Academic Paper Reproduction and Improvements

Analysis of Bitcoin Price Prediction Using Machine Learning

Group 10: Yuhan Liu, Yangyige Qiu, Kaili Shi, Chuyi Wang, Yanran Xu
Professor: Steven Kou
Course: MF 740 Economics of Fintech

Abstract: This project expands on Junwei Chen's seminal work, "Analysis of Bitcoin Price Prediction Using Machine Learning," by reproducing and extending its methodology, which employs Random Forest Regression and Long Short-Term Memory (LSTM) models for predicting Bitcoin prices. Our study spans from March 31, 2015, to April 1, 2022, and collects 47 variables across diverse categories, including Bitcoin's market behavior, technical aspects, comparative cryptocurrency data, and broader economic indicators. We meticulously replicate Chen's methodology to establish a benchmark upon which we build by introducing new data preprocessing techniques and models to elevate predictive accuracy. Our replication aligns closely with the original findings, while our enhancements provide a more nuanced understanding of the Bitcoin market dynamics and improved predictive performance. Our enhanced XGBoost model demonstrates the best accuracy among all models and differential/log differential datasets. This paper contributes to the fintech field by deepening the understanding of Bitcoin price determinants and refining the approaches used in cryptocurrency market analysis.

Keywords: Bitcoin Price Prediction, Machine Learning, Random Forest Regression, Long Short-Term Memory (LSTM), Feature Analysis

Content

1 Paper Introduction

In the realm of financial technology, the prediction of Bitcoin prices using machine learning techniques has garnered significant interest. The original paper by Junwei Chen, "Analysis of Bitcoin Price Prediction Using Machine Learning" published in the Journal of Risk and Financial Management, serves as a basic study in this domain. Chen's work primarily focuses on leveraging Random Forest Regression and Long Short-Term Memory (LSTM) models to predict Bitcoin's future prices.

Our project aims to replicate and extend Chen's study, incorporating additional data and applying improved methodologies. We have meticulously replicated the original study's methodology, ensuring a thorough understanding and application of the techniques used. We compare the newly reproduced figures with the original figures and analyze their differences.

Building upon this foundation, we introduce several novel aspects to improve prediction accuracy. These include incorporating new models, a refined selection of variables, and an extended analysis period that considers recent market trends and events. Our modifications are designed to provide a more comprehensive understanding of the factors influencing Bitcoin prices and enhance the models' predictive performance.

The structure of our project paper mirrors the logical flow of our research process. It begins with a detailed data analysis, followed by an exposition of our replication of the original study. We then present our improvements, detailing the methodologies and rationale behind each. The paper concludes with a comparative analysis of the original research and our enhanced models, offering insights into the advancements we have achieved.

In summary, our project not only reaffirms the original study's findings but also pushes the boundaries of Bitcoin price prediction research by introducing refined techniques and a broader analytical perspective. Our work contributes to the growing body of knowledge in financial technology and offers practical insights for those interested in the dynamics of cryptocurrency markets.

Data Introduction

Our study utilizes an extensive dataset from March 31, 2015, to April 1, 2022. This period captures significant events and transitions in the cryptocurrency market, including the emergence of Bitcoin Cash, the 2017 price bubble, and the recent surge in Bitcoin's value, surpassing $40,000 for the first time. The comprehensive dataset allows for a deep dive into the factors influencing Bitcoin's price over different market phases.

We collected forty-seven variables across eight categories, aligning with the original study's framework. These categories include Bitcoin price variables (such as open, high, low, and close prices), specific technical features of Bitcoin, other cryptocurrencies prices, commodities, market indices, foreign exchange rates, public attention metrics (including Google Trends and daily Tweets), and dummy variables representing weekdays.

In processing this data, we emphasized ensuring accuracy and consistency. The data were sourced from reliable financial and public platforms, ensuring a high level of credibility and relevance. We paid particular attention to handling missing data, particularly for variables like ETH that were introduced later in the timeframe. Our approach was to fill missing values with the previous period's data, maintaining the continuity and integrity of our time series analysis. Here is the quantitative analysis of the original data:

	count	mean	std	min	max
BTC_Open	2559.0	1.262814e+04	1.668978e+04	2.100680e+02	6.754973e+04
BTC_High	2559.0	1.296549e+04	1.713374e+04	2.238330e+02	6.878962e+04
BTC_Low	2559.0	1.225905e+04	1.618448e+04	1.995670e+02	6.638206e+04
BTC_Close	2559.0	1.264427e+04	1.669706e+04	2.104950e+02	6.756683e+04
BTC_Volume	2559.0	1.601909e+10	2.024297e+10	1.060090e+07	3.510000e+11
Active Addr Cnt	2559.0	7.151230e+05	2.359796e+05	2.226280e+05	1.366494e+06
Xfer Cnt	2559.0	6.464933e+05	1.838259e+05	2.348060e+05	2.041653e+06
Mean Tx Size (native units)	2559.0	2.092273e+00	3.507530e+00	3.070392e-01	1.267199e+02
Total Fees (USD)	2559.0	9.367344e+05	1.971955e+06	2.850355e+03	2.139776e+07
Mean Hash Rate	2559.0	6.057145e+07	6.155013e+07	2.717381e+05	2.481103e+08
Difficulty	2559.0	8.371336e+12	8.497941e+12	4.671755e+10	2.860000e+13
Mean Block Size (in bytes)	2559.0	9.685166e+05	2.584561e+05	2.929293e+05	1.523656e+06
Sum Block Weight	2559.0	4.816130e+08	1.049458e+08	1.912665e+08	7.584308e+08
LTC	2559.0	7.187075e+01	7.081633e+01	1.321170e+00	3.864508e+02
XRP	2559.0	3.544873e-01	3.814098e-01	3.560000e-03	2.780000e+00
DASH	2559.0	1.421313e+02	1.824392e+02	2.060000e+00	1.550850e+03
DOGE	2559.0	3.587345e-02	8.775410e-02	8.730000e-05	6.848000e-01
ETH	2430.0	7.088693e+02	1.107578e+03	4.348000e-01	4.812090e+03
Gold	2559.0	1.489620e+03	2.453960e+02	1.070800e+03	2.117100e+03
Silver	2559.0	1.919696e+01	3.737286e+00	1.197800e+01	3.013500e+01
Copper	2559.0	2.998034e+00	6.934497e-01	1.994000e+00	4.937500e+00
Oil	2559.0	5.483713e+01	1.446541e+01	-3.763000e+01	1.237000e+02
Treasury Yield 10 Years	2559.0	1.951992e+00	6.558973e-01	4.990000e-01	3.234000e+00
S&P500	2559.0	2.905920e+03	7.789848e+02	1.829080e+03	4.796560e+03
DJI	2559.0	2.482381e+04	5.700070e+03	1.566018e+04	3.679965e+04

	count	mean	std	min	max
CBOE	2559.0	95.019945	21.613803	55.500000	137.160004
NASDAQ	2559.0	8334.481796	3307.996972	4266.839844	16057.440430
JP225	2559.0	22005.423200	3759.131988	14952.020000	30670.100000
CSI300	2559.0	3978.366663	670.747293	2853.760000	5807.720000
DXY	2559.0	95.650754	2.973578	88.589996	103.290001
EUR	2559.0	1.343499	0.088222	1.149439	1.588512
GBP	2559.0	0.747387	0.046783	0.629520	0.869990
JPY	2559.0	111.049275	5.123320	99.905998	125.628998
CAD	2559.0	1.303572	0.044419	1.195400	1.457800
AUD	2559.0	1.367293	0.072520	1.232000	1.741281
SGD	2559.0	1.367161	0.029391	1.306590	1.456300
CNY	2559.0	0.733341	0.037264	0.574290	0.811688
RUB	2559.0	66.558836	8.633415	0.716200	138.965103
Tweets	2559.0	50500.825322	43438.574634	13294.000000	363566.000000
Google	2559.0	495.820606	519.210217	64.000000	6064.503589

Additionally, recognizing the importance of data normalization in machine learning, we applied min-max scaling to transform all variables to a [0, 1] range. This step was crucial for our LSTM model, given its sensitivity to the scale of input data. This preprocessing step helps in mitigating the influence of disparate metric units and scales, allowing for a more accurate and fair comparison across different variables.

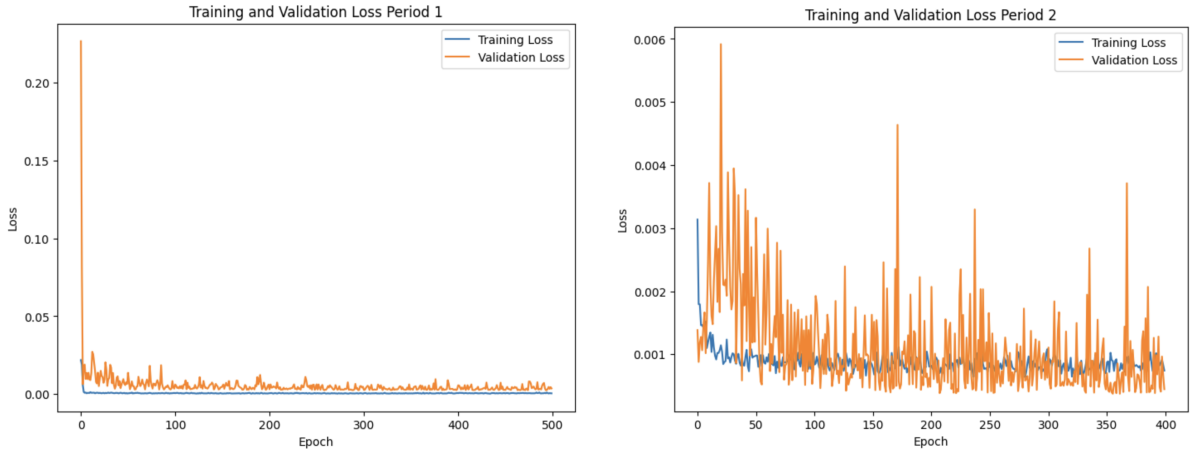
Through this meticulous data preparation process, we laid a strong foundation for our subsequent analysis, ensuring that our models are fed with high-quality, well-structured data. This, in turn, enhances the reliability of our findings and the robustness of our predictive models.

2 Paper Reproduction

In this part of our project, we present reproduction of the analysis by Junwei Chen on predicting Bitcoin prices using machine learning algorithms. Our meticulous replication process focused on the original paper's important plots. Here, we emphasize our replication fidelity by showcasing our reproduced figures that closely mimic the original results, and analyze the slight differences between them. Notably, we only demonstrate the newly created figures here, and put the original ones in Appendix.

Training and validation loss of LSTM

Reproduction of Figure 3:

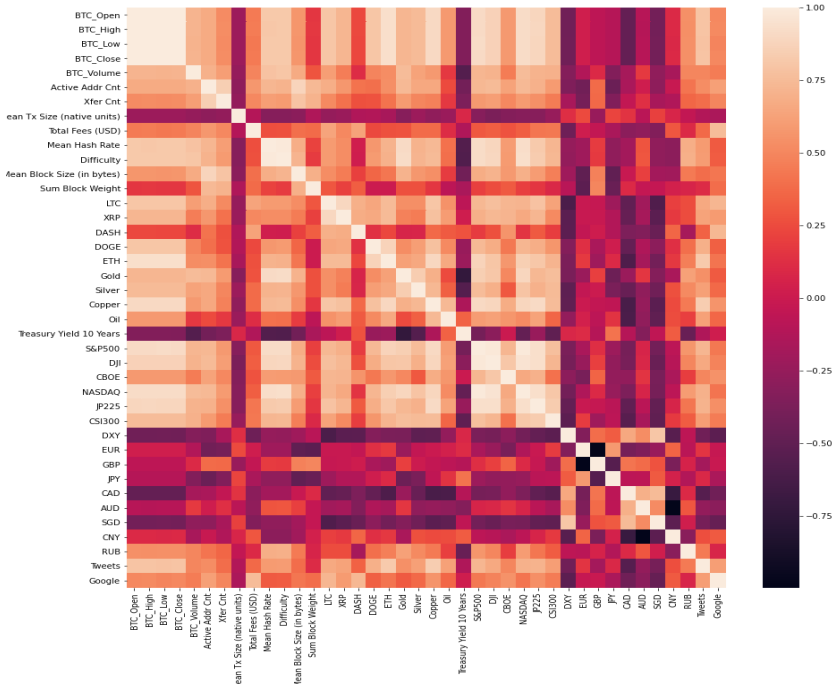


Our reproduced Figure 3, is quite similar to the original study's figure, delineates the training and validation loss across epochs for two distinct periods of Bitcoin price data. In Period 1, our reproduced results align closely with the original study, exhibiting a typical convergence of training and validation losses. This alignment reinforces the reliability of the LSTM model in learning from the training data and accurately generalizing to the validation set.

It is worth noting that a slight divergence emerges in the second period. Our plot reveals a downward trajectory in validation loss as the number of epochs increases, which diverges from the original plot. However, we assert the correctness of our reproduced results. We contend that our plot is an accurate reflection because, with a larger epoch count, the validation loss is expected to decrease, indicating that the model is effectively capturing the underlying patterns in the data without overfitting.

Variable analysis

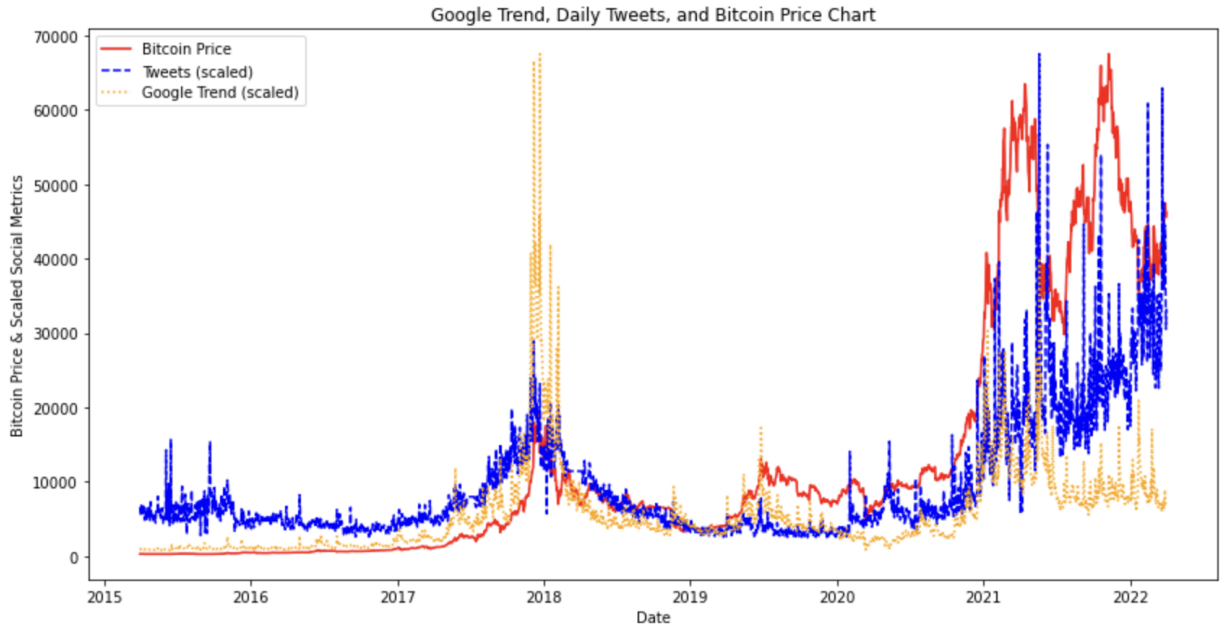
Reproduction of Figure 4



In Figure 4, we recreate a correlation heat map. The original map displays a matrix of correlations among various financial indicators and Bitcoin prices, revealing a diverse range of positive and negative interactions. Our replicated version, however, indicates subtle yet significant differences. While the overall patterns of correlation remain intact, the contrast in the intensity of certain correlations is evident. Notably, the negative correlation between Bitcoin and the 10-year U.S. Treasury yield appears more pronounced in our rendition. This heightened inverse relationship could suggest a stronger sentiment of Bitcoin being an alternative investment when traditional yields falter.

Google Trend, Daily Tweets and Bitcoin Price Prediction

Reproduction of Figure 5



This plot reveals significant patterns in public interest and Bitcoin's price, which aligns with the original plot. Firstly, the surge in Google searches and daily tweets coincides with the periods where Bitcoin reached unprecedented price peaks, indicating heightened public engagement during these pivotal moments. Secondly, the zenith of Google search interest for Bitcoin, observed at the end of 2017, remains unparalleled. Notably, even as Bitcoin prices soared above USD 60,000 in 2021, the volume of searches did not eclipse the high watermark set in 2017. This suggests that while the public's responsiveness to Bitcoin's price milestones remains robust, the peak curiosity or novelty factor—as captured by Google Trends—may have its own distinct lifecycle, separate from the currency's market valuation.

Error results for random forest regression

In order to reflect the result of each prediction model, we calculated three error results like paper, shown in Table 4. The first one is RMSE, Root Mean Squared Error. And the second one is MAPE, Mean Absolute Percentage Error. And the last one is DA, Decision Accuracy. These three errors will not only help us compare the result we got with that from the paper, but help us to compare the prediction from different models, especially those that appeared in our owo improvement. The following are the equations for each error.

$$MAPE = \frac{1}{m}\sum_{t=1}^{m}\left|\frac{y(t)-\hat{y}(t)}{y(t)}\right|$$

$$RMSE = \sqrt{\frac{1}{m}\sum_{t=1}^{m}(y(t)-\hat{y}(t))^2}$$

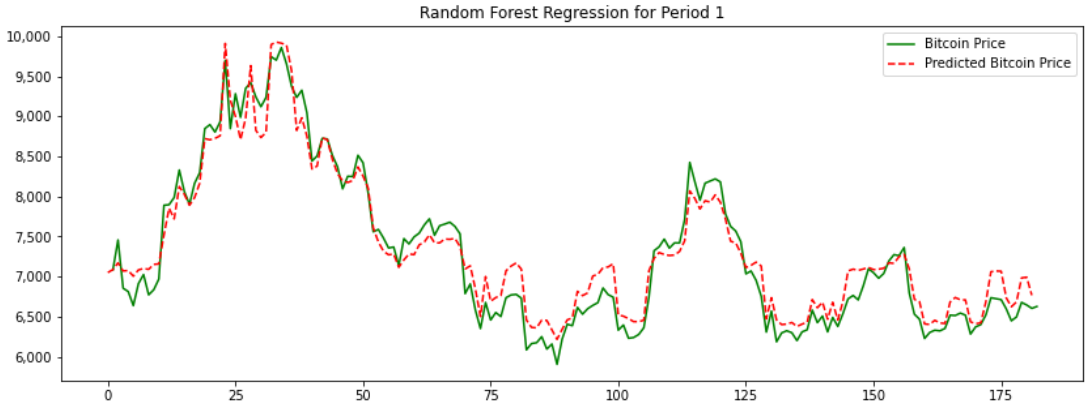
$$DA = \frac{1}{m}\sum_{t=1}^{m}a(t)\times 100\%$$

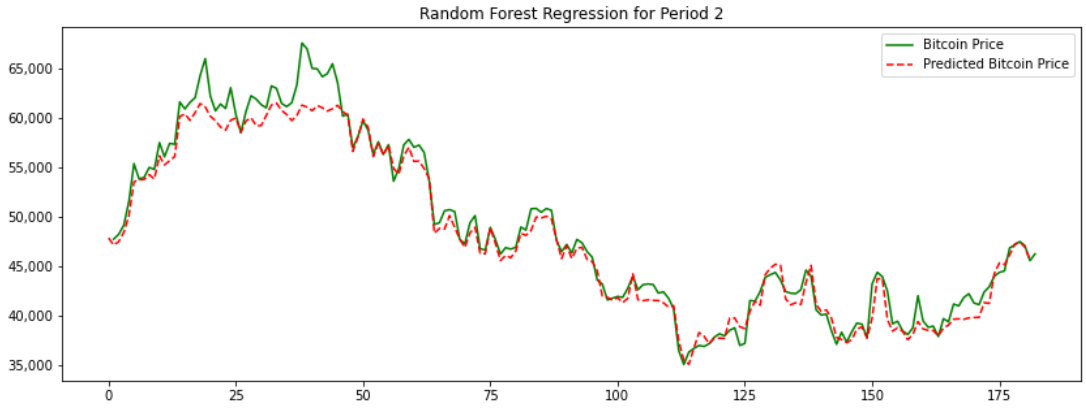
After choosing suitable evaluation criteria, we started to predict the bitcoin price with the first model appearing in the paper. Here we use the random forest algorithm to predict the price for both periods. The following table is the replica of Table 4 from paper (Found in Appendix). It's not difficult to find that we have quite similar numbers.

	Period 1	Period 2
RMSE	320.433	2107.97
MAPE	0.0337091	0.0329074
DA	51.93	51.38

Predicted price based on random forest regression and actual price comparison

In addition to calculating three errors of Random Forest Model, we also generate the plot of both real price and prediction from the model to reflect price predictions more intuitively. The following is the graph of the comparison. To recognize it from the original graph from the paper, I swap the color of two lines, where the green line represents the real price of Bitcoin from the market while the red dashed line represents the prediction price from the random forest model. Comparing our replica with Figure 8 from the paper (Found in Appendix), we also get the similar results, reflecting the authenticity of three error results.

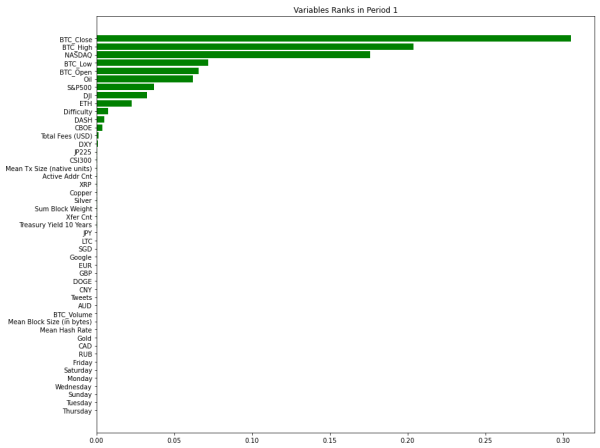


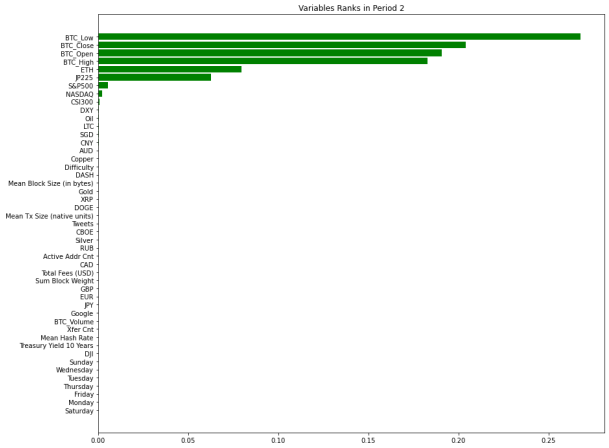


Explanatory variable importance ranks using random forest regression

After predicting the price based on all variables, we are thinking whether we can obtain better results through controlling the variables used for prediction. Hence, the first step here is to find important variables. Here, we still use the random forest algorithm as it provides the important of each explanatory variable through the statistics of occurrences' number of boundary variables in 500 sub-regression trees.

The following is the rank of explanatory variable importance, which is also the replica of Figure 9 in paper (Found in Appendix). To ensure the authenticity of the replica, I swap the color of the bar from blue to green.



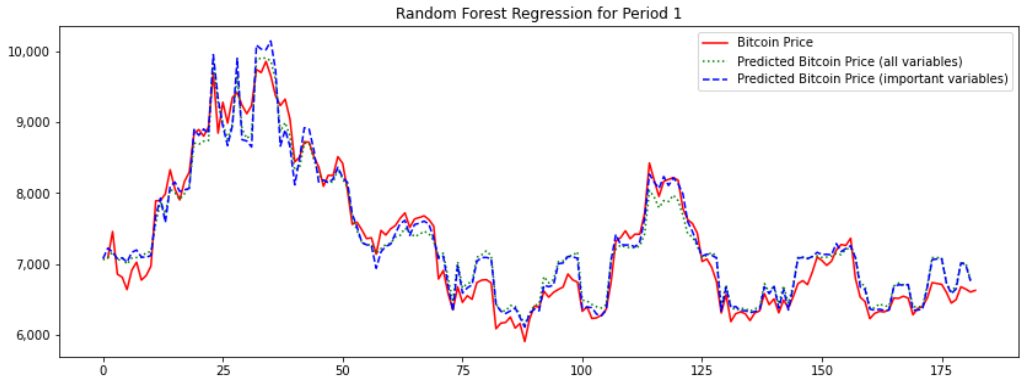


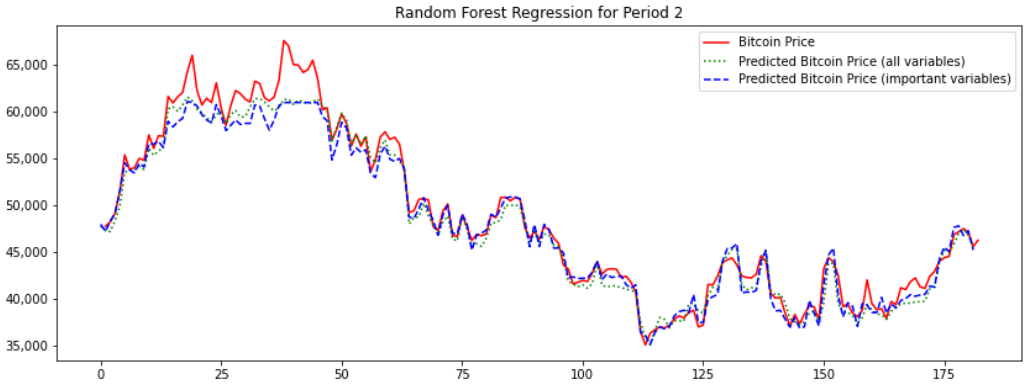
Random forest regression results by all variables and only important variables

According to the rank we got in the previous step, we choose the first six explanatory variables in each period to do the prediction with the same model again to see whether we could get a better result by controlling the variables.

The following graph is the comparison among real price, prediction based on all variables, and prediction based on only important variables. It is also the replica of Figure 11 from paper. Here we swap the colors of prediction based on variables and that of only important variables to recognize. But we can still observe the similar trends from the graph in paper.

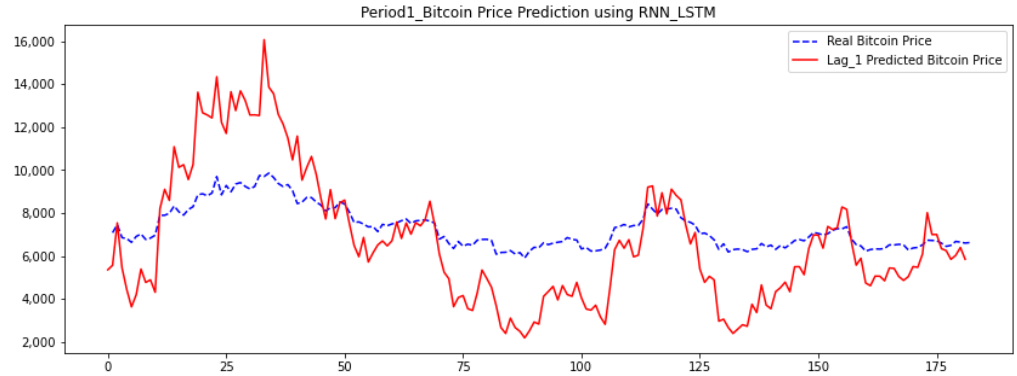
In addition, we compared the result of RMSE from the model with that from the model for all variables. And we found that the RMSE with all variables is better since it’s around 3% smaller than the results using only important variables. So, after the testing we will continue to use all variables to do the prediction in the models afterwards.

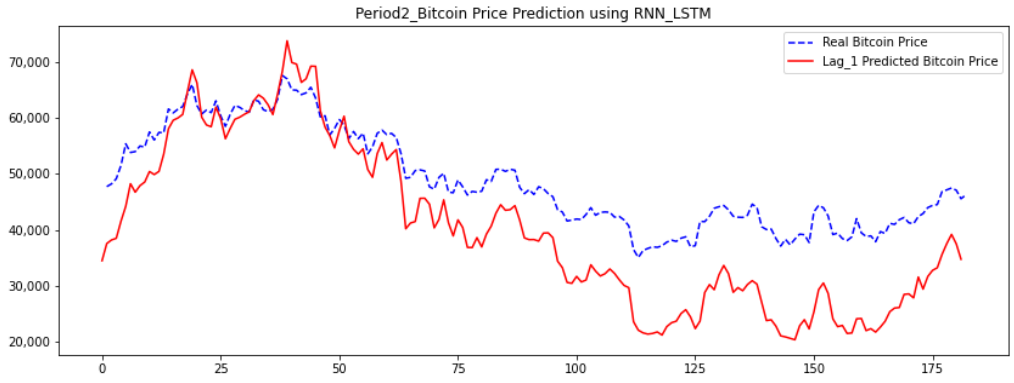




Comparison of the true price of Bitcoin and predicted price based on one-lag LSTM model

The second model used to predict the Bitcoin price is the LSTM model. For the sake of simplicity, we do the prediction with one lag first. The following is the plot of comparison between real price in two periods and the predicted price based on the one-lag LSTM model. Compared with the Figure 12 from the paper (Found in Appendix), we also get the similar trends with the paper.





Errors of the LSTM models with one-lag

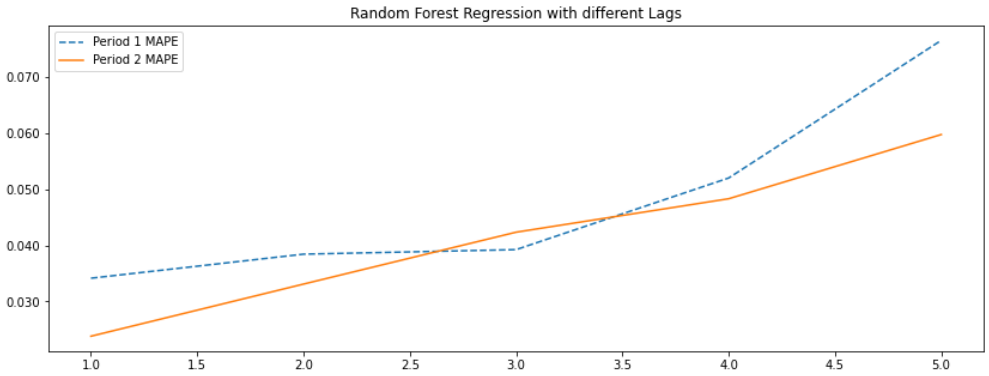
To compare the results better, we also calculated the RMSE and MAPE from the LSTM model. Compared with the result from Table 8 in the paper (Found in Appendix), we got very close numbers.

However, we also found the result from the LSTM model is worse than errors from the previous random forest model. And we are considering whether it is the problem of the number of lags. Thus, we tried different lags in two models in the next step to find out whether we could have a closer price prediction.

	Period 1	Period 2
RMSE	360.48	3168.71
MAPE	0.0399781	0.0497339

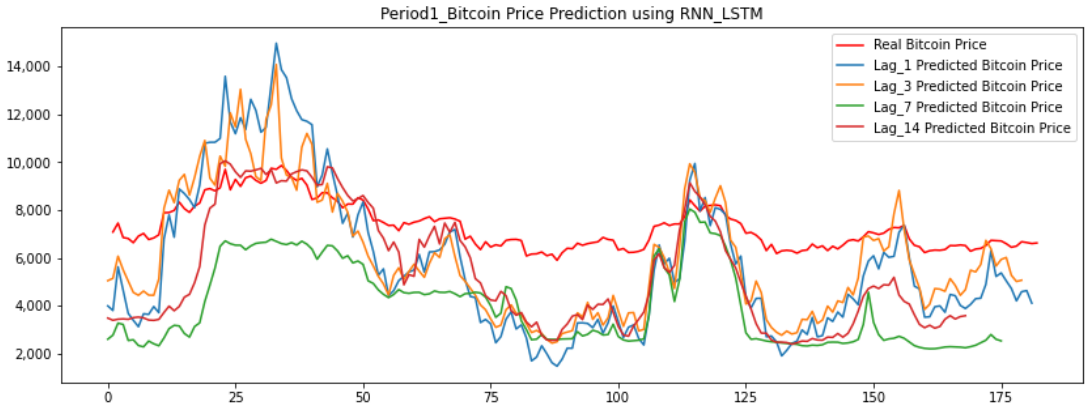
Relationship between MAPE and the number of lags (random forest regression)

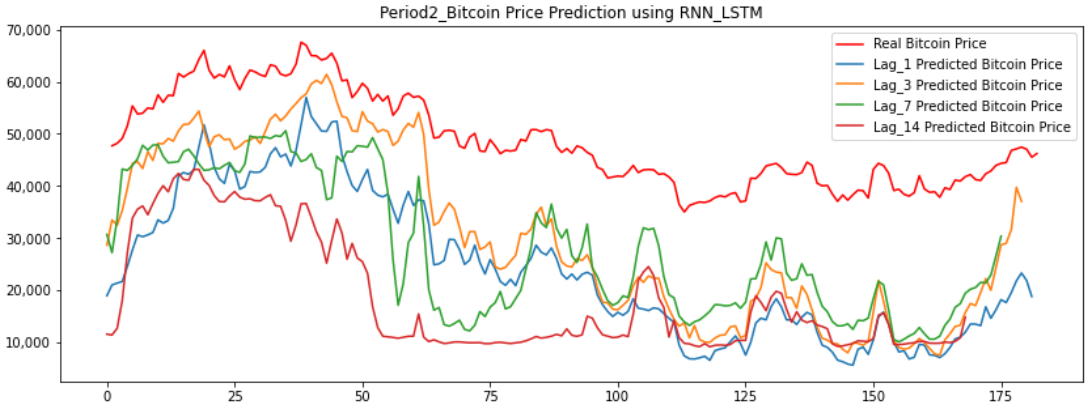
The following is the regression after testing the lag number from 1 to 5, which is the replica of Figure 13 (found in Appendix). I swap the line style of two periods from the paper. Here, we could observe that the MAPE increases with the number of lags increases in both period 1 and period 2.



Relationship between accuracy and the number of lags (LSTM)

Just as mentioned, we also tried to figure out the influence of lag numbers on the prediction in the LSTM model. Thus, we replace the number of lags with 1, 3, 7, 14 and compare these four prediction accuracies with the real price in the following replica of Figure 14. We could clearly see that the more lags we used in the model the more periods of data will be substituted into the model. Then the smoother the curve of the forecast data becomes. But the prediction will be also more deviating from the real price.





3 Improvements

Date Preprocessing

We employ prediction models to support trading decision making. If prediction shows the price will go up, we can long the instrument, otherwise, short. Thus, we specifically looked at the differential of each of these variables, in order to predict the sign of the price change or the sign of the return, as opposed to the actual price itself. Instead of using the original magnitude dataset, we processed the data into differential dataset and log differential dataset.

Differential dataset: uses price change [$P_t - P_{t-1}$] as the target

Date	BTC_Open	BTC_High	BTC_Low
2015/4/1	-3.230988	-1.188996	-1.578994
2015/4/2	2.865998	6.919999	4.255996
2015/4/3	5.985	1.582	6.462998
2015/4/4	1.216995	-0.785003	-0.778992
2015/4/5	-0.529998	5.416992	0.841995

Log differential dataset: uses log return [$log\ r_t = log\ P_t - log\ P_{t-1}$] as the target

Date	BTC_Open	BTC_High	BTC_Low
2015/4/1	-0.01314291495	-0.00479172986	-0.006526153589
2015/4/2	0.01166684485	0.02757135381	0.01749409962
2015/4/3	0.02393333999	0.006197816551	0.0259940757
2015/4/4	0.004797324695	-0.003070612514	-0.003097515469
2015/4/5	-0.00208639348	0.02099959194	0.003347616201

New Models

The original paper uses Random Forest and LSTM on magnitude dataset which gives around 50% accuracy. Some other machine learning algorithms are commonly used in bitcoin price prediction, including KNN, Binomial GLM and XGBoost. Thus, we will test the three models on the datasets above to improve the accuracy.

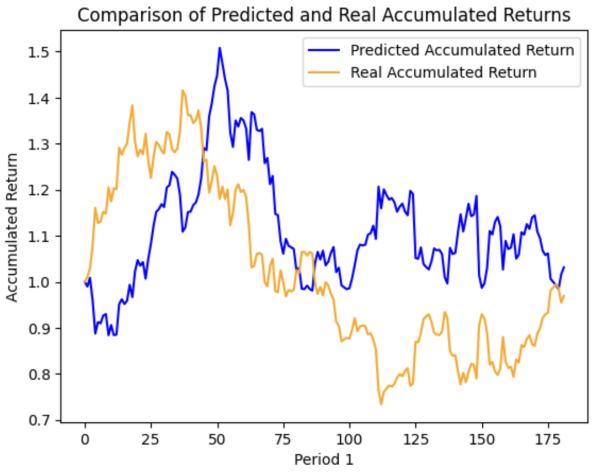
Period		Original Paper		New Model with magnitude data			New Model with differential data			New Model with log diff data		
		Random Forest	LSTM	Binomial GLM	KNN	XGBoost	Binomial GLM	KNN	XGBoost	Binomial GLM	KNN	XGBoost
P1	RMSE	321.61	330.26	N/A	1271.09	408.45	N/A	545.72	351.44	N/A	0.03	0.04
	MAPE	3.39%	3.57%	N/A	3.67%	4.52%	N/A	3.05%	4.06%	N/A	2.52%	3.56%
	DA	51.93%	49.98%	50.55%	53.04%	48.62%	53.85%	58.56%	**65.75%**	53.85%	**67.4%**	**66.85%**
P2	RMSE	2096.24	3045.87	N/A	1650.79	3374.69	N/A	1901.25	1727.42	N/A	0.06	0.04
	MAPE	3.29%	4.68%	N/A	2.13%	4.83%	N/A	2.23%	2.00%	N/A	6.70%	2.09%
	DA	52.49%	48.09%	50.27%	51.38%	52.49%	50.28%	71.67%	**72.78%**	50.28%	52.22%	**70.56%**

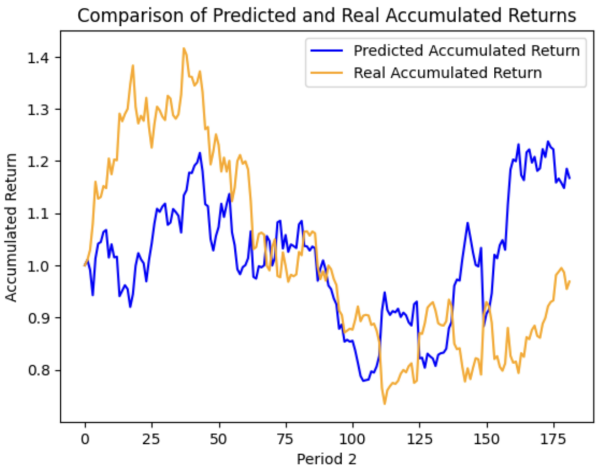
Binomial GLM, giving binary results of 1 for positive price change/log return and 0 for negative, is not applicable with metrics including RMSE and MAPE. So, only Decision Accuracy (DA) will be shown in the table. By comparison, XGBoost gives the best accuracy among all models and differential/log differential datasets show significant improvement with the same model.

Trading Strategy and Performance

Depending on the model performance, we will use XGBoost and log differential datasets to construct a trading strategy. We will long Bitcoin if the prediction gives positive results and short otherwise. For comparison, we will also show the Long-Term Holding Strategy which means to buy and hold bitcoin until the end of the test period.

Performance Metrics	**Period 1**	**Period 2**
Sharpe Ratio	0.02	0.04
Annual Return	30.98%	68.08%





Our strategy has the potential to have a stable return. Holistically, our strategy can not only capture the alpha but also avoid the great loss of long-term holding strategy

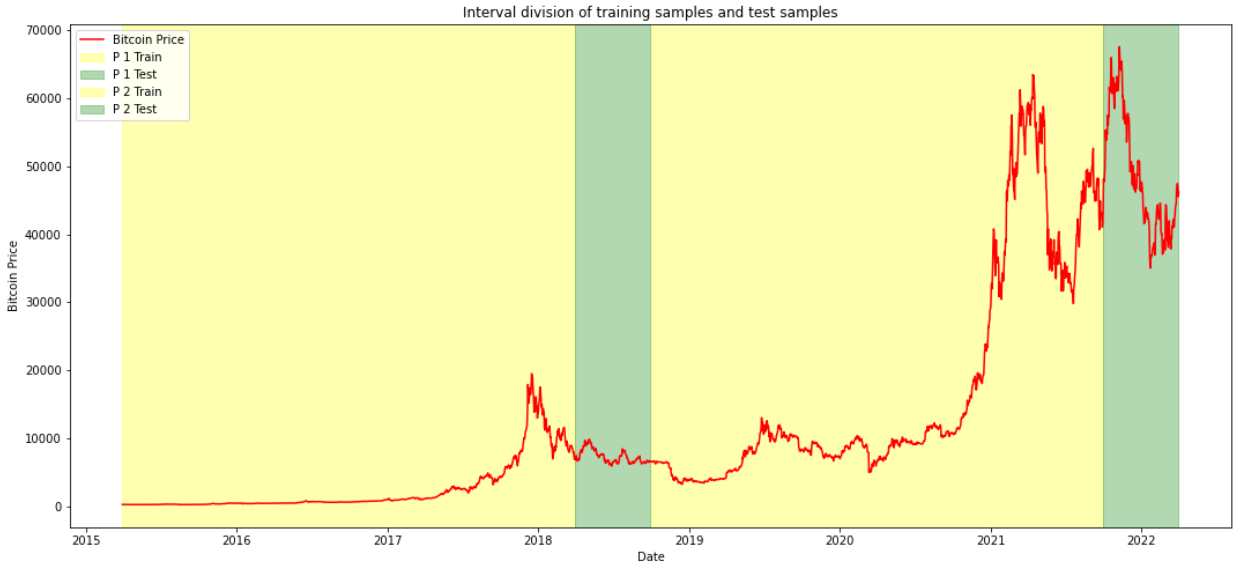
Period Selection

Old: 2 Periods

Table 3(Percentage of Train Data have differences)

	Train Data	Test Data	Percentage of Train Data
Period 1	31 March 2015–31 March 2018	1 April 2018–30 September 2018	0.841601
Period 2	1 October 2018–30 September 2021	October 2021–1 April 2022	0.857478

Reproduce Figure6 of the original paper



New: 4 Periods

The paper divided the space from March 31, 2015, to April 1, 2022, by two periods. However, according to the big events of cryptocurrency, we think we should divide it into more periods.

Period 1: March 31, 2015 - July 31, 2017

This period covers the early growth phase of Bitcoin leading up to the creation of Bitcoin Cash in August 2017. It includes the moment when Bitcoin first crossed the $2,000 mark in May 2017.

Period 2: August 1, 2017 - December 31, 2018

A significant phase includes the Bitcoin Cash hard fork, the 2017 price bubble, and the subsequent market crash at the end of 2017, followed by a year of market correction in 2018.

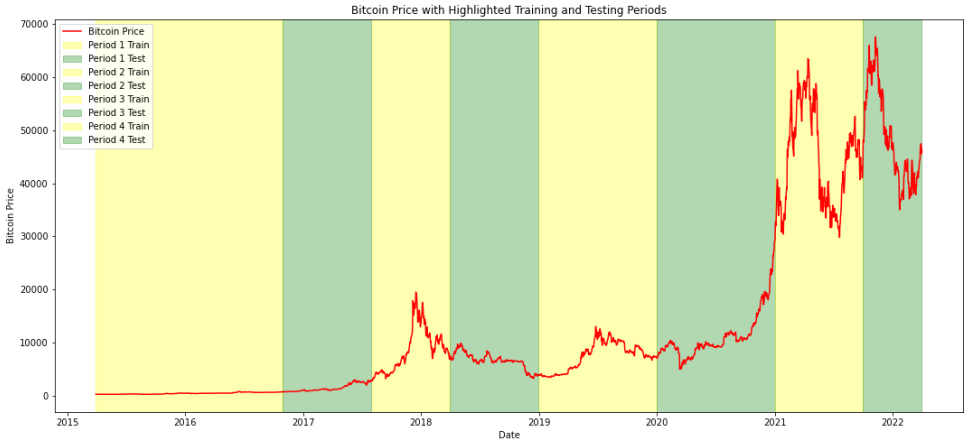
Period 3: January 1, 2019 - December 31, 2020

This period covers the post-crash recovery, various legal and regulatory developments, growing institutional interest in Bitcoin, and the initial impact of the COVID-19 pandemic. Key events include the legal disputes involving Craig Wright and the introduction of crypto services by PayPal.

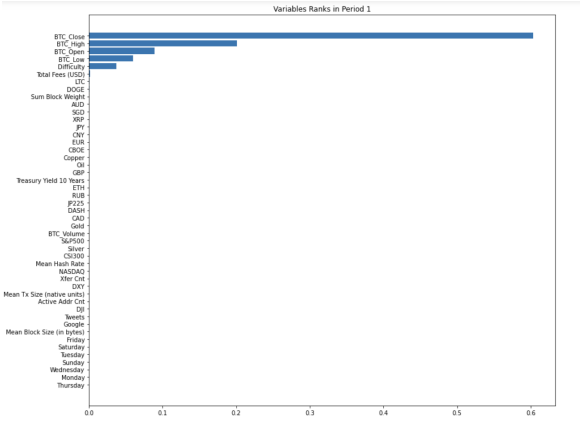
Period 4: January 1, 2021 - April 1, 2022

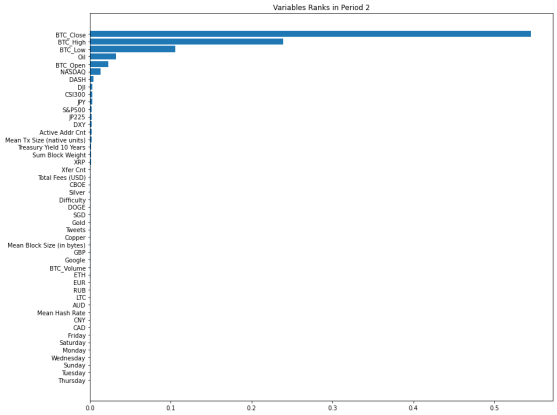
The most recent phase, marked by Bitcoin's surge past $40,000 for the first time and continuing developments and adoption in the cryptocurrency space, including increased institutional interest and regulatory discussions.

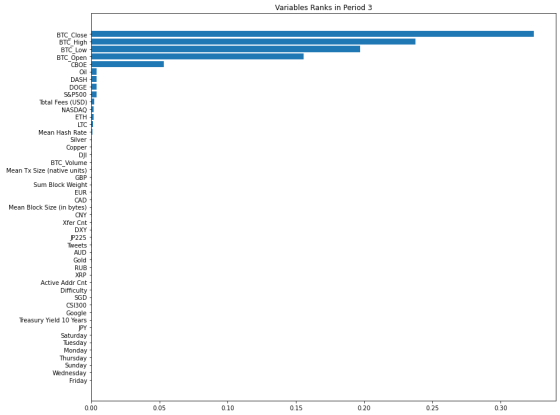
Reproduction of Figure 6 for new periods

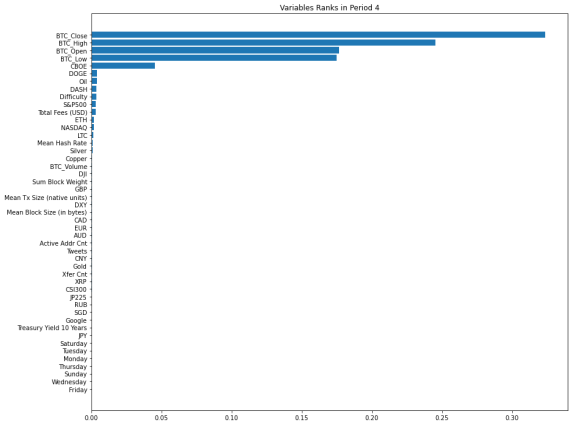


Variables Ranks in 4 periods







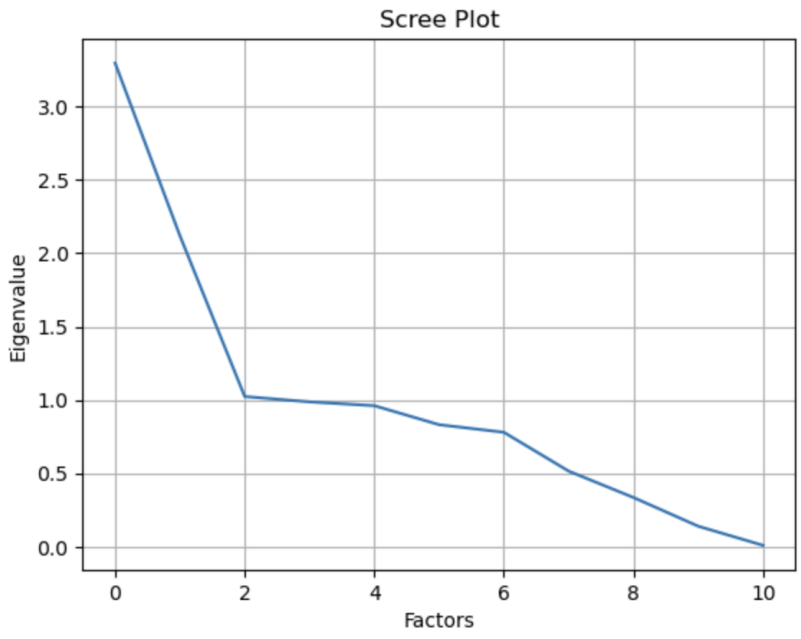


Compared to the original period, the new period gives more variety in factors. For instance, Total Fees have never occurred in the original period but it has explanatory power in period 3. ETH, which ranks high in the origin period, is hard to explain the BTC in new periods.

Feature Analysis

In this section, we conducted our analysis by selecting the top 10 features exhibiting the highest correlation, as evidenced by the heatmap. These features were then subjected to a principal component analysis to extract pertinent factors. Prior to executing the factor analysis, we employed Bartlett's Test of Sphericity and the Kaiser-Meyer-Olkin (KMO) tests, both of which substantiated the suitability of our data for factor analysis. After these tests, we performed principal component analysis (PCA) to distill the factors.

Using a scree plot was instrumental in determining the optimal number of factors to retain. While five factors were initially identified, further scrutiny led to Factor 1 and Factor 2 selection based on their eigenvalues. These factors ostensibly represent latent trends and patterns prevalent across the financial variables under consideration.



Notably, the regression analysis, focusing on the relationship between Bitcoin's closing prices and the two factors extracted from the factor analysis, revealed a significantly high R-squared value. This suggests that these factors account for a substantial portion of the variance in Bitcoin closing prices. Specifically, the coefficients associated with Factor 2 denote a more pronounced impact. However, it is pertinent to note that our analysis did not pass the Jarque-Bera (JB) test, which can be attributed to the residual instability caused by specific events such as Elon Musk’s Twitter activities and the COVID-19 outbreak.

Further, we identified the top and bottom two variables most closely related to Factor 1 and Factor 2, respectively. Intriguingly, the results indicate that Factor 1 likely encapsulates broader market trends influencing both traditional and cryptocurrency markets. Conversely, Factor 2 appears to indicate aspects uniquely pertinent to cryptocurrencies or specific economic indicators.

Finally, we assessed the predictive efficacy of Factors 1 and 2. The findings suggest that Factor 2, derived from three distinct variables, exhibits greater predictive power. While these factors were not incorporated into our models, we propose their

consideration for enhancing future predictive models, as this methodology was not addressed in the original paper.

```
Intercept    0.002881
factor1      0.001011
factor2      0.039078
dtype: float64
                            OLS Regression Results
==============================================================================
Dep. Variable:              BTC_Close   R-squared:                       0.826
Model:                            OLS   Adj. R-squared:                  0.825
Method:                 Least Squares   F-statistic:                     5738.
Date:                Wed, 06 Dec 2023   Prob (F-statistic):               0.00
Time:                        16:45:08   Log-Likelihood:                6540.3
No. Observations:                2429   AIC:                         -1.307e+04
Df Residuals:                    2426   BIC:                         -1.306e+04
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      0.0029      0.000      8.661      0.000       0.002       0.004
factor1        0.0010      0.000      3.146      0.002       0.000       0.002
factor2        0.0391      0.000    106.984      0.000       0.038       0.040
==============================================================================
Omnibus:                      928.678   Durbin-Watson:                   1.870
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            33576.111
Skew:                          -1.130   Prob(JB):                         0.00
Kurtosis:                      21.073   Cond. No.                         1.14
==============================================================================
```

4 Conclusion

This paper aims to replicate and enhance Junwei Chen's influential study, "Analysis of Bitcoin Price Prediction Using Machine Learning," leveraging a rich dataset and advanced methodologies. Our reproduction effort faithfully adhered to Chen's original methods, using Random Forest Regression and Long Short-Term Memory (LSTM) models, and our findings echoed the reliability and effectiveness of these approaches in predicting Bitcoin prices. The replication served as a foundational benchmark, against which we measured our innovative improvements.

The introduction of new data preprocessing techniques and the implementation of additional models, notably the XGBoost model, marked a significant stride in our research. These innovations not only demonstrated an increased predictive accuracy but also offered fresh insights into the multifaceted nature of Bitcoin's price determinants. The XGBoost model, in particular, emerged as the superior model in terms of accuracy, showcasing the value of exploring and integrating diverse modeling techniques.

Furthermore, our feature analysis using principal component analysis (PCA) unearthed critical underlying factors that influence Bitcoin prices. The identification of these factors and their correlation with Bitcoin's closing prices underscores the complexity and multi-dimensionality of the cryptocurrency market. The analysis highlighted the impact of broader market trends and unique cryptocurrency-specific indicators on Bitcoin's valuation.

While our study made significant headways, it also recognized certain limitations. The failure for PCA to pass the Jarque-Bera (JB) test pointed towards residual instabilities, possibly stemming from specific unpredictable events. These findings highlight the inherent volatility and unpredictability of cryptocurrency markets and the challenges they pose for predictive modeling.

In conclusion, our research not only reaffirms the validity of Chen's original findings but also pushes the boundaries of knowledge in Bitcoin price prediction. By introducing refined techniques and a broader analytical lens, we have contributed a substantial body of work to the fintech field. This study serves as a stepping stone for future research, offering a more nuanced understanding of Bitcoin market dynamics and a robust framework for cryptocurrency market analysis.

Appendix

Figure 3: Training and validation loss of LSTM

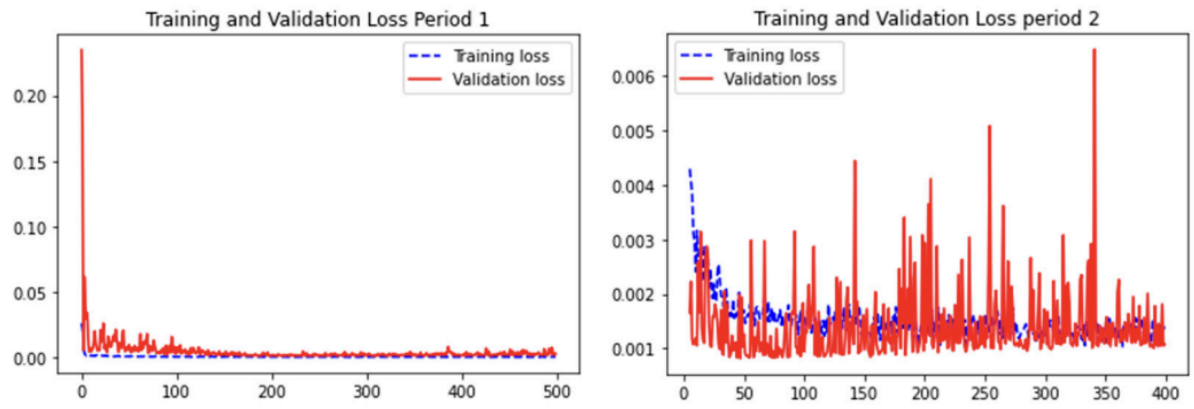


Figure 4: Correlation heat map

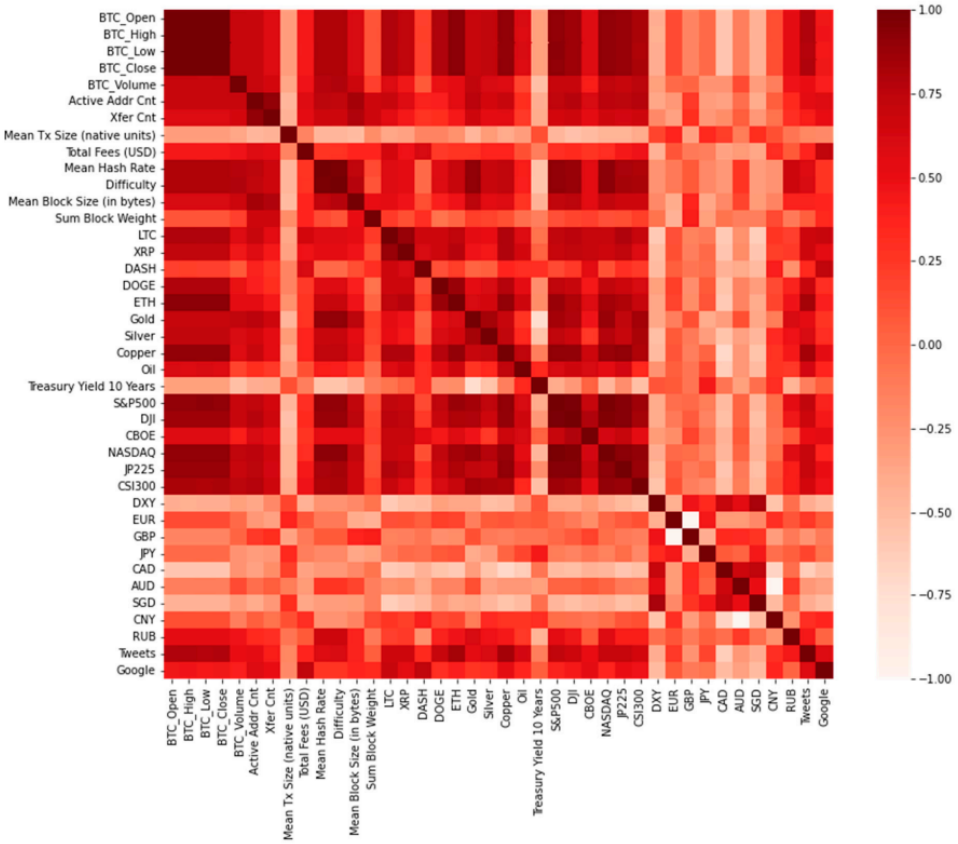


Figure 8: Predicted price based on random forest regression and actual price comparison

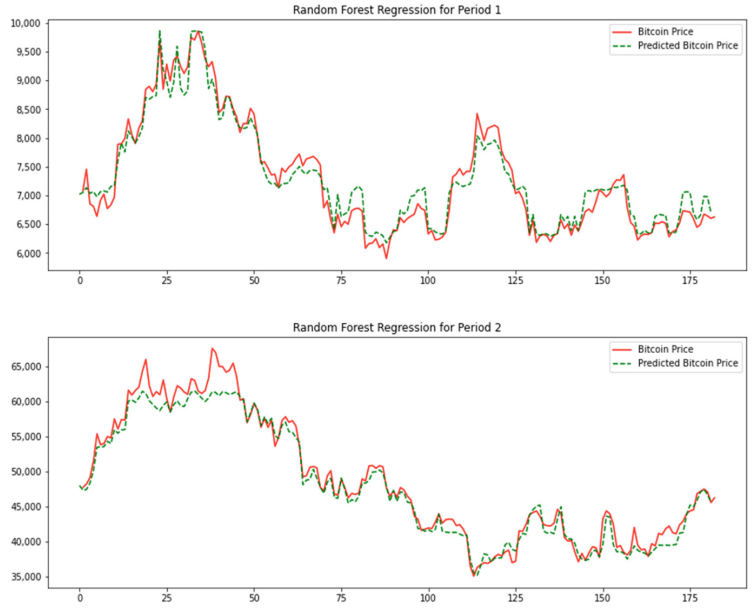


Figure 8. Predicted price based on random forest regression and actual price comparison.

Table 4: Error results from random forest regression

Table 4. Error results for random forest regression.

	Period 1	**Period 2**
RMSE	321.61	2096.24
MAPE	3.39%	3.29%
DA	51.93%	52.49%

Figure 9: Explanatory variable importance ranks using random forest regression

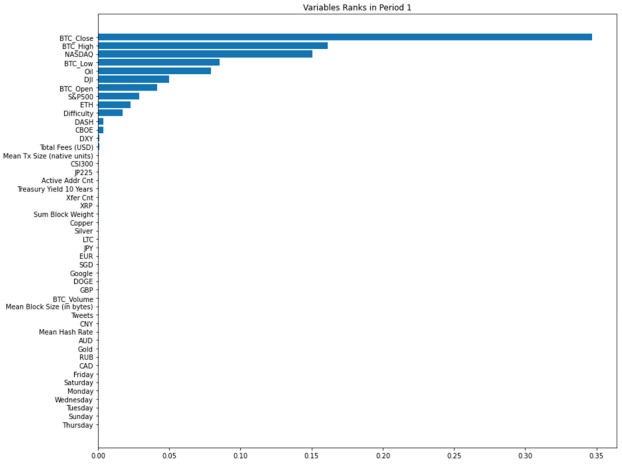


Figure 9. *Cont.*

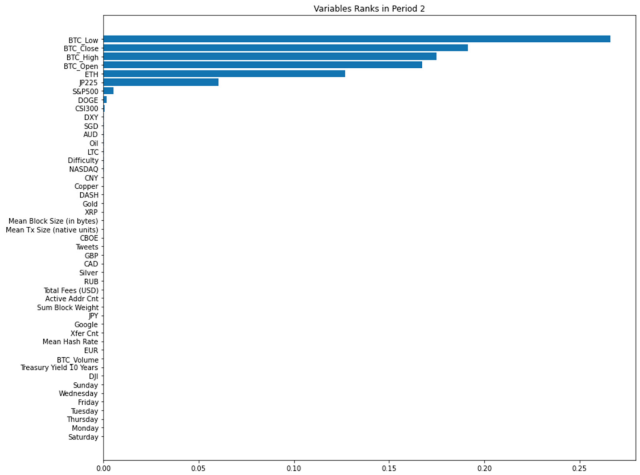


Figure 9. Explanatory variable importance ranks using random forest regression.

Figure 11: RFR results by all variables and only important variables

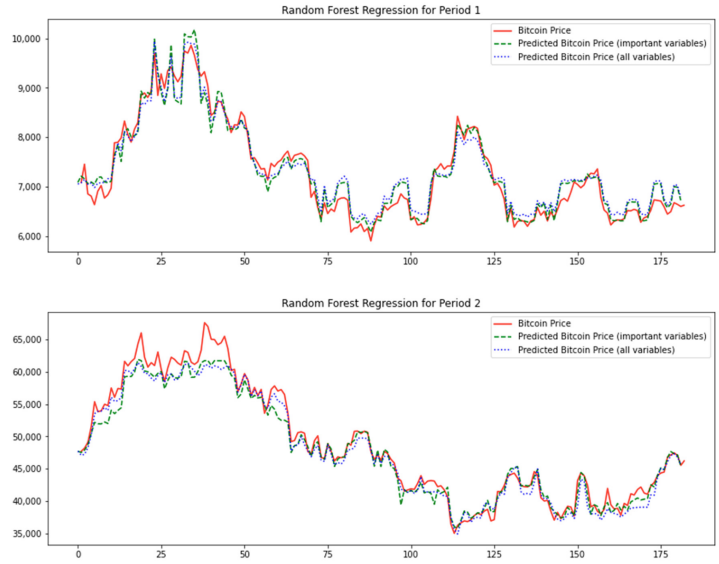


Figure 11. RFR results by all variables and only important variables.

Figure 12: Comparison of the true price of Bitcoin and predicted price based on different models

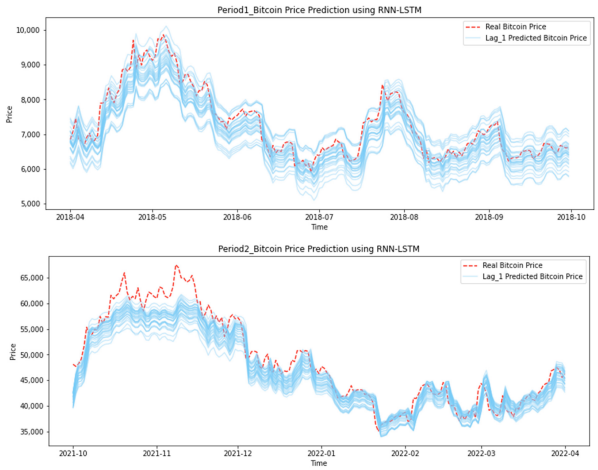


Figure 12. Comparison of the true price of Bitcoin and predicted price based on different models. (LSTM).

Table 8: Errors of the LSTM models

Table 8. Errors of the LSTM models.

	Period 1	Period 2
RMSE	330.26	3045.87
MAPE	3.57%	4.68%

Figure 13: Relationship between MAPE and the number of lags (RF)

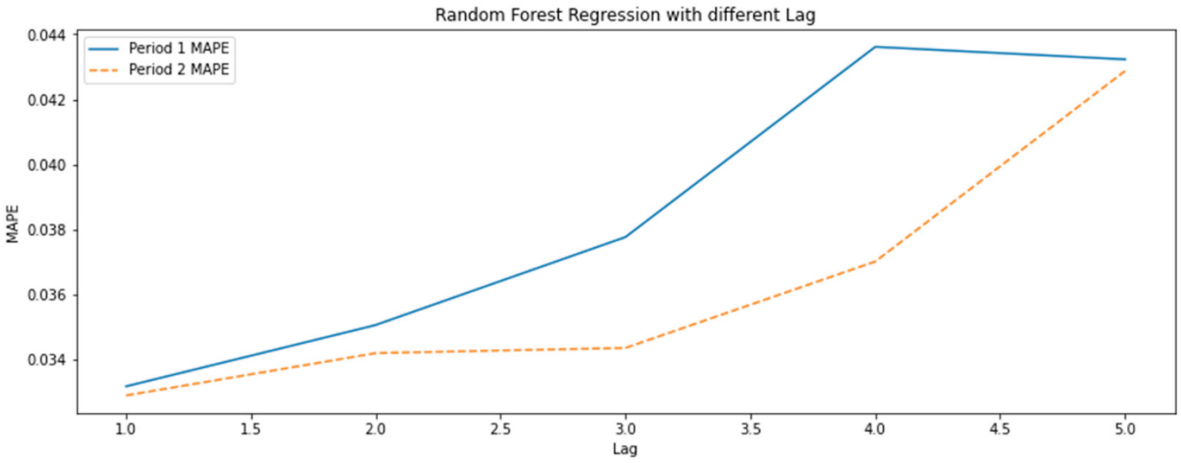


Figure 13. Relationship between MAPE and the number of lags (random forest regression).

Figure 14: Relationship between accuracy and and the number of lags (LSTM)

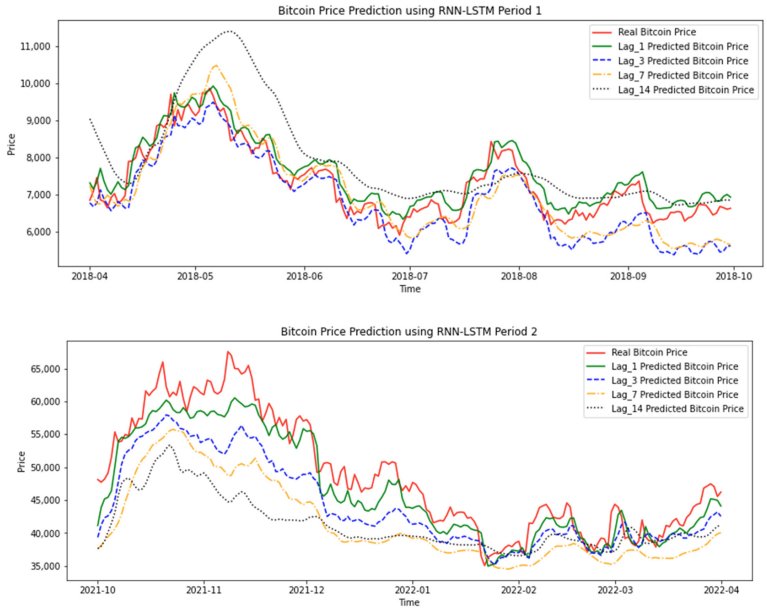


Figure 14. Relationship between accuracy and the number of lags (LSTM).