Stock Prediction using Factor-Based Fundamental Approaches together with Neural Networks

Ritvik Dhupkar

Abstract—This paper is inspired from Alberg and Lipton [9]. The paper attempts to combine Deep learning techniques with Fundamental Factor based approaches found in Economic Literature. It Models Stock Price over the long term (12 months) using, historical data on the Price, the Market Index and Fundamental Factors. The Factor model' is then compared with a Univariate Deep Learning model using only the historical data on Stock, and a Bivariate Deep Learning model that uses the Market Index (XOI) with the Stock Data.

1 Introduction

Prediction of Stock market has gathered attention from both Industry and Academia. Various Machine learning models like Neural Networks, Support Vector Machines and LSTMs have been used to predict Stock prices. Recurrent Neural Networks are an extremely powerful tool for processing Sequential data. The Changes and fluctuations of Stock Prices occur on multiple time scales. In the short run, price fluctuations are usually determined by the dynamics of order executions and the dynamics of high frequency traders. In the medium term, on the scale of days, the price of stocks may be determined by news cycles. Stock prices may rise or fall, depending how the company is portrayed in the news media. In the long run however, Stock prices are believed to converge towards the true value of the company, which is given by fundamentals or financial and accounting information that reflects the true value of the company. Deep learning models are used to predict Stock fluctuations at all three levels. In the short run, predictions may be made only using the past values of the stock. In the medium term, natural language processing techniques are used to parse news and social media to provide inputs to deep learning models for stock prediction. In the long run, stock fundamentals are used as inputs to predict the stock prices. Academic Research has identified some factors that are used in factor models to compute the 'Intrinsic Value' of Stock. These are Price/Earnings Ratio, Price/Book Ratio, Book Value of Equity per share, the Market Capitalization of the Company, the Dividend Payout Ratio and the prevalent value of the Stock/Industry index.

This paper compares the performance of three models for the prediction of Stock Price in the Oil & Gas Sector. The Univariate Model uses only the historical data of Stock price to make long-term (12 months) predictions of the Price. the Bi-Variate model uses historical data from the Stock Price and the Market Index i.e. XOI (The NYSE index for Oil & Gas companies). The Multivariate Model uses historical data on a number of fundamental ratios and parameters in addition to the historic data on the Stock Price and the Market Index (XOI).

2 Literature Review

2.1 Investigation of financial market prediction by Recurrent Neural Networks

Nijole, Aleksandras and Algirdas [7] explored effects of different epochs and number of neurons on time series prediction results. The authors aimed to investigate the best parameter settings to achieve a good result in the financial market and studied the effects on RMSEs under different epoch settings of a Recurrent Neural Network model. It was observed, that a small number of epochs (around 16) did not allow the network to learn sufficiently and provided poor results. Beyond 164 epochs, it was observed that no significant gains were made in terms of increase in accuracy.

2.2 Improving Factor-Based Quantitative Investing by Forecasting Company Fundamentals

Alberg and Lipton uses 'fundamental stock information' like revenue, operating income and debt to project financial health of a stock. Academic research has identified some factors, i.e. computed features of the reported data, that are known through retrospective analysis to outperform the market average. Two of these features are book value normalized by market capitalization and the operating income normalized by enterprise value (EBIT/EV). Traditional Quantitative investment strategies score stock according to some factors calculated on 'Stock Fundamentals and the current price of the stock. The stocks that score best on these factors (henceforth called 'value factors) are selected and are called 'value stock.

Value factors can be used to identify companies stock that are low priced with respect to current fundamentals (earnings or book value). This paper suggests that the long term success of an investment depends on how well priced the stock is with respect to future fundamentals. This paper projects company ‘fundamentals in the future and attempts to use these to compute value factors based on future fundamentals. The paper predicts future fundamental data based on a trailing time series of five years of fundamental data. The paper considers all stocks publicly traded on NYSE, NASDAQ or AMEX exchanges for at least 12 consecutive months from January 1970 to September 2017. The paper excludes stocks of Non-US based companies and any companies whose inflation adjusted market capitalization is less then 100 million are excluded. This results in a dataset of 11815 stocks. The ‘Fundamentals are obtained from the CompStat North America or the CompStat Snapshot datasets. The paper considers 20 input features at each timestep t.

These features include features from the income statement, which a are cumulative and are (denoted as TTM or Trailing 12 Months) and features from the balance sheet from the previous quarter (denoted as MRQ or Most Recent Quarter). The Features are: Revenue (TTM); Cost of goods Sold (TTM); selling, general and admin expense (TTM); earnings before interest and taxes or EBIT (TTM); net income (TTM); cash and cash equivalents (MRQ); receivables (MRQ); inventories (MRQ); other current assets (MRQ); property plant and equipment (MRQ); other assets (MRQ); debt in current liabilities (MRQ); accounts payable (MRQ); taxes payable (MRQ); other current liabilities (MRQ); total liabilities (MRQ). Four momentum features were used in addition to the features mentioned. Using the ‘Lookahead model, that uses projected fundamental prices, the paper calculates the value factors and creates a list of 50 ‘value stock with amounts invested equally in each stock. When a stock falls out of this list, it is sold and the proceeds are invested in the stock with the best ‘value factor. The results show that in nearly all months, the look ahead model outperforms traditional models, Multi-layer perceptron models and RNN models, irrespective of market conditions.

2.3 Stock Market Forecasting using Recurrent Neural Networks

Gao and He [8] used Raw Stock data consisting of open price, highest price, lowest price, close price and transaction volume of a specific time point, to model stock price behavior. Six different Stocks, were chosen randomly from six different industries. Hourly Stock Price data was used, from iqfeed.net, and each stock had a data set size of 3703 entries. This data was deemed insufficient to train the LSTM model. To mitigate this, the raw data was pre-processed 369 features including information on competitor stocks and index data, which were then forwarded into the LSTM model.

The Indices used were the NASDAQ index, and the S&P 500 Index. Two Competitor Stocks were chosen for each stock target stock. 369 features were created using lag values of the target Stock, Lag values of the Market Indices and lag values of the competitor stock. The performance of the Stock was evaluated in terms of percentage increase of next three hours’ highest price compare with current hour open price. The stock performance was classified into One of three categories- Increasing between 0-1%, increasing above 1% and not increasing. The architecture used for the classification problem was a hidden layer, followed by a LSTM block, which was followed by another hidden layer and an output layer.

The model was used to build a trading engine, and an average return of 6% over 400 hours was achieved by the simulator.

2.4 Fundamental Factor Models

The common-factor variables are determined using fundamental, asset-specific attributes such as; 1) the sector/ Industry the stock is present 2) Market Capitalization 3) Dividend Yield 4) Accounting measurement technique - Book to market Ratio, Price Earnings Ratio, etc. Fama developed a Factor model technique called Fama-Macbeth Regressions, which could be used to predict price of Stock based on a number of fundamental variables - like P/E ratio, B/P ratio, ratios involving dividend payouts, Book Value of Equity per share, the Market Capitalization of the Company.

2.4.1 Fundamental Factor Models: Fama and French, 1992

In ”Cross Section of the Expected Stock Returns”, fama and French [4] investigated factors affecting market $\beta's$ and established that the Firm Market capitalization and Book to market Equity are able to explain most of the variation in returns attributed to β. Fama-Macbeth (FM) regessions are performed with Cross Section returns and size, E/P, leverage and BE/ME for each stock. The results of these regressions are as follows; 1) When β is regressed keeping size constant, there is no relationship between β and the average return. 2) The opposite roles of market leverage and book leverage in average returns are captured well by book-to-market equity. 3) The relation between E/P and average return seems is absorbed by the combination of size and book-to-market equity. Fama and French create a chart based on book to market ratios and size of the firm. It was observed that there is a negative relationship between the size of the firm and the returns.

The fact that small firms can suffer a long earnings depression that bypasses big firms suggests that size is associated with a common risk factor that might explain the negative relation between size and average return. Similarly. the relation between book-to-market equity and earnings suggests that relative profitability is the source of a common risk factor in returns that might explain the positive relation between BE/ME and average return.

Fama and French performed time series regression of returns with factors. The following variable constructions were performed:

RM-RF - Average premium per unit of market β. This is the difference between the expected market return and the risk free return (RF).

SMB- Average premium for size related factors in returns.

HML- High minus Low is he difference, each month, between the simple average of the returns on the two high BE/ME portfolios and the average of the returns on the two low BE/ME portfolios.

The paper shows that the excess return on the market portfolio of Stock RM-RF captures more variation than any other variable. This variable achieves an R-Square of 0.9 for low BE/ME stocks and achieves an R-Square of 0.7 for low size - high BE-ME stocks. When SMB and HML are regressed independently on returns, without RM-RF, they achieve an R-Square of 0.25. When SMB, HML and RM-RF are regressed together, the predictive power for low size high BE/ME stocks increases significantly, achieving R-Square of 0.8-0.9. This is shown in the model below:

$R(t) - RF(t) = a + b[RM(t) - RF(t)] + sSMB(t) + hHML(t)$

2.4.2 Arbitrage Pricing Theory applied to Fundamental Factors

Arbitrage pricing theory (APT) is a multi-factor asset pricing model based on the idea that an asset's returns can be predicted using the linear relationship between the assets expected return and a number of fundamental variables that capture systematic risk. It is a useful tool for analyzing portfolios from a value investing perspective, in order to identify securities that may be temporarily mispriced. In "Cross Section of Expected Price Returns" Fama and French performed regressions between fundamental factors and stock returns. They established that there were significant relationships between fundamental factors like Earnings per share, firm size, book to market equity, and leverage. This section expands on the the Fama French three factor and five factor models.

Kalmazzi and Borri [10] applied the following fundamental factor models to stock price:

(1) $R_{it} = a_t + b_{1t} * ln(ME_{it}) + b_{2t} * ln(A/ME_{it}) + b_{3t} * ln(A/BE_{it}) + b_{4t} * (E/P) * Dummy_{it} + b_{5t} * (E/P)(+)_{it} + u_{it}$

(2) $R_{it} = a_t + b_{1t} * ln(ME_{it}) + b_{2t} * ln(BE/ME_{it}) + b_{3t} * (E/P) * Dummy_{it} + b_{3t} * (E/P)(+)_{it} + u_{it}$

R_{it} = Stock Returns
ME = Market Capitalization. The total Value of Shares traded.
BE/ME = Book-to Market. Book value of common equity over market capitalization. This is equal to the P/B ratio which calculates the ratio of price to book value of equity per share.
A/ME = Market Leverage. The ratio of the book value of total assets over book equity. This is equal to the Equity to Assets Ratio.
E/P = Earnings/Price Ratio
Market Excess Returns = Stock Return - Risk Free Return

3 Data Description

The data consists of two datasets; The Stock Price data and the fundamentals data. The Stock Price dataset was obtained from Yahoo Finance while the Fundamentals dataset was obtained from an online source (stockpups.com). The Stock Price Dataset consists of daily Stock Price data. The four relevant price measurements used in the paper were 'Open', 'High', 'Low' and 'Close'. 'High' is the highest trading price of the security on the given day. 'Open' is price at which the security trades when the market opens. 'Low' is the lowest trading price of a security on the given day. 'Close' is the closing price of the security on the given day. This paper focuses on prediction of securities in the Oil & Gas sector.

The Fundamentals dataset consists of quarterly data from Balance Sheet, Cash Flow, and Income statements of Companies. The dataset compiled on stockpups.com, was obtained from XBRL filings and filings with the US securities and exchange commission. Some of the important fundamental factors used in the model are given below:

P/E ratio - The ratio of Price to EPS diluted TTM as of the previous quarter.
P/B ratio - The ratio of Price to Book value of equity per share as of the previous quarter.
Book Value of Equity per Share (BVPS) - Common stockholders' equity per share.
Dividend Payout Ratio - The ratio of Dividends TTM to Earnings (available to common stockholders) TTM.
Long-term debt to equity ratio - The ratio of Long-term debt to common shareholders' equity.
XOI - It is the NYSE market index for securities in the Oil & Gas sector.

The Dataset is not uniform. It contains 20 years of quarterly fundamentals data for some of the Companies (not all). XOI is the NYSE market index for Oil & Gas companies. It is obtained from Yahoo Finance.

4 Methodology

Securities Classified under the Oil & Gas sector, which have more than 20 years of Fundamentals data are selected. There are 16 Securities that satisfy this condition. The Fundamentals Data is Quarterly and consists of 80 data points between January 1997 and March 2018. The Stock Price data set, including XOI is selected from 1 Jan 1997 to 1st April 2018.

The following Operations are performed:
1) A daily Average price is calculated from the Price dataset from the Open, High, Low and Close Prices.
2) A Monthly Average price is calculated from this daily average price and the Price Dataset, including XOI are merged with the fundamentals dataset.
3) A Linear Interpolation is performed on the Fundamental dataset to obtain monthly data for fundamentals.
4) A LSTM model is applied to the dataset. Iterations are performed over the model to choose the most Significant Predictors. This is called the Multivariate Model. Since the Total number of data points for 20 years is only 240 for a particular security, A large number of predictors cannot be selected for the Multivariate model.
5) Tuning and Optimization of hyperparameters and the number of hidden layers are performed.
6) The Results obtained from the Multivariate model are compared with the Bivariate model- Consisting only of the security and Market Index (XOI) and a Univariate model consisting only of the security.

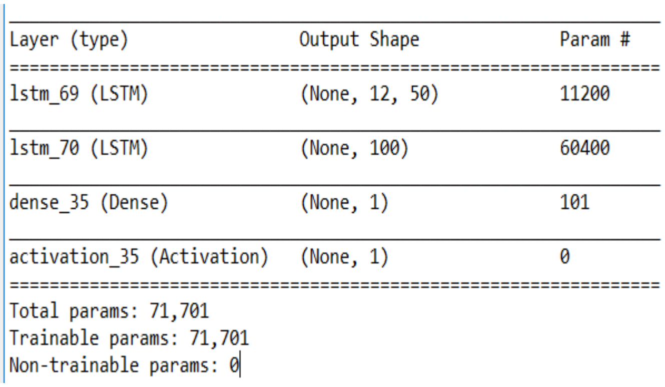
The Justification for the last step is that there are a number of Deep-Learning approaches used to predict short-term Stock Behaviour using historical data on the price of security, or using historical data on the market index and price of security. However, Long term Predictions of Stock Price using Machine Learning have been Unsuccessful. The Market Index is a commonly used benchmark for the performance of the sector and can give indications for the overall performance of the Sector. Hence, it is included in the Multivariate Model.

5 Results

Analysis is performed and presented for three models:

1) Multivariate Model: Monthly average price is predicted using the features. These features were selected based on a survey of literature and based on multiple iterations on parameters.The features are - P/E ratio, P/B ratio, BVPS and XOI. Since the Data size is small, a large number of predictors cannot be selected as inputs to the model.
2) Bivariate Model: Predictors are the historical data of security and XOI index.
3) Univariate Model: Predictors are the historical data of the security.

The following Architecture was Implemented for the multivariate case. The Same Architecture was used for the Bivariate models.



```
Layer (type)                 Output Shape              Param #
=================================================================
lstm_69 (LSTM)               (None, 12, 50)            11200
_________________________________________________________________
lstm_70 (LSTM)               (None, 100)               60400
_________________________________________________________________
dense_35 (Dense)             (None, 1)                 101
_________________________________________________________________
activation_35 (Activation)   (None, 1)                 0
=================================================================
Total params: 71,701
Trainable params: 71,701
Non-trainable params: 0
```

All values are in % RMSE (Expressed as a percentage)

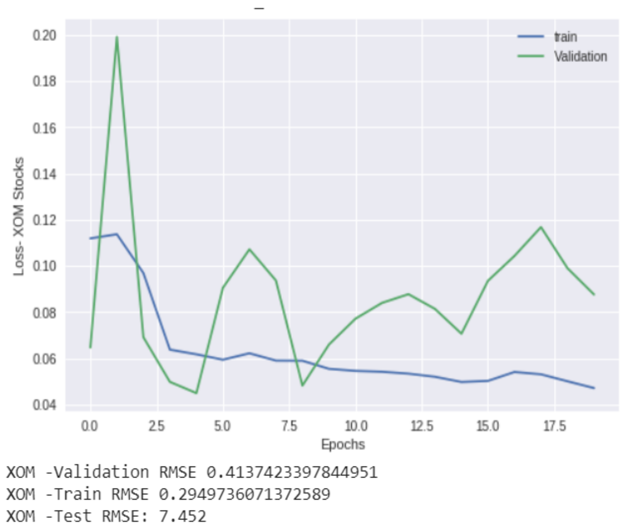
Security	Multivariate	Bivariate	Univariate
RDC	72	4.3	33.5
DO	11.8	58.5	16.9
ESV	322.3	149.5	75
NBL	9.2	11.6	17.9
RRC	37.0	75.9	153.5
MUR	37.2	13.5	40.6
OKE	48	21.8	13.1
SWN	83.8	268.9	176
HP	2.3	6.3	32.3
OXY	16.9	11.9	10
EQT	71.4	94.8	54.6
NFX	20.4	32.8	52.9
PXD	54.2	5.55	2.6
EOG	43.8	4.0	20.4
CVX	8.6	23.6	24.7
XOM	27.5	8	17.4

The average % RMSE for the Multivariate case is 54.1%, for the Bivariate case is 57.4% and for the Univariate case is 46.3%. Out of the 16 cases given above, the Multivariate model outperforms the Univariate model 8 out of 16 times, while the Bivariate model gives a better performance than the Univariate model for 8 out of 16 companies. This suggests that overall, the Univariate Model has the best performance.

The Plots for XOM (Exxon Mobil) using the above mentioned three methods is given below. The Tables on training RMSE for the securities using each of the above mentioned techniques are given in the Appendix A. It was observed during hyper-parameter tuning that epochs between 20-50 provide the best results on the test dataset. Epochs higher than 50 do not provide sufficient gains in learning to the model. Dropouts were implemented from 0-0.5. The best results were achieved for dropouts ratio ranging from 0-0.1. This may be due to the limited number of data points. An attempt was made to incorporate time distributed output techniques, however, this did not give satisfactory results. This may have been because of the small size of the Test dataset.

5.1 Multivariate Model

The Figure 1 plot gives the results of multivariate model for XOM (Exxon Mobil) security and plots the mean square error for training and validation dataset and the Figure 2 plots the predicted value of test data with the original values.

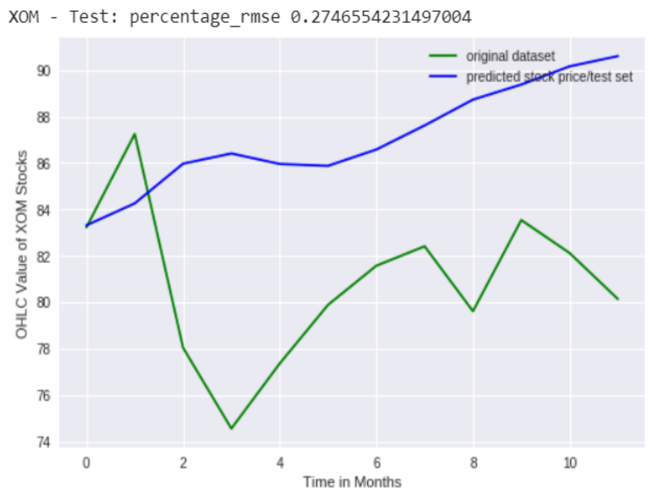


```
XOM -Validation RMSE 0.4137423397844951
XOM -Train RMSE 0.2949736071372589
XOM -Test RMSE: 7.452
```

Fig. 1. Multivariate Model RMSE plots for training and Validation Data for XOM stock

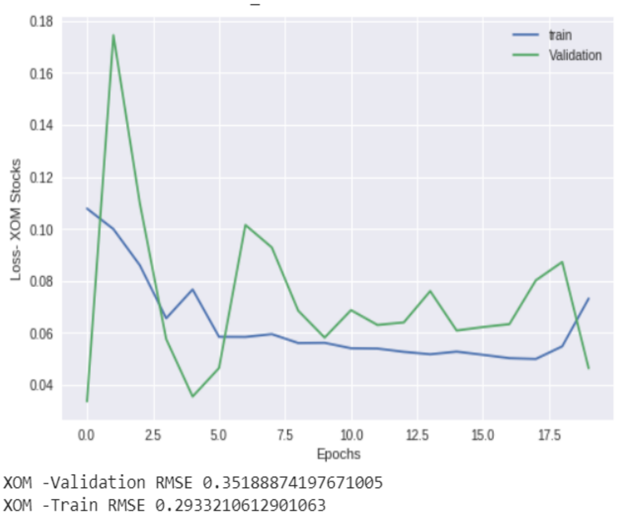
5.2 Bivariate Model

The plot gives the results of bivariate model for XOM (Exxon Mobil) security. Figure 3 plots the mean square error for training and validation dataset and Figure 4 plots the predicted value of test data with the original values.



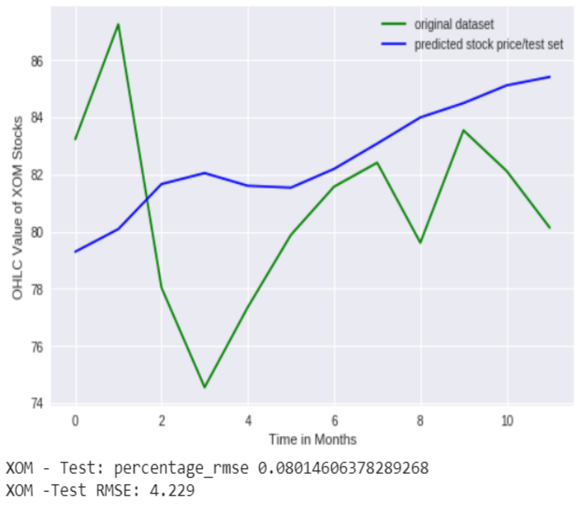
```
XOM - Test: percentage_rmse 0.2746554231497004
```

Fig. 2. Multivariate Model Predicted stock price with the original price for the test data for XOM stock



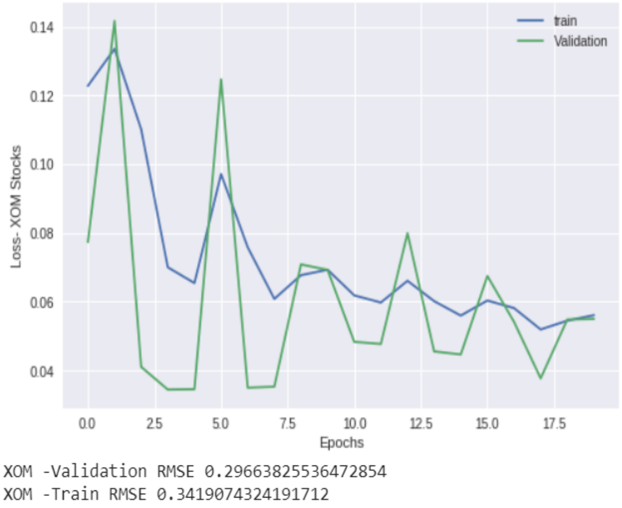
```
XOM -Validation RMSE 0.35188874197671005
XOM -Train RMSE 0.2933210612901063
```

Fig. 3. Bivariate Model RMSE plots for training and Validation Data for XOM stock



```
XOM - Test: percentage_rmse 0.08014606378289268
XOM -Test RMSE: 4.229
```

Fig. 4. Bivariate Model Predicted stock price with the original price for the test data for XOM stock



```
XOM -Validation RMSE 0.29663825536472854
XOM -Train RMSE 0.3419074324191712
```

Fig. 5. Univariate Model RMSE plots for training and Validation Data for XOM stock

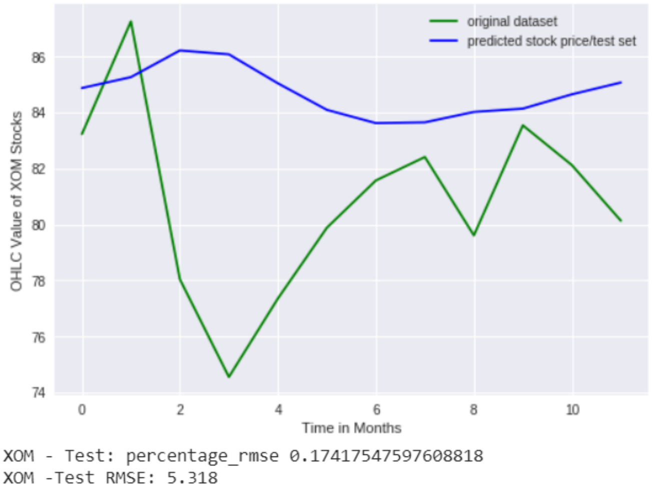
5.3 Univariate Model

The results of mean square error for training and validation dataset under univariate model for XOM (Exxon Mobil) security are plotted in Figure 5. The Figure 6 shows the predicted value of test data with the original values for the XOM company.

6 Conclusion

This Paper is unable to establish that deep learning models, when combined with Fundamental factors, improve the prediction accuracy significantly. Alberg and Lipton, are able to achieve a normalized value of Mean Square error of 47% for both the LSTM and the LSTM factor models. The superior accuracy of their factor model may be because, they considered data from January 1970- September 2016.

The quality of Data used in this analysis is unreliable for professional investments. For this project, we tried to get access to Wharton Research Data Services (WRDS) database [11], however the request was rejected by the owners with a statement that only professors can get access to that dataset. The methods mentioned above, should be replicated with higher quality data that can give superior results.It is to be noted that companies having a large Market Capitalization, which are members of S&P 100 (XOM, CVX and OXY), have low RMSE as compared to other companies. This is in confirmation with Fama and French (1992) who determined after a study of market returns, that companies with lower market capitalization have higher volatility and higher returns as compared with Companies higher market capitalization. This may be a contributing to the higher RMSE achieved by the multivariate model in this paper.



```
XOM - Test: percentage_rmse 0.17417547597608818
XOM -Test RMSE: 5.318
```

Fig. 6. Univariate Model Predicted stock price with the original price for the test data for XOM stock

7 References

[1] Kempthorne, Lee, Strela, Xia, ”Topics in Mathematics with Applications in Finance Topics in Mathematics with Applications in Finance”, Lecture 14,15,16 , MIT Open-Courseware, Course No-18.S096, Fall 2013

[2] Cochrane, ”Asset Pricing” (June 2000)

[3] Fama, French ” The Cross Section of Expected Stock Returns”, Journal of Finance, June 1992 .

[4] Fama, French ”Common Risk Factors in the returns of Stocks and Bonds”, University of Chicago, September 1992.

[5] Rosenberg, Mckibben, ” Prediction of Systematic and Specific Risk in common Stocks”, Journal of Quantitative and Financial Analysis, March 1973

[7] N. maknickiene, A. V. Rutkauskas and A. maknicksas, ”Investigation of financial market prediction by recurrent neural network,” Innovative Infotechnologies for Science, Business and Education, vol. 2, no. 11, pp. 3-8, 2011.

[8]-Qiyan Gao, Zhihai He Stockmarket Forecasting using Recurring Neural Network Masters Thesis, University of Missouri at Columbia.

[9]Alberg, Lipton ImprovingFactor-Based Quantitative Investing by Forecasting Company Fundamentals, 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

[10]-Borri, Kalmazzi, Sylvestri, ”Asset Pricing Models, Arbitrage Pricing Theory and Fundamental Analysis: Main Applications and the European Market Case”, Dipartimento di Economia e Finanza.

[11]-Borri, Kalmazzi, Sylvestri, ”Asset Pricing Models, Arbitrage Pricing Theory and Fundamental Analysis: Main Applications and the European Market Case”, Dipartimento di Economia e Finanza.

Appendix A
Training Results

All values are in RMSE (Expressed as absolute value)

Security	Multivariate	Bivariate	Univariate
RDC	0.56	0.57	0.56
DO	0.5	0.45	0.365
ESV	0.41	0.5	0.335
NBL	0.37	0.45	0.37
RRC	0.37	0.86	0.39
MUR	0.47	0.46	0.41
OKE	0.32	0.33	0.315
SWN	0.47	0.51	0.478
HP	0.31	0.32	0.34
OXY	0.51	0.4	0.41
EQT	0.31	0.56	0.41
NFX	0.34	0.44	0.45
PXD	0.34	0.34	0.34
EOG	0.37	0.30	0.29
CVX	0.34	0.35	0.3
XOM	0.29	0.29	0.34