Statistical Consulting Group 1 Project 2

Alexander Bonham, Weijie Qi, Erzhen Hu, Mingyu Qi

Brief Summary and Conclusions

In this paper, we found that there is no best model across all industries when doing long term 126 day predictions of stock prices. In other words, there is no general model, and different industries require different optimal models. This is not unexpected because stock price data is diverse and situational, and it is unlikely that any single model will be uniformly best across all industries or contexts. Based on our results, ARIMA GARCH methods are better for Consumer Discretionary and Financial industries, and LSTM models are better for Healthcare and Industrials. Specifically, we found Cumulative Year (CumYr) ARIMA GARCH performs best for the Consumer Discretionary industry, year by year (YrByYr) ARIMA GARCH performs best for Financials, YrByYr multivariate LSTM performs best for Healthcare, and YrByYr univariate LSTM performs best for Industrials. Overall, the LSTM models with YrByYr under multivariate condition perform better than LSTM models under other conditions.

After finding that there was no general model, we still attempted to find a potential model that might apply all the industries. So we transformed our data and fitted a classification model. We converted our data into a series of -1,0, and 1 depending on whether prices increased or decreased by a certain amount. We then ran an LSTM on this transformed data set predicting 126 new time points. This method did not work well, with a low accuracy rate around 40% for classification, as model predicted almost exclusively 0s. Thus, we concluded that LSTM is not fit for this situation and stock price data may generally be resistant to this kind of transformation.

Finally, we ran LSTM and ARIMA GARCH models predicting 20 days of stock prices. We found that ARIMA GARCH always outperformed LSTM models, which is reasonable as time series tend to predict well on the first several time points but predicts worse as more time points are added.

Description of the problem and data set

Our problem was to use basic information about stock values to predict future values based only on time-dependent, highly correlated stock price data. By its nature, our data posed unique challenges that required specifically designed methods to overcome.

Each year is 253 business days long, and we used the first 16 business years worth of stock price data to predict the following 126 days of stock price data. To this end we applied univariate methods, namely ARIMA GARCH and Univariate LSTM, and multivariate approaches, namely VAR and Multivariate LSTM. We were assigned twelve stocks, three stocks from each of four industries: Consumer Discretionary (CD), Finance, Healthcare (Hcare), and Industrials (Indstrls). In particular, our group was assigned CD stocks 1-AMZN, 3-DIS and 9-MCD, Finance stocks 1-AIG, 3-AXP and 13-USB, Hcare stocks 10-LLY, 11-MDT and 13-PFE, and Indstrls 2-CAT, 3-EMR and 11-UNP. It should be noted that AIG was an outlier, as its stock prices suffered an enormous drop during the 2008 financial crisis and never recovered.

Description of Methods we tried

We tried methods such as ARIMA GARCH, VAR, and Univariate and Multivariate LSTM. Different methods transform the data in various ways, such as standardization or taking the logarithm, so we transformed the data back for some models before calculating the RMSE to keep the measure consistent and comparable between methods. Specifically, for the ARIMA GARCH method, we had to exponentiate the predicted values before calculating the RMSE. For LSTM method, the prediction was multiplied by the standard deviation(sigma) of the training data, and the mean of the training data was added to it. Once they were on the same scale, we compared the predictions of ARIMA GARCH models to LSTM models under univariate and multivariate conditions. Given that LSTM model parameterization is flexible, we chose to run several different models with different features, namely hidden units and prediction days, to find the optimal model.

In addition, we also tried to use another transformation to predict the price. We calculated the price change for each stock and treated the 30th and 70th percentiles as the thresholds. The price change larger than 70% and smaller than 30% were converted into 1 and -1 respectively, whereas the middle 40% were coded as 0. Under this transformation, we tried to investigate the price change in a way that could lend itself intuitively to classification via LSTM.

For the rejected models, we attempted a variety of methods in R such as LSTM and ARIMA models. Taking AMAZON as an example in LSTM models, first we transformed the data by getting the difference between two consecutive values in the series and lagged the series by having the value at time (t-1) as the input and value at time t as the output, which leads to a 1-step lagged dataset. The whole data was split so that the first 70% of the series as training set and the remaining 30% as test set. Then we normalized the train data to [0, 1], which is the range of the activation function sigmoid. In this part, the statistics (max and min values) of the training data are the scaling coefficients used to scale both the training and testing data as well as the predicted values in order to keep the data consistent. The batch size was set as 1 and only one previous day was used as the feature. The MSE and Adam method were specified as the loss function and the optimization algorithm respectively. After setting the learning rate as well as the learning rate decay over each update, the accuracy was considered as the metric to assess the model performance. The model predicts so well that the error rate is no more than 1%. However, this model has an obvious flaw. Considering we only used 1 previous point to form a 1-lagged dataset, the prediction of the model looks like a lagged version of the initial curve. For long term prediction, this model would make a huge underestimation of the difference between each point and not perform well as a result. More points need to be taken into consideration.

We also fit a Vector Auto Regression model to our data as well. Here we set the parameter as three. However, the VAR(3) model performed poorly when predicting stock values. In fact, we can see from the acf/pacf plot of residuals, prediction graph, and sample autocorrelation that VAR has done a poor job of fitting our data and that its predictions are essentially white noise (Appendix: Fig. 3). Thus the VAR model was quickly rejected as a multivariate model of interest for this kind of data.

The Description of Chosen Methods and Conclusions:

To compare different models we selected Root Mean Squared Error (RMSE) as the best metric to compare the different models. Because the data has been split into 16 years, 16 RMSE were generated, one for each year. Then the 16 numbers were converted to 1 mean RMSE by squaring each of the 16 RMSE, taking the average of the squared RMSE, and then taking the square root of the average. For model comparisons see Table 1.

First we compared the models by ARMA GARCH and LSTM with 3 previous days, 50 hidden units and 126 prediction days. For the ARMA GARCH model, we selected AICC as the criterion to deal with the condition that training data from different years might lead to different optimal parameters. In this case, we tried models ARIMA(1,1,1), ARIMA(2,1,2), ARIMA(3,1,3) and ultimately chose model ARIMA(2,1,2) ARCH(5) to calculate the RMSE based on the lowest AICC value. For LSTM model, as a Neural Network method, it typically outperforms series models with the longer memories, but one drawback it has is that it needs larger dataset to train than other models. When running LSTM we mainly focused on differences between LSTM using YrByYr or CumYr as the training data, LSTM trained on univariate or multivariate stocks, and LSTM with different feature number and hidden unit parameters.

When comparing and predicting 126 new prices from the data, we observed that these models perform differently depending on various industries. Cumulative year ARIMA GARCH performs best for the Consumer Discretionary industry (appendix Fig. 5), YrByYr multivariate LSTM performs best for Health Care (appendix Fig. 4), YrByYr univariate LSTM performs best for industrials, and YrByYr ARIMA GARCH performs best for financials; however, if AIG, whose stock prices tanked massively in 2008 financial crisis and never recovered (see Fig. 1), is not considered, cumulative ARIMA GARCH performs better.

This summary demonstrates that we could not find a general model that performs best over all the industrials. It makes sense because in practice, different models should be considered for different conditions. For example, according to the mean RMSE, we conclude that the univariate method cumulative year ARIMA GARCH performs best for the Consumer Discretionary industry. The reason may be that the CD industry stocks take on a wide range of large prices (the range of CD1- AMZN is from 0 to 800, the range of CD3 - DIS and CD9 - MCD are from 0 to 150). Moreover, although there are small fluctuations, each stock exhibits an overall tendency to increase significantly. Taking these characteristics into consideration, the cumulative year ARIMA GARCH model can deal with this kind of condition best.

In addition, it is easy to find different types of industries have various, differing characteristics. For the Financial industry, the ranges of different stocks also vary widely (the range of Fin1 - AIG is from 0 to 1500, the range of Fin3 - AXP is from 0 to 100, and the range of Fin13 - USB is from 0 to 60). However, the main difference between Fin and CD is that the stock price of Fin fluctuates more frequently and drastically. Under some extreme condition, we even need to consider the worldwide incident such as the 2008 economic crisis (e.g. the stock price of AIG is strongly influenced by the economic crisis).

Healthcare and Industrials stock prices cover much smaller ranges than CD and Fin, but also experience frequent and sometimes drastic fluctuations. Also note, these two industries are strongly influenced by public events and societal developments. These differences between industries suggest that we should apply different models to different industries.

Table 1. RMSE for 126 days prediction

RMSE	ARIMA GARCH	ARIMA GARCH	LSTM	LSTM	LSTM	LSTM
Yr.type	YrByYr	CumYr	YrByYr	CumYr	YrByYr	CumYr
Single or multi	univar	univar	univar	univar	multi	multi
Num.features	-	-	3	3	3	3
Hidden units	-	-	50	50	50	50
Predict.days	126	126	126	126	126	126
CD1 - AMZN	46.68	34.63	65.98	31.99	46.73	46.93
CD3 - DIS	4.10	3.72	6.19	6.63	5.75	6.54
CD9 - MCD	5.67	3.72	4.53	4.82	4.50	6.02
Fin1 – AIG	99.18	120.94	133.17	130.57	111.62	195.93
Fin3 - AXP	5.62	4.50	6.64	6.58	6.13	10.01
Fin13 - USB	2.01	1.68	2.44	2.18	2.00	4.33
Hcare10 - LLY	4.25	5.27	5.28	6.37	4.42	5.72
Hcare11 - MDT	5.04	3.60	4.45	4.99	4.01	5.14
Hcare13 - PFE	2.44	2.02	2.27	2.40	1.98	2.47
Inds2 - CAT	8.95	8.39	7.31	7.29	8.13	10.02
Inds3 - EMR	3.73	3.26	2.97	3.43	3.50	4.04
Inds11 - UNP	10.15	6.70	6.12	6.53	6.49	6.96
Average - CD	18.82	14.02	25.57	14.48	18.99	19.83
Average - Fin	35.60	42.37	47.42	46.44	39.92	70.09
Average- Fin	3.82	3.09	4.54	4.38	4.07	7.17

(without AIG)						
Average - Hcar	3.91	3.63	4	4.59	3.47	4.44
Average - Inds	7.61	6.13	5.47	5.75	6.04	7.01
Average - all	16.49	16.54	20.61	17.82	17.11	25.34

Because the ranges of these four industries differ significantly, we attempted to transform the data in a uniform way. We calculated the price changes for each stock and sorted them to get the 30% and 70% quantile (Fig.2), after which the lowest 30% decreasing price changes and highest 30% increasing price changes were converted to -1 and 1 respectively, and the price changes in the middle 40% were set to 0. Therefore, each price stock change was transformed to 30% -1, 40% 0, and 30% 1. This new data set was one observation smaller than the original as price change for the first observation could not be assessed.

Fig. 1 Stock Price of Financial 1- AIG

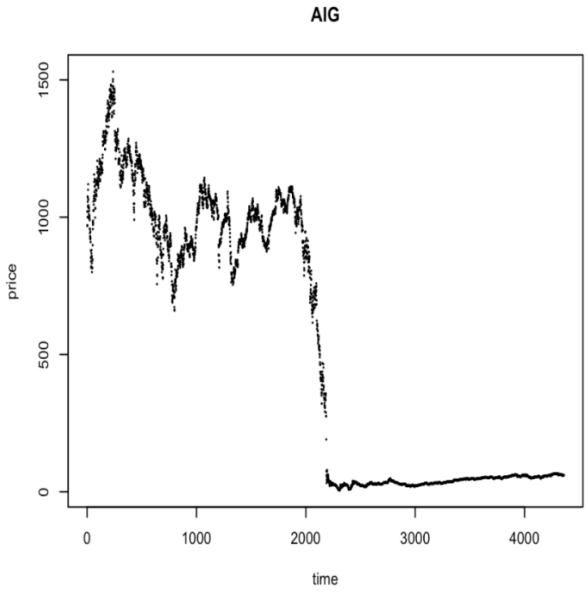

Fig. 2 The quantile of price change

```
> quan.price.change
             30%       70%
AMZN -0.7699912 0.8699958
DIS  -0.2107514 0.2430732
MCD  -0.1904190 0.2390234
AIG  -0.8948352 0.7635110
AXP  -0.2884540 0.3369848
USB  -0.1370578 0.1559474
LLY  -0.2319750 0.2483210
MDT  -0.2366608 0.2819670
PFE  -0.1111930 0.1225116
CAT  -0.3030986 0.3570118
EMR  -0.1985236 0.2181260
UNP  -0.1482940 0.1900094
```

We still used the same LSTM model with three previous days and 50 hidden points under univariate condition to predict 126 following days. Considering the prediction values were not exactly -1, 0 and 1, we set -0.2 and 0.2 as thresholds (according to Fig. 2, most of the quantiles distribute at -0.2 and 0.2) so that the prediction values larger than 0.2 and smaller than -0.2 were set to 1 and -1 respectively, whereas the values between -0.2 and 0.2 were set as 0. Then we compared the prediction values with the test data by using the number of (-1 to -1), (0 to 0) and (1 to 1) matches. Taking CD3 - DIS and Hcare13 - PFE as examples, the vast majority of our prediction values were 0 and the classification accuracy was only about 40%. One potential explanation for this scenario is LSTM model shrinks each input and distributing most of them nearly at 0, significantly decreasing the accuracy of prediction.

Finally we failed to find a uniform way to fit a general model for the four industries. However, considering that ARIMA GARCH models perform better on testing small intervals as well as the high flexibility of the LSTM models, we attempted to predict 20 days (20 business days represent about a month) from the previous 16 years using ARIMA GARCH models (see Table 2 for 20 day prediction results). Additionally, in the previous analysis, multivariate LSTM often outperformed univariate LSTM, so we also predicted on 20 days with multivariate LSTM models tuned with different previous days and hidden points parameters. Additionally, controlling variates is important to guarantee the effectiveness of the comparison, so for each LSTM only one parameter was changed from the previous LSTM model.

Table 2. RMSE for 20 day predictions

RMSE	ARIMA GARCH	ARIMA GARCH	LSTM	LSTM	LSTM	LSTM	LSTM	LSTM
Yr.type	YrByYr	CumYr	YrByYr	CumYr	YrByYr	CumYr	YrByYr	CumYr
uni or multi	univar	univar	multi	multi	multi	multi	multi	multi
Num.features	-	-	3	3	3	3	5	5
Hidden units	-	-	50	50	100	100	50	50
Predict.days	20	20	20	20	20	20	20	20
CD1- AMZN	12.31	10.49	17.85	22.03	17.26	26.79	19.67	17.03
CD3 - DIS	2.75	2.97	2.87	3.93	3.53	3.87	3.16	3.76
CD9 - MCD	1.90	1.58	1.85	2.72	1.89	2.22	2.17	2.08
Fin1 - AIG	55.30	55.60	71.92	79.93	74.82	81.33	72.3	80.88
Fin3 - AXP	1.93	2.02	3.24	2.95	2.82	3.81	3.30	2.92
Fin13 - USB	0.92	0.92	1.17	1.38	1.21	1.56	1.22	1.31
Hcare10 - LLY	2.07	2.07	1.85	2.16	2.34	1.93	2.17	2.38
Hcare11-MDT	1.60	1.63	1.80	2.02	1.92	1.92	1.78	1.82
Hcare13-PFE	1.07	1.10	1.22	1.13	1.14	1.06	1.13	1.29

Inds2 - CAT	2.95	2.89	3.98	3.39	3.88	3.76	3.98	4.07
Inds3 - EMR	1.47	1.37	2.09	1.58	1.82	1.49	1.88	1.93
Inds11 - UNP	2.32	1.48	1.73	2.58	1.65	2.11	1.71	2.83
Average -CD	5.66	5.01	7.52	9.56	7.56	10.96	8.33	7.62
Average - Fin	19.38	19.52	25.44	28.09	26.28	28.9	25.61	28.37
Average - Fin (without AIG)	1.43	1.47	2.21	2.17	2.02	2.69	2.26	2.12
Average-Hcar	1.58	1.60	1.62	1.77	1.80	1.64	1.69	1.83
Average-Inds	2.25	1.91	2.6	2.52	2.45	2.45	2.52	2.94
Average - all	7.22	7.01	9.30	10.48	9.52	10.99	9.54	10.19

When predicting 20 new stock prices from the data, ARIMA GARCH always outperforms LSTM, which makes sense as time series models typically predict well for the first several time points, but perform worse as time points extend further into the future. It is possible that a more finely tuned LSTM could outperform ARIMA GARCH, but given the above data we are comfortable saying ARIMA GARCH models perform the best. YrByYr ARIMA GARCH was optimal for Fin and HCare, and CumYear ARIMA GARCH was best for CD and Inds.

Our findings, support that for long-term prediction of price data, no one model is best. Depending on the industry, different ARIMA GARCH and LSTM, with varying modifications and parameters, had the greatest predictive power, which reflects the diverse and complicated nature of stock price data. We also found that ARIMA GARCH methods performed better than LSTM in predicting short-term price data (20 days). While we acknowledge that it is possible that a more finely tuned LSTM model could potentially outperform the aforementioned ARIMA GARCH models, ultimately our results support using a more diverse set of models for predicting stock price data. Finally, we found that the stock price data was resistant to our attempts to treat price changes as a classification problem, and the LSTM model was not well suited to classifying new stock price data as significantly increasing or decreasing.

Appendix Figures:

Fig3. Var(3) YrByYr CD 1,3,9 acf/pacf, standardized residuals, conditional variance

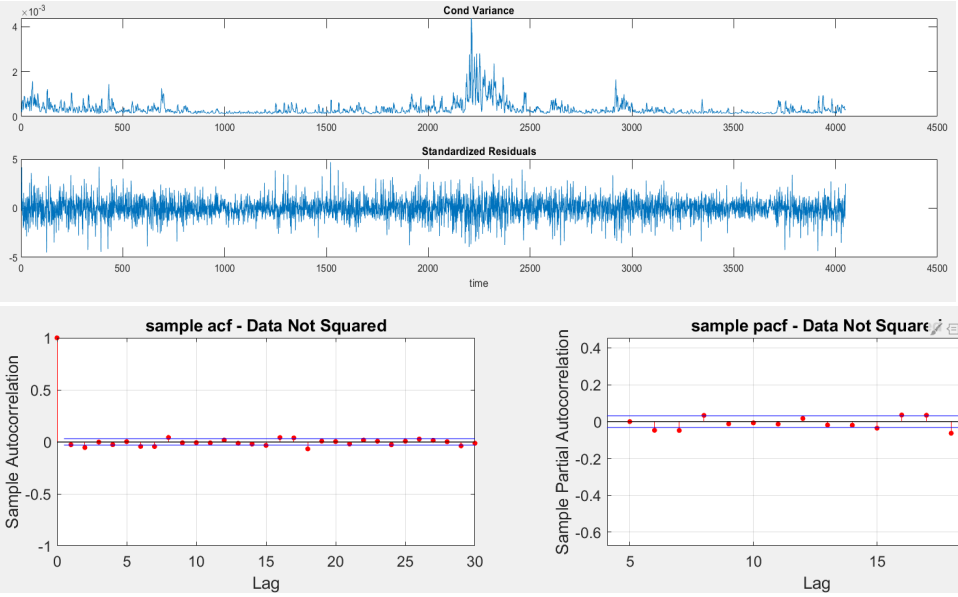

Fig3 continued. Var(3) YrByYr CD 1,3,9 Predictions

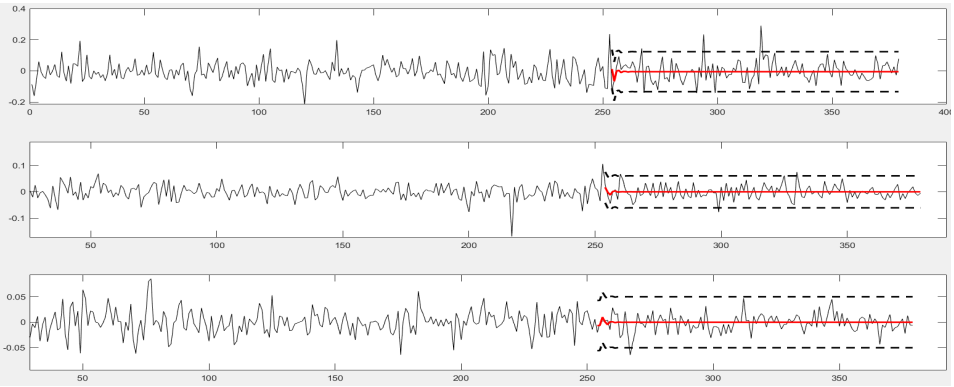

Fig.4 LTSM MTS YrByYr HCare 10,11,13 Predictions

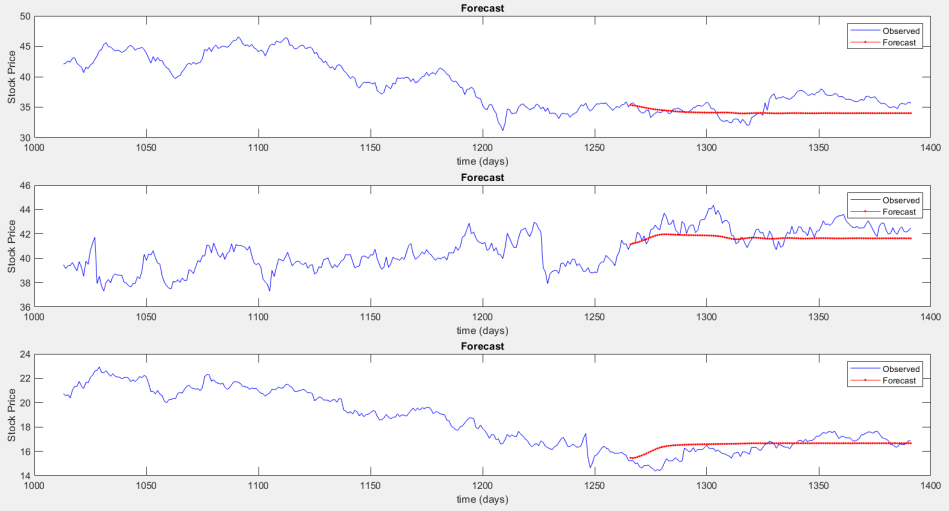

Fig. 5 Cumulative Yr ARIMA GARCH for CD 3 Disney Predictions