ranges from
60 Aug 2006 to Feb 2012 (80%), and the test data ranges from Mar 2012 to Aug 2013 (20%). Input
61 features are all S&P 500 stock prices and daily stock change ratio at time $T=t$ and $t-1$ (i.e., previous
62 two days), and output is the stock price at time $T=t+1$. Our model includes 4 hidden layers with 1024,
63 512, 256 and 128 neurons at 1st, 2nd, 3rd and 4th layers, respectively.

64 3 Result

65 3.1 Risk management of stock investment by PCA

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

84 resulted from the PC change along each PCAs direction. In other words, if there is no changes
 85 on PC (projection), then stock market is constant. With these information, we propose that the
 86 return due to the change of PC1 (the projection of stock prices on PCA1 direction) could be written as:
 87

$$88 \quad IR_i = \frac{PCA_1}{S_{price}} P_{portfolio} R_{PCA_1} = Risk_i R_{PCA_1} \dots\dots\dots (1)$$

89
 90 where IR_i is the investment return ($d\$/dt$) $\in R^1$, PCA_1 is the eigenvector of PCA1 ($d\$/dPC$)
 91 $\in R^{467}$ (i.e., 467 stocks are selected), S_{price} is the stock prices (\$) $\in R^{467}$, $P_{portfolio}$ is the portfolio
 92 for each stock investment combination (\$) $\in R^{467}$, R_{PCA_1} is the PC change rate (dPC/dt) $\in R^1$
 93 (which is the moving rate of projection of stock price on PCA1), and $Risk_i$ is the investment risk
 94 due to the market variability caused by PCA1. Note that same formula could be applied to PCA2,
 95 PCA3 and so on. Therefore, based on the information of PCA, a simple risk index as a function of
 96 investment portfolio, stock price and PCA eigenvector is derived. The equation 1 above indicates
 97 that the risk due to certain PC (i.e., projection) is worth considered only when the change rate of
 98 projection R_{PCA_1} is large enough (second panel of Fig. 4).
 99

100 From Eq. (1), it is known that R_{PCA_i} is the key parameter determining the sign of the return (i.e.,
 101 gain or loss). The second panel of Fig. 4 suggests that there might be some periodic cycle/relationship
 102 among R_{PCA_i} , and might be useful for the investment strategy. Figure 5 shows that PCA scatter
 103 plot (color bar is the time) of the PC projection moving rate of stock prices from 2007 to 2013 (left
 104 panel) and from 2010 to 2013 (right panel) (i.e., PCA analysis on the bottom panel of Fig. 4), which
 105 showing that during the stable market after 2010, there is period cycle in the market controlled by
 106 PCA1 and PCA2 of PC moving rate of stock prices, and during the financial crisis of 2007 to 2008,
 107 there is abnormal cycle of R_{PCA_i} . The result here suggests that the metric R_{PCA_i} could potentially
 108 be a useful tool to determine the stock investment strategy and predict the future market trend. Figure
 109 6 shows the PCA analysis on the stock trade volume historical data, which indicate that after the
 110 financial crisis of 2007 to 2008, the stock market is completely transformed from a financial-leading
 111 market to the market that is leaded by high technology companies (see bottom panels of Fig. 6).

112 3.2 Application of deep learning to trend prediction

113 In the application of deep learning to trend prediction, the input features are all selected 467 stock
 114 prices and daily stock change ratio at time $T=t$ and $t-1$ (i.e., previous two days), and output (the goal)
 115 is the stock price at time $T=t+1$ (i.e. tomorrow's stock price). There are 9 stocks that are selected here
 116 for the prediction. Figure 7 shows the accuracy of training set and the testing set. As one can imagine,
 117 training accuracy could reach 95% for the selected stocks. However, the testing set accuracy could
 118 only reach 65% in average for the selected stocks. Figure 8 summarizes the accuracy for the test
 119 set, indicating that the testing accuracy of predicting up or down binary questions is varied among
 120 stocks, and could reach 65% in average. Given that the random guessing accuracy could reach 50%
 121 in average, the results from deep learning can beat random guessing by 15%, which could be a useful
 122 reference for the development of stock trading strategy.

123 4 Summary

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

136 **5 References**

- 137 [1] JP Morgan (2017): Big Data and AI Strategies: Machine Learning and Alternative Data Approach to
138 Investing.
- 139 [2] Hargreaves, C. A., & Mani, C. K. (2015): The Selection of Winning Stocks Using Principal Component
140 Analysis. American Journal of Marketing Research, 1(3), 183-188.
- 141 [3] Yang, L. (2015): An Application of Principal Component Analysis to Stock Portfolio Management.

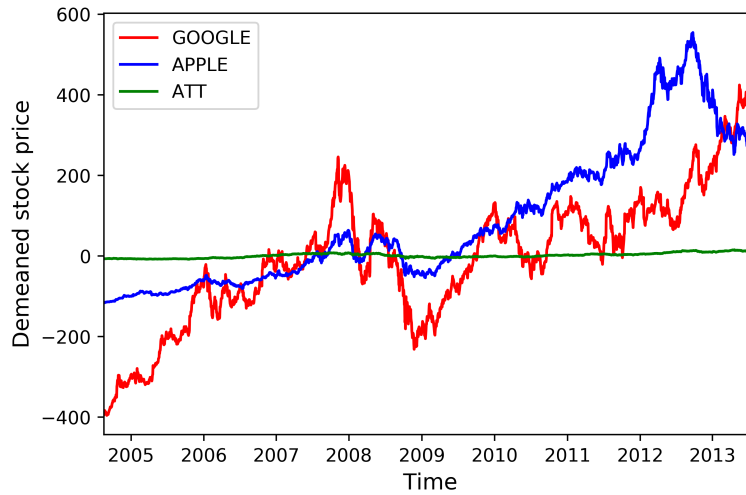


Figure 1: The trend of demeaned stock price of Google, Apple and AT&T.

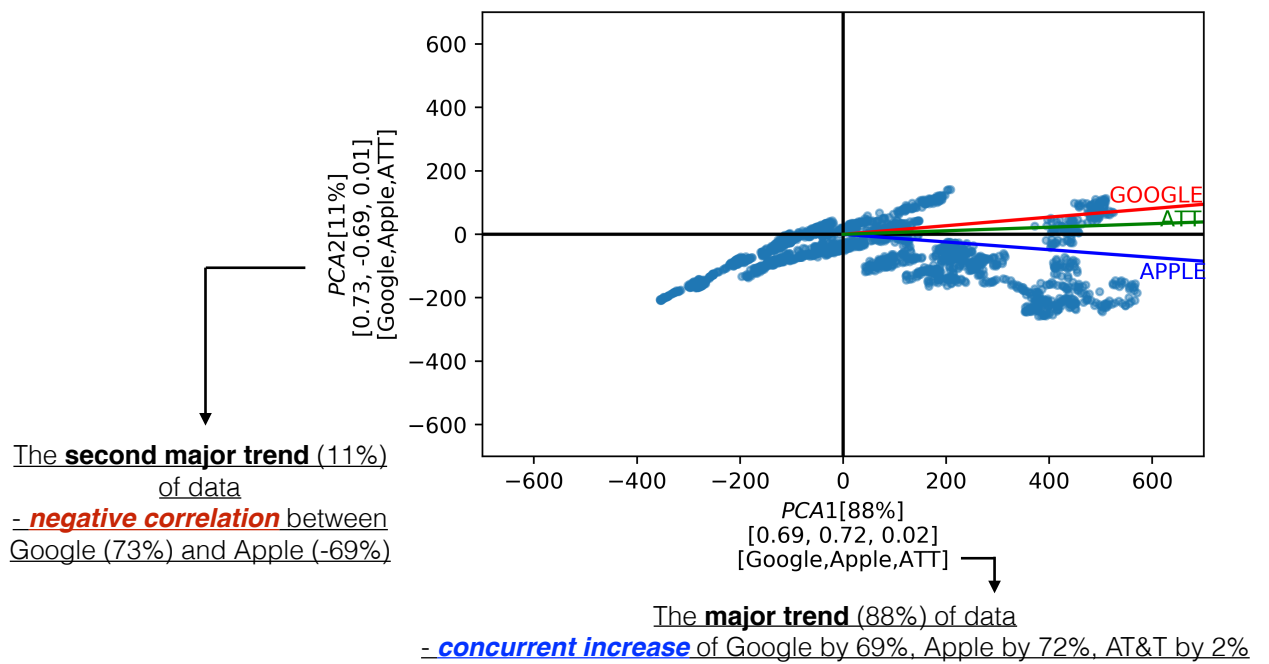


Figure 2: PCA1 and PCA2 of demeaned Google, Apple and AT&T stock price from August 2006 to August 2013.

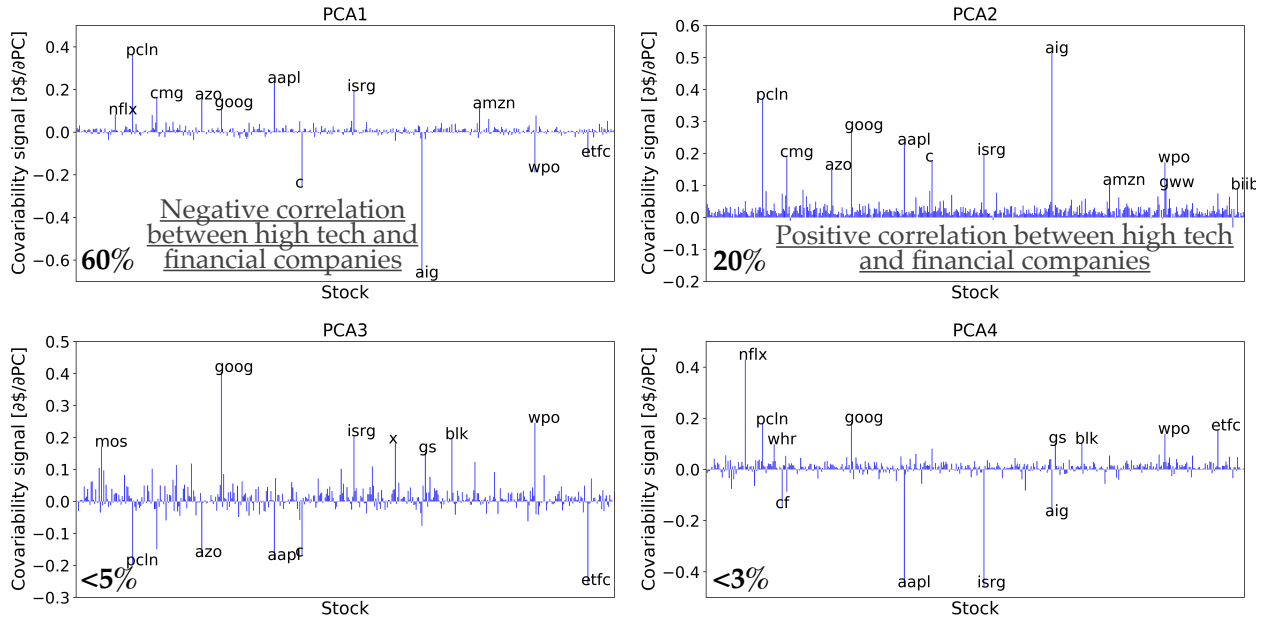


Figure 3: PCA1 to 4 of all available stocks from S&P 500 stock from August 2006 to August 2013.

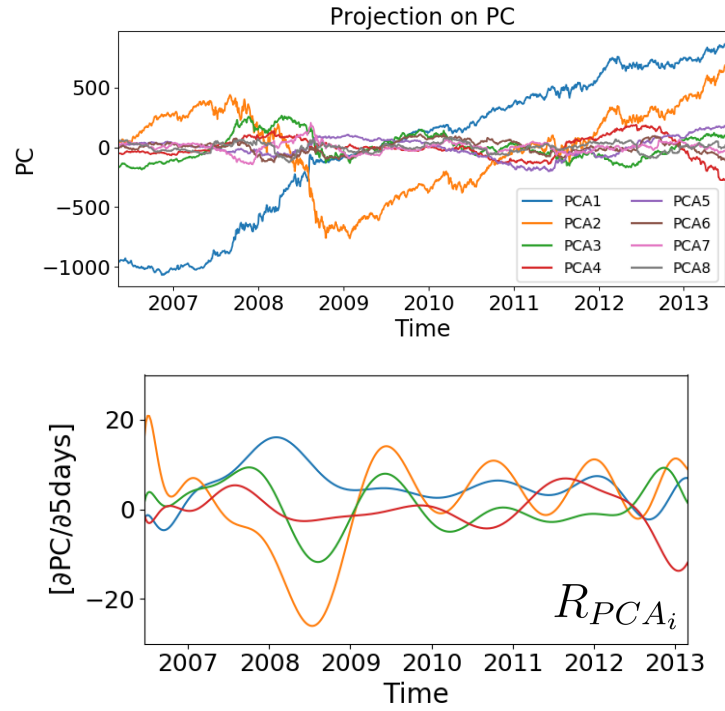


Figure 4: Projection of S&P500 stocks price on the PCAs (top panel) and their moving rate (gradient of upper panel) (bottom panel).

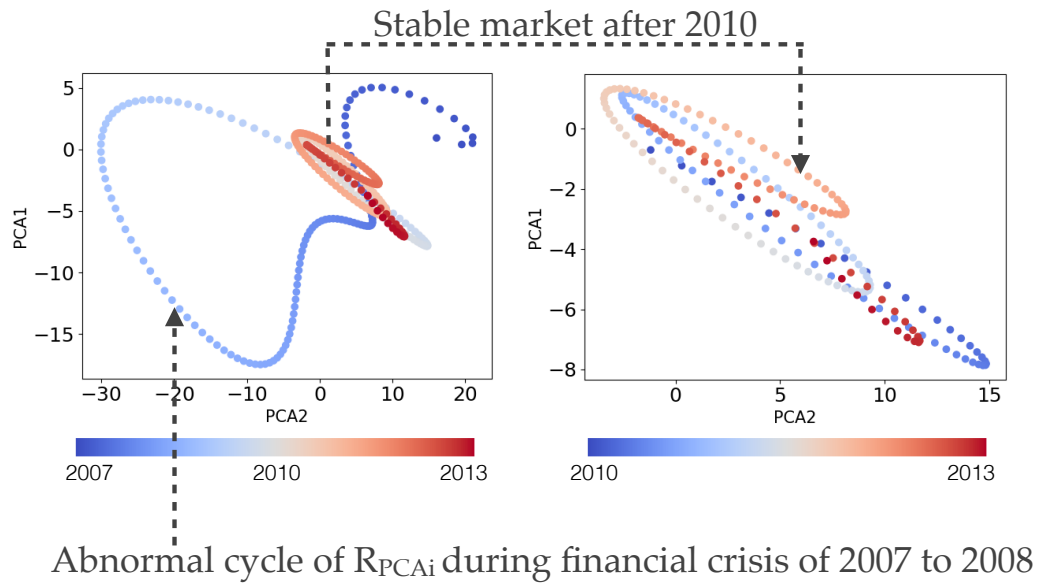


Figure 5: PCA of the PC projection moving rate of stock prices from 2007 to 2013 (left panel) and from 2010 to 2013 (right panel) (i.e., PCA analysis on the second panel of Fig. 4). PCA1 can explain 60% of the variance, PCA2 can explain 20% of the variance.

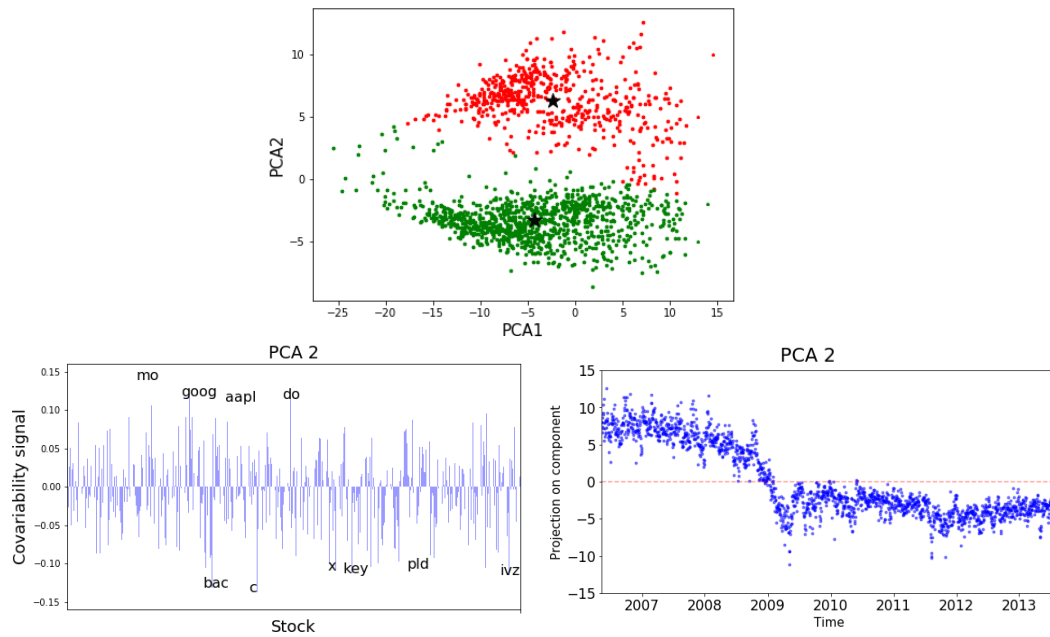


Figure 6: PCA of stock trade volume history. Scatter plot of PCAs (top panel), eigenvector of PCA2 (bottom left panel) and time series of PCA2 projection (bottom right panel).

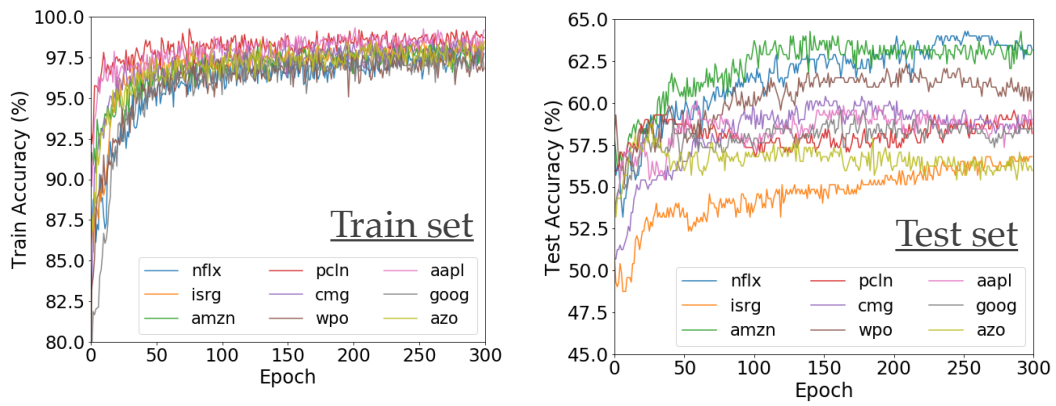


Figure 7: Training accuracy and testing accuracy of daily price trend prediction (up or down compared to yesterday's price) of deep learning for 9 selected stocks.

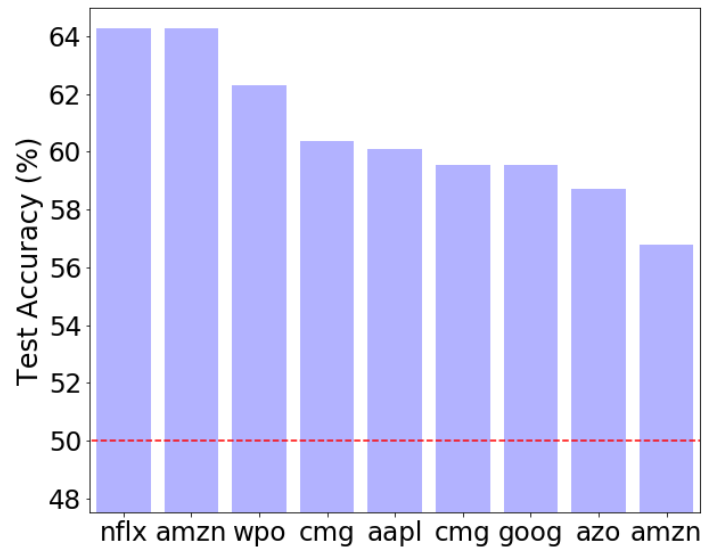


Figure 8: Test accuracy of daily price trend prediction (up or down compared to yesterday's price) of deep learning for 9 selected stocks.