2021

Part I - Price Prediction for Cryptocurrencies with Machine Learning Methods

STUDENT PROJECT

Price Prediction for Cryptocurrencies with Machine Learning Methods

Student: Phan Nguyen Khoi[1]

Supervisor: Keith Wong[2], Raymond Tsang[3]

1. Introduction:

There have been a lot of research works on stock price prediction analysis, applying the techniques both from the classical statistical methods (e.g. [1], [2]) and the trendy machine learning methods (e.g. [3]), particularly those involving neural networks (e.g. [4], [5], [6]). With the arising market appetite on cryptocurrencies, it is of interest to revisit those machine learning methods on the digital assets. Driven by a practical interest in trading, this project aims at predicting the price of several popular cryptocurrencies based on machine learning methods. The effectiveness of Long Short-Term Memory (LSTM) was studied and compared with other traditional machine learning models, including Linear Regression, Decision Tree, SVM, Random Forest and Stochastic Gradient Boosting. Based on our numerical experiment, it turned out that linear regression appeared to outperform some sophisticated models including LSTM in most of the scenarios.

The rest of this report will be organized as follows. In Section 2, we will describe the two groups of data used in this project. The methodology used to evaluate and compare the performances of different models is discussed in section 3, while section 4 gives the results of our experiments. Section 5 concludes our experimentation and propose some ideas to extend our work.

2. Dataset:

Our numerical experiments were conducted in two phases. In Phase 1, we compared and analysed the performance of the machine learning methods on both stocks and cryptocurrencies. We want to make use of the stock movements as a control because price predictions on stock have been analysed in the past. And owing to the availability of the stock data in open sources such as Yahoo Finance, both the stock and cryptocurrency data are in daily frequency. Then we focus our analysis on cryptocurrencies in Phase 2. In this phase, the data frequency is 10 seconds. The details of the sets of data in both phases are as follows:

a. Phase 1:

[1] Department of System Engineering & Engineering Management, The Chinese University of Hong Kong

[2] Department of System Engineering & Engineering Management cum Centre for Financial Engineering, The Chinese University of Hong Kong

[3] Centre for Financial Engineering, The Chinese University of Hong Kong

In the first phase, we use daily data from 2 types of assets, namely cryptocurrencies and stocks. The latter was selected from three popular stock markets, The United States, Hong Kong, Shanghai:

Type of Asset	Number of Assets	List of Assets	Time Period
Cryptocurrencies	8	ADA-USD, BCH-USD, BNB-USD, BTC-USD, DOGE-USD, ETH-USD, LTC-USD, XRP-USD	2016 – 2021
US Stocks	8	AAPL, BRK-B, FB, GOOGL, JNJ, JPM, SPY, TSLA	2011 - 2021
Hong Kong Stocks	10	Tencent, Petrochina, CCB, ABC, AIA, ICBC, Ping An, CM Bank, Bank of China, CISCO-T	2011 - 2021
Shanghai Stocks	7	China Merchants Bank, Kweichow Moutai, Ping An Insurance, Industrial and Commercial Bank of China, China Life Insurance, Petrochina, China Construction Bank	2011 - 2021

In this phase, this daily data is utilized for its availability on Yahoo Finance, which helps us easily obtain a large amount of data. This also enables us to test the accuracy of multiple models without a tedious training process. Apart from cryptocurrencies, stocks were chosen to test the same set of models in order to verify our conclusion that is made based on cryptocurrency data.

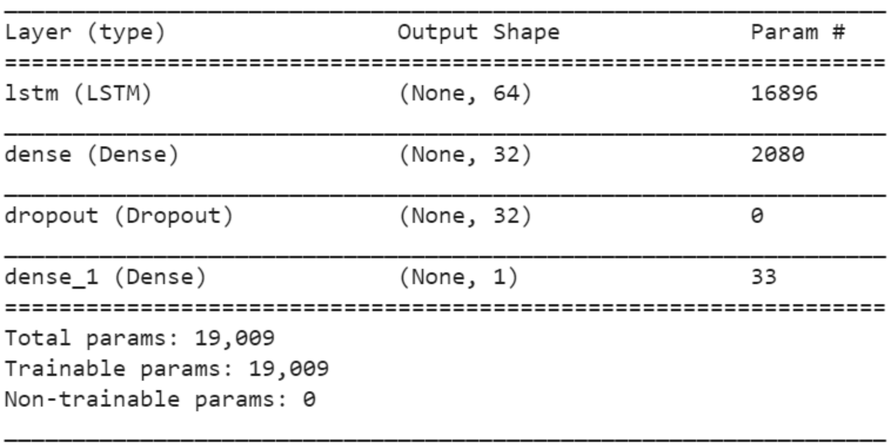
b. Phase 2:

In the second phase, we use cryptocurrencies' close price obtained from BitFinex with a time step of 10 seconds between 30/05/2021 and 13/06/2021. The dataset includes 10 assets with 63442 records each, namely ADA-USD, BTC-USD, DOG-USD, DOT-USD, ETH-USD, LINK-USD, LTC-USD, UNI-USD, UST-USD, YFI-USD. These cryptocurrencies were selected among the top market cap[4] at the start of the experiment and constrained by the availability in BitFinex. The time step of 10 seconds, which was our best available source, also helps us explore the potential advantage of applying machine learning for daily trading to most non-institutional investors. In practice, we understand that institutional traders may take advantage of obtaining higher frequency data for minimizing the latency.

3. Methodology:

In our experimentation, we test the LSTM model as well as 5 other Machine Learning models, including both single machine learning algorithms (Linear Regression, Decision Tree and SVM) and ensemble machine learning algorithms (Random Forest and Stochastic Gradient Boosting). Our LSTM model consists of an input layer, an LSTM layer, a hidden dense layer, a dropout layer and an output layer. The following figure gives some details of the LSTM model:

[4] For example, see https://coinmarketcap.com/



```
_________________________________________________________________
Layer (type)                 Output Shape              Param #   
=================================================================
lstm (LSTM)                  (None, 64)                16896     
_________________________________________________________________
dense (Dense)                (None, 32)                2080      
_________________________________________________________________
dropout (Dropout)            (None, 32)                0         
_________________________________________________________________
dense_1 (Dense)              (None, 1)                 33        
=================================================================
Total params: 19,009
Trainable params: 19,009
Non-trainable params: 0
_________________________________________________________________
```

Fig. 1. Architecture of LSTM model

In this project, two statistical quantitative indicators are utilized to evaluate and compare the performance of the LSTM model against the performances of other machine learning models. These metrics are: relative root mean square error (rRMSE) and mean absolute percentage error (MAPE). Since our models have multiple assets with different price ranges that are used for testing, the aforementioned metrics enable us to test the performances of these models in an accurate manner. Particularly, the rRMSE is calculated by dividing root mean square error (RMSE) by average value of measured data. Meanwhile, the MAPE is the average of the absolute percentage errors of forecasts, where error is defined as actual value minus forecasted value. For ease of comparison, both metrics are measured in percentage, which can be interpreted as following: The smaller the number (i.e., error), the better the model. In addition, the accuracy of predicting the trend of price movement (upward or downward) by each model is also discussed in our report although it is not a selecting criterion.

$$rRMSE = \sqrt{\frac{\sum_{t=1}^{n}(A_t - F_t)^2}{\sum_{t=1}^{n} A_t}}$$

$rRMSE =$ relative root mean square error
$n =$ number of observations
$A_t =$ actual value
$F_t =$ forecasted value

Fig. 2-1. Relative Root Mean Square Error

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right|$$

$MAPE =$ mean absolute percentage error
$n =$ number of observations
$A_t =$ actual value
$F_t =$ forecasted value

Fig. 2-2. Mean Absolute Percentage Error

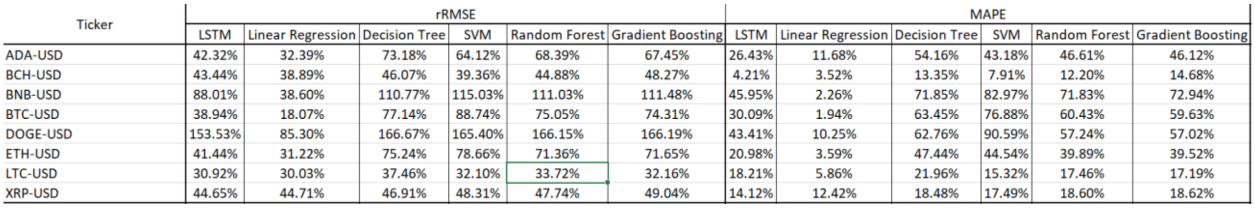
Using the two metrics above as selecting criteria, we conduct our experimentation in two phases as follows:

- In phase 1, we apply all 6 models on the first set of data mentioned in Section 2. For each asset, we train each model using 90% of the data and report the testing result using the remaining 10%. After that, we shortlist some of the better models for Phase 2 using the above criteria.

- In phase 2, we repeat the training and testing process on the second set of data mentioned in Section 2, using only the shortlisted models based on the performance in phase 1. Finally, we compare their testing errors and trend prediction accuracies.

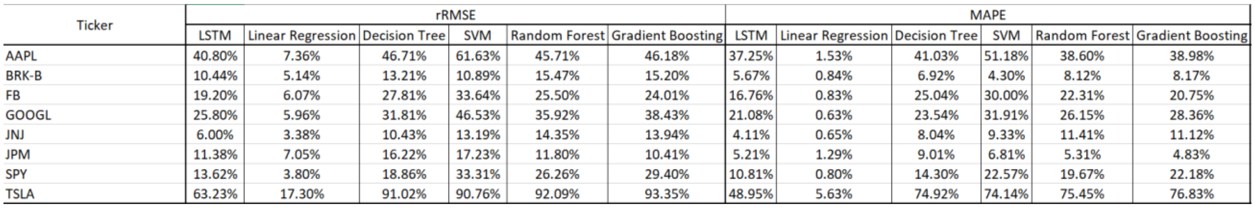
4. Numerical Results:

a. Phase 1:



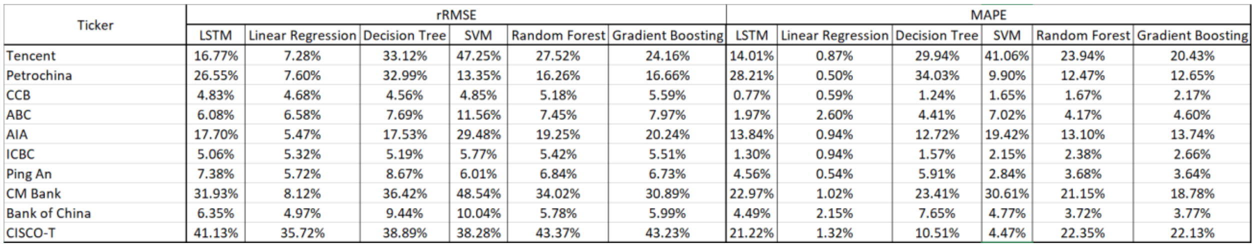
Ticker	rRMSE						MAPE					
	LSTM	Linear Regression	Decision Tree	SVM	Random Forest	Gradient Boosting	LSTM	Linear Regression	Decision Tree	SVM	Random Forest	Gradient Boosting
ADA-USD	42.32%	32.39%	73.18%	64.12%	68.39%	67.45%	26.43%	11.68%	54.16%	43.18%	46.61%	46.12%
BCH-USD	43.44%	38.89%	46.07%	39.36%	44.88%	48.27%	4.21%	3.52%	13.35%	7.91%	12.20%	14.68%
BNB-USD	88.01%	38.60%	110.77%	115.03%	111.03%	111.48%	45.95%	2.26%	71.85%	82.97%	71.83%	72.94%
BTC-USD	38.94%	18.07%	77.14%	88.74%	75.05%	74.31%	30.09%	1.94%	63.45%	76.88%	60.43%	59.63%
DOGE-USD	153.53%	85.30%	166.67%	165.40%	166.15%	166.19%	43.41%	10.25%	62.76%	90.59%	57.24%	57.02%
ETH-USD	41.44%	31.22%	75.24%	78.66%	71.36%	71.65%	20.98%	3.59%	47.44%	44.54%	39.89%	39.52%
LTC-USD	30.92%	30.03%	37.46%	32.10%	33.72%	32.16%	18.21%	5.86%	21.96%	15.32%	17.46%	17.19%
XRP-USD	44.65%	44.71%	46.91%	48.31%	47.74%	49.04%	14.12%	12.42%	18.48%	17.49%	18.60%	18.62%

Fig. 3-1. Test errors on cryptocurrencies



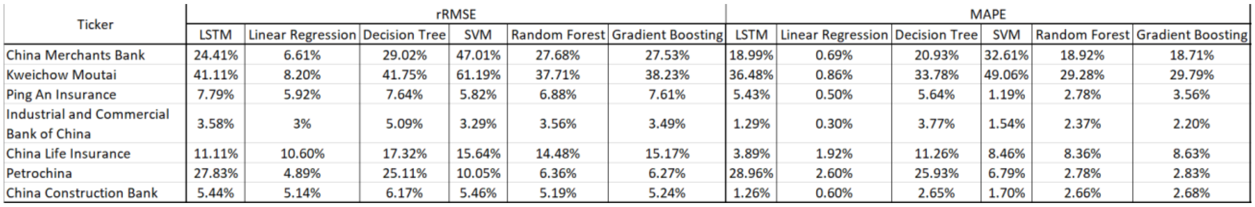
Ticker	rRMSE						MAPE					
	LSTM	Linear Regression	Decision Tree	SVM	Random Forest	Gradient Boosting	LSTM	Linear Regression	Decision Tree	SVM	Random Forest	Gradient Boosting
AAPL	40.80%	7.36%	46.71%	61.63%	45.71%	46.18%	37.25%	1.53%	41.03%	51.18%	38.60%	38.98%
BRK-B	10.44%	5.14%	13.21%	10.89%	15.47%	15.20%	5.67%	0.84%	6.92%	4.30%	8.12%	8.17%
FB	19.20%	6.07%	27.81%	33.64%	25.50%	24.01%	16.76%	0.83%	25.04%	30.00%	22.31%	20.75%
GOOGL	25.80%	5.96%	31.81%	46.53%	35.92%	38.43%	21.08%	0.63%	23.54%	31.91%	26.15%	28.36%
JNJ	6.00%	3.38%	10.43%	13.19%	14.35%	13.94%	4.11%	0.65%	8.04%	9.33%	11.41%	11.12%
JPM	11.38%	7.05%	16.22%	17.23%	11.80%	10.41%	5.21%	1.29%	9.01%	6.81%	5.31%	4.83%
SPY	13.62%	3.80%	18.86%	33.31%	26.26%	29.40%	10.81%	0.80%	14.30%	22.57%	19.67%	22.18%
TSLA	63.23%	17.30%	91.02%	90.76%	92.09%	93.35%	48.95%	5.63%	74.92%	74.14%	75.45%	76.83%

Fig. 3-2. Test errors on US stocks



Ticker	rRMSE						MAPE					
	LSTM	Linear Regression	Decision Tree	SVM	Random Forest	Gradient Boosting	LSTM	Linear Regression	Decision Tree	SVM	Random Forest	Gradient Boosting
Tencent	16.77%	7.28%	33.12%	47.25%	27.52%	24.16%	14.01%	0.87%	29.94%	41.06%	23.94%	20.43%
Petrochina	26.55%	7.60%	32.99%	13.35%	16.26%	16.66%	28.21%	0.50%	34.03%	9.90%	12.47%	12.65%
CCB	4.83%	4.68%	4.56%	4.85%	5.18%	5.59%	0.77%	0.59%	1.24%	1.65%	1.67%	2.17%
ABC	6.08%	6.58%	7.69%	11.56%	7.45%	7.97%	1.97%	2.60%	4.41%	7.02%	4.17%	4.60%
AIA	17.70%	5.47%	17.53%	29.48%	19.25%	20.24%	13.84%	0.94%	12.72%	19.42%	13.10%	13.74%
ICBC	5.06%	5.32%	5.19%	5.77%	5.42%	5.51%	1.30%	0.94%	1.57%	2.15%	2.38%	2.66%
Ping An	7.38%	5.72%	8.67%	6.01%	6.84%	6.73%	4.56%	0.54%	5.91%	2.84%	3.68%	3.64%
CM Bank	31.93%	8.12%	36.42%	48.54%	34.02%	30.89%	22.97%	1.02%	23.41%	30.61%	21.15%	18.78%
Bank of China	6.35%	4.97%	9.44%	10.04%	5.78%	5.99%	4.49%	2.15%	7.65%	4.77%	3.72%	3.77%
CISCO-T	41.13%	35.72%	38.89%	38.28%	43.37%	43.23%	21.22%	1.32%	10.51%	4.47%	22.35%	22.13%

Fig. 3-3. Test errors on Hong Kong stocks

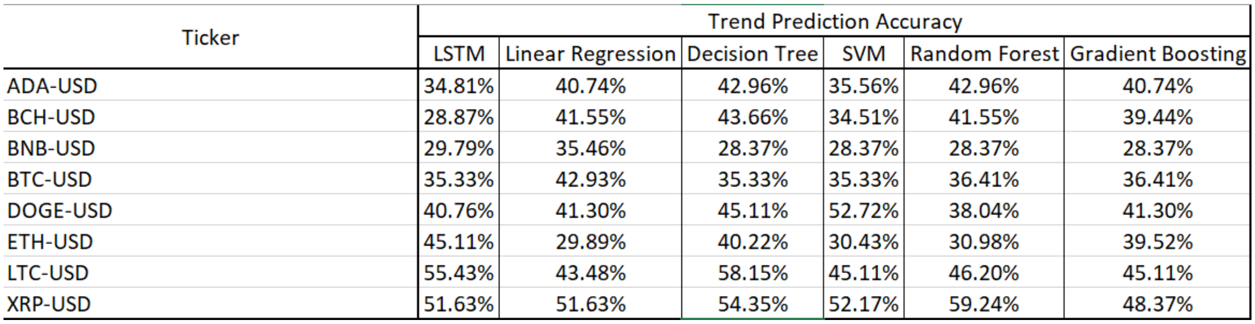


Ticker	rRMSE						MAPE					
	LSTM	Linear Regression	Decision Tree	SVM	Random Forest	Gradient Boosting	LSTM	Linear Regression	Decision Tree	SVM	Random Forest	Gradient Boosting
China Merchants Bank	24.41%	6.61%	29.02%	47.01%	27.68%	27.53%	18.99%	0.69%	20.93%	32.61%	18.92%	18.71%
Kweichow Moutai	41.11%	8.20%	41.75%	61.19%	37.71%	38.23%	36.48%	0.86%	33.78%	49.06%	29.28%	29.79%
Ping An Insurance	7.79%	5.92%	7.64%	5.82%	6.88%	7.61%	5.43%	0.50%	5.64%	1.19%	2.78%	3.56%
Industrial and Commercial Bank of China	3.58%	3%	5.09%	3.29%	3.56%	3.49%	1.29%	0.30%	3.77%	1.54%	2.37%	2.20%
China Life Insurance	11.11%	10.60%	17.32%	15.64%	14.48%	15.17%	3.89%	1.92%	11.26%	8.46%	8.36%	8.63%
Petrochina	27.83%	4.89%	25.11%	10.05%	6.36%	6.27%	28.96%	2.60%	25.93%	6.79%	2.78%	2.83%
China Construction Bank	5.44%	5.14%	6.17%	5.46%	5.19%	5.24%	1.26%	0.60%	2.65%	1.70%	2.66%	2.68%

Fig 3-4. Test errors on Shanghai stocks

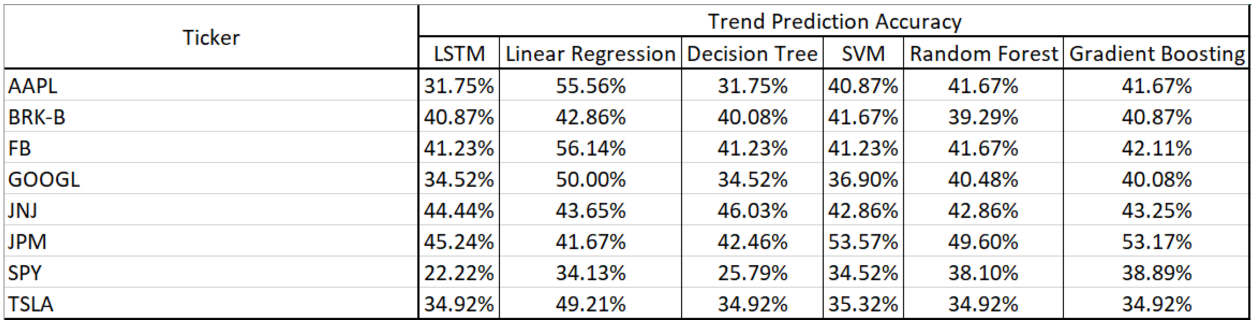
In terms of rRMSE and MAPE, as it is shown from Figure 3-1, 3-2, 3-3 and 3-4, Linear Regression, the simplest model, achieved the lowest error (generally below 10% MAPE and below 30% rRMSE) among 6 models for most of the assets, while LSTM is generally the second-best model for three groups of assets including cryptocurrencies, US stocks and Hong Kong stocks, as its MAPE was often below 20%. On the other hand, all other Machine Learning models did not perform as consistently well as the two aforementioned models did. Indeed, Decision Tree, SVM and Random Forest are in top 3 for at most 2 groups of assets. For example, the error of SVM was often in top 3 regarding cryptocurrencies, while Decision

Tree's error was in top 3 for US stocks. It is also noteworthy that Linear Regression's error was significantly lower than others most of the time, which implies that Linear Regression outperformed all other models. For instance, its MAPE was below 3%, which was the smallest among the Hong Kong stock market.



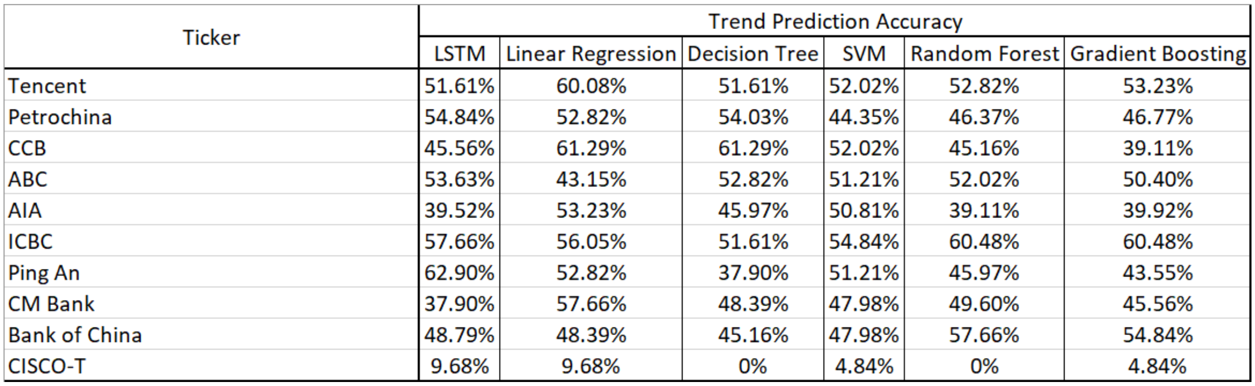
Ticker	Trend Prediction Accuracy					
	LSTM	Linear Regression	Decision Tree	SVM	Random Forest	Gradient Boosting
ADA-USD	34.81%	40.74%	42.96%	35.56%	42.96%	40.74%
BCH-USD	28.87%	41.55%	43.66%	34.51%	41.55%	39.44%
BNB-USD	29.79%	35.46%	28.37%	28.37%	28.37%	28.37%
BTC-USD	35.33%	42.93%	35.33%	35.33%	36.41%	36.41%
DOGE-USD	40.76%	41.30%	45.11%	52.72%	38.04%	41.30%
ETH-USD	45.11%	29.89%	40.22%	30.43%	30.98%	39.52%
LTC-USD	55.43%	43.48%	58.15%	45.11%	46.20%	45.11%
XRP-USD	51.63%	51.63%	54.35%	52.17%	59.24%	48.37%

Fig. 4-1. Trend prediction accuracy on cryptocurrencies



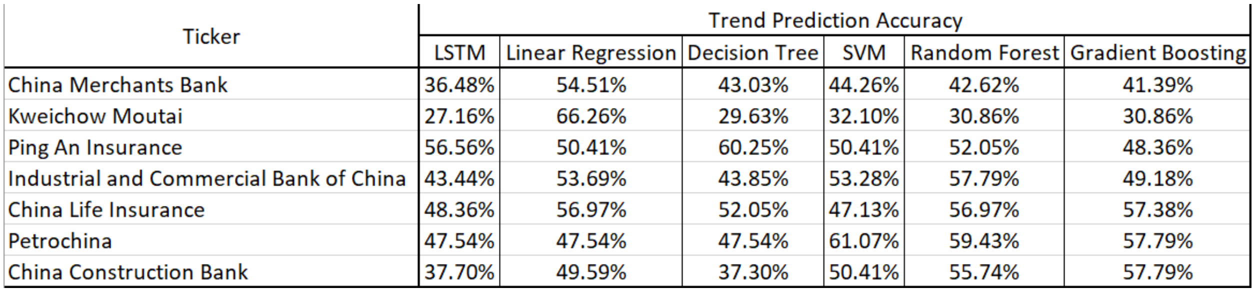
Ticker	Trend Prediction Accuracy					
	LSTM	Linear Regression	Decision Tree	SVM	Random Forest	Gradient Boosting
AAPL	31.75%	55.56%	31.75%	40.87%	41.67%	41.67%
BRK-B	40.87%	42.86%	40.08%	41.67%	39.29%	40.87%
FB	41.23%	56.14%	41.23%	41.23%	41.67%	42.11%
GOOGL	34.52%	50.00%	34.52%	36.90%	40.48%	40.08%
JNJ	44.44%	43.65%	46.03%	42.86%	42.86%	43.25%
JPM	45.24%	41.67%	42.46%	53.57%	49.60%	53.17%
SPY	22.22%	34.13%	25.79%	34.52%	38.10%	38.89%
TSLA	34.92%	49.21%	34.92%	35.32%	34.92%	34.92%

Fig. 4-2. Trend prediction accuracy on US stocks



Ticker	Trend Prediction Accuracy					
	LSTM	Linear Regression	Decision Tree	SVM	Random Forest	Gradient Boosting
Tencent	51.61%	60.08%	51.61%	52.02%	52.82%	53.23%
Petrochina	54.84%	52.82%	54.03%	44.35%	46.37%	46.77%
CCB	45.56%	61.29%	61.29%	52.02%	45.16%	39.11%
ABC	53.63%	43.15%	52.82%	51.21%	52.02%	50.40%
AIA	39.52%	53.23%	45.97%	50.81%	39.11%	39.92%
ICBC	57.66%	56.05%	51.61%	54.84%	60.48%	60.48%
Ping An	62.90%	52.82%	37.90%	51.21%	45.97%	43.55%
CM Bank	37.90%	57.66%	48.39%	47.98%	49.60%	45.56%
Bank of China	48.79%	48.39%	45.16%	47.98%	57.66%	54.84%
CISCO-T	9.68%	9.68%	0%	4.84%	0%	4.84%

Fig. 4-3. Trend prediction accuracy on Hong Kong stocks

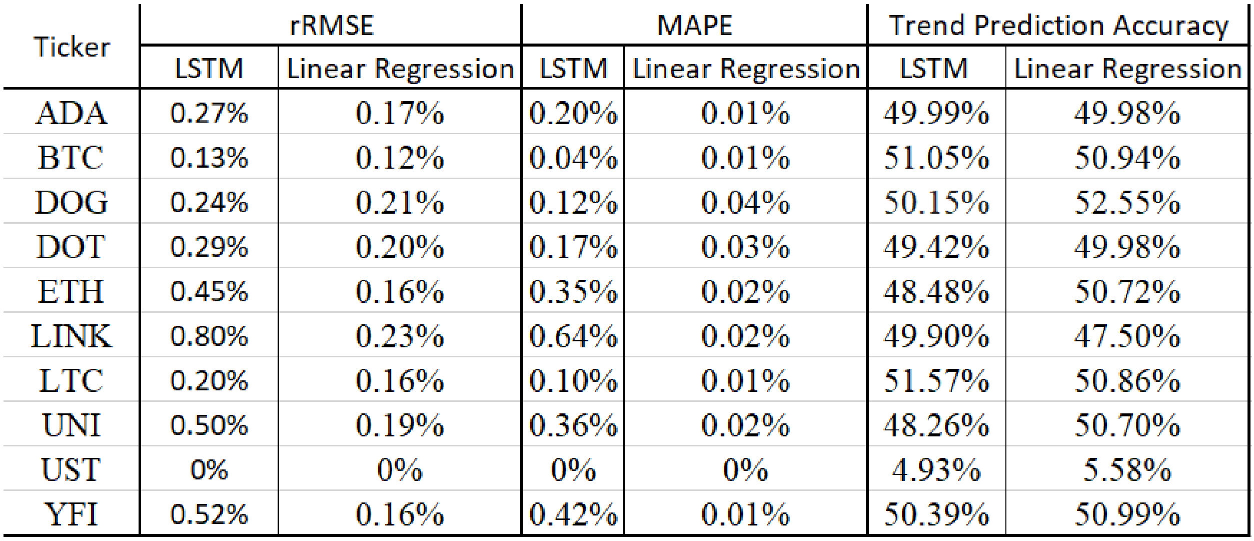


Ticker	Trend Prediction Accuracy					
	LSTM	Linear Regression	Decision Tree	SVM	Random Forest	Gradient Boosting
China Merchants Bank	36.48%	54.51%	43.03%	44.26%	42.62%	41.39%
Kweichow Moutai	27.16%	66.26%	29.63%	32.10%	30.86%	30.86%
Ping An Insurance	56.56%	50.41%	60.25%	50.41%	52.05%	48.36%
Industrial and Commercial Bank of China	43.44%	53.69%	43.85%	53.28%	57.79%	49.18%
China Life Insurance	48.36%	56.97%	52.05%	47.13%	56.97%	57.38%
Petrochina	47.54%	47.54%	47.54%	61.07%	59.43%	57.79%
China Construction Bank	37.70%	49.59%	37.30%	50.41%	55.74%	57.79%

Fig. 4-4. Trend prediction accuracy on Shanghai stocks

Regarding the accuracy of trend prediction, as we can see from Figure 4-1, 4-2, 4-3 and 4-4, while Linear Regression is still the best-performing model, LSTM has the worst accuracy. However, in many cases, all 6 models produced much lower trend prediction accuracies than 50%, which was the chance of a random pick.

b. <u>Phase 2:</u>



Ticker	rRMSE		MAPE		Trend Prediction Accuracy	
	LSTM	Linear Regression	LSTM	Linear Regression	LSTM	Linear Regression
ADA	0.27%	0.17%	0.20%	0.01%	49.99%	49.98%
BTC	0.13%	0.12%	0.04%	0.01%	51.05%	50.94%
DOG	0.24%	0.21%	0.12%	0.04%	50.15%	52.55%
DOT	0.29%	0.20%	0.17%	0.03%	49.42%	49.98%
ETH	0.45%	0.16%	0.35%	0.02%	48.48%	50.72%
LINK	0.80%	0.23%	0.64%	0.02%	49.90%	47.50%
LTC	0.20%	0.16%	0.10%	0.01%	51.57%	50.86%
UNI	0.50%	0.19%	0.36%	0.02%	48.26%	50.70%
UST	0%	0%	0%	0%	4.93%	5.58%
YFI	0.52%	0.16%	0.42%	0.01%	50.39%	50.99%

Fig. 5. LSTM vs Linear Regression

Compared to Phase 1, in this phase, both Linear Regression and LSTM achieved considerably small errors – below 1% in all cases – with the former model being the better one for all assets. This can possibly be explained by the smaller time step of this dataset (10 seconds versus 1 day): Given the same period, a smaller time step captured more information of the market than a larger time step (i.e. less frequent data). As a result, it was expected that the training of the models could be improved with the additional amount of information.

Another interesting observation from the table is that the accuracy of trend prediction by both models is approximately 50%, which is the probability of random guesses.

5. <u>Conclusion:</u>

In this report, we presented various sets of the data of different types of assets and trading frequencies. Multiple steps were conducted to train and evaluate all 6 models across many assets. After two phases of experiments, it is shown that Linear Regression provides the best performance, while LSTM is questionably the second-best model. Additionally, from Phase 2, it is suggested that for the data that was up to the frequency of 10 seconds, the probability of making a correct prediction of price movement is about as high as that of making a random guess.

An extension of our research can focus on the confidence level in each trend prediction made by each of these two models, especially LSTM. To calculate the confidence level of the LSTM model, we used the method proposed in [7]. During our research, we found that LSTM appeared to be overconfident of its prediction. For some of the cryptocurrencies, the confidence level in each prediction by LSTM was significantly higher than the accuracy that we see in Phase 2 of our experimentation, where the number produced by both models was around 50%. For example, there were multiple times that LSTM predicted a price movement with over 90% confidence in the next time step (see Figure 6-2 & Figure 6-3). The distribution of confidence level in each prediction made by LSTM in those cases are shown below, in which the x-axis represents the confidence level, and the y-axis represents the number of times such confidence level appeared in a trend prediction, as examples of such observation. For example, the highest bar in Fig 6-1 means that the model is 75% certain of its predictions in above 1600 times out of 6345 simulations.

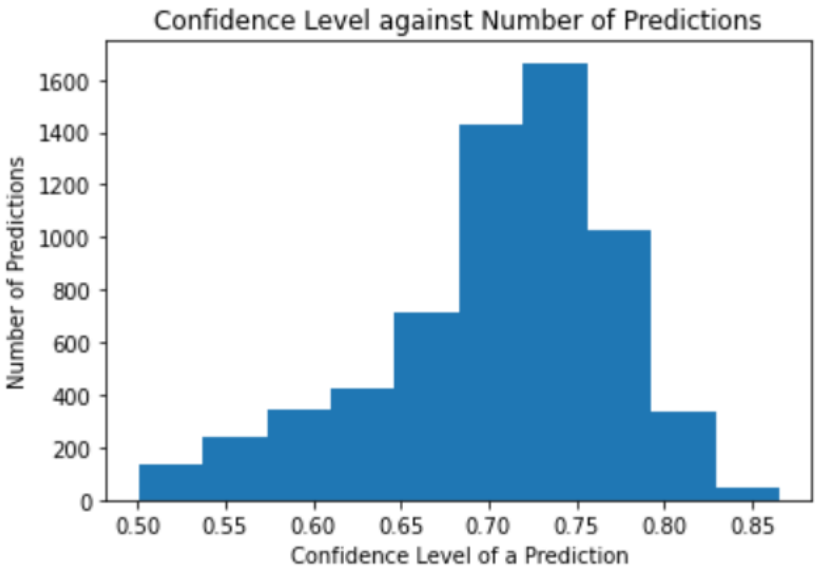


Fig. 6-1. BTC-USD

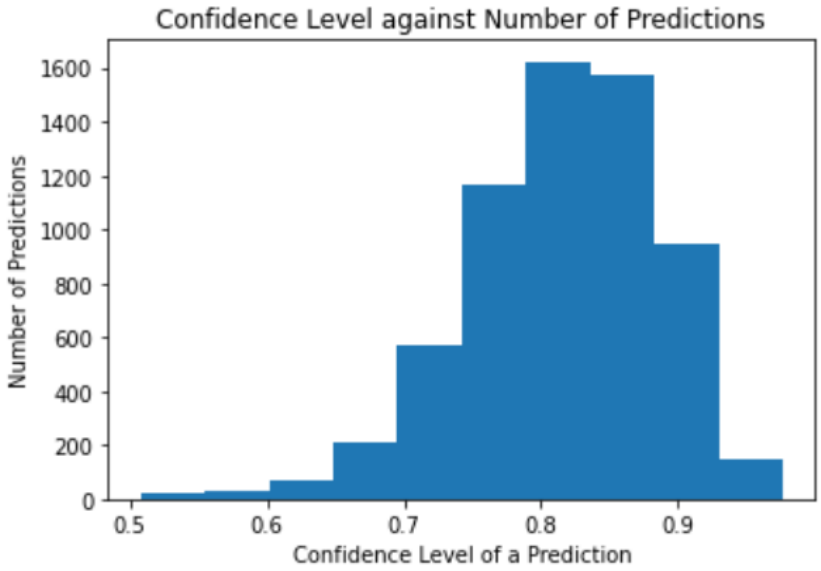


Fig. 6-2. DOG-USD

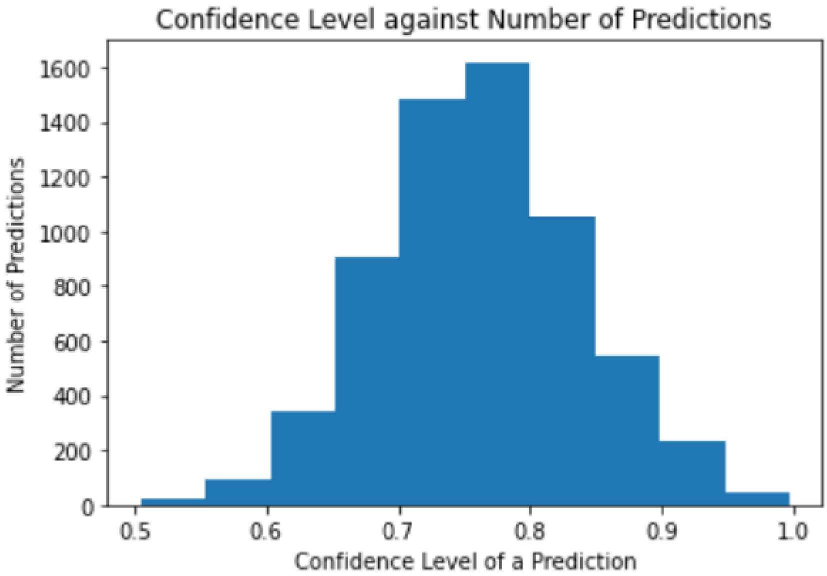


Fig. 6-3. ETH-USD

References

[1] A. A. Ariyo, A. O. Adewumi and C. K. Ayo, "Stock Price Prediction Using the ARIMA Model," in *2014 UKSim-AMSS 16th International Conference on Computer Modelling and Simulation*, Cambridge, UK, 2014.

[2] H. Herwartz, "Stock return prediction under GARCH — An empirical assessment," *International Journal of Forecasting,* vol. 33, no. 3, pp. 569-580, 2017.

[3] D. Zhang and L. Zhou, "Discovering Golden Nuggets: Data Mining in Financial Application," *IEEE Transactions on Systems Man and Cybernetics Part C (Applications and Reviews),* vol. 34, no. 4, pp. 513-522, 2004.

[4] E. DEZSI and I. A. NISTOR, "Can Deep Machine Learning Outsmart The Market? A Comparison Between Econometric Modelling And Long- Short Term Memory," *Romanian Economic Business Review,* vol. 11, no. 4.1, pp. 54-73, 2016.

[5] S. Liu, C. Zhang and J. Ma, "CNN-LSTM Neural Network Model for Quantitative Strategy Analysis in Stock Markets," in *International Conference on Neural Information Processing*, Guangzhou, China, 2017.

[6] S. Loukas, "Time-Series Forecasting: Predicting Stock Prices Using An LSTM Model," towards data science, 2020.

[7] L. Zhu and N. Laptev, "Deep and Confident Prediction for Time Series at Uber," in *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, New Orleans, LA, USA, 2017.