Proceedings of the 4th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/55/20241404

Litecoin price prediction based on random forest regression, LightGBM and LSTM

Shaohui Lang

Department of Statistics, University of California-Santa Barbara, Santa Barbara CA 93106, The United States

shaohui_lang@ucsb.edu

Abstract. Cryptocurrency has caught huge amounts of investment attentions and interests ever since its first introduction. However, due to the highly instable nature of its price, it is crucial for investors to avoid risks when investing in cryptocurrency. Studies has been conducted to predict the price of cryptocurrency with different price influencing factors using various machine learning models. On the other hand, most of them only focused on major types of cryptocurrencies, e.g., Bitcoin and overlooked minor ones. This study focuses on one type of minor cryptocurrency, Litecoin. Three machine learning algorithms, random forest regression, light gradient boosting machine (LightGBM), and long short-term memory (LSTM) are used to predict long-term Litecoin price. The effects of 19 price influencing factors are considered, including Litecoin price variables, other popular cryptocurrency prices, major foreign exchange prices, market indices, and major commodities prices. Root mean squared error (RMSE), mean absolute percentage error (MAPE), and R-squared score are used to evaluate model performances. The results suggest that, among the 3 models, random forest model shows the best prediction with the least error, while LSTM model has the most error. Such result can provide insights for investors to avoid risks in Litecoin investments. Future studies are still necessary to take more types of cryptocurrency and price influencing factors into consideration.

Keywords: Cryptocurrency, Litecoin, random forest, LGBM, LSTM.

1. Introduction

Cryptocurrency refers to the concept of digital currency, which is an alternative online transaction method without authority regularization. The first cryptocurrency, Bitcoin, was first introduced in 2009 by Satoshi Nakamoto [1]. Since then, huge amounts of cryptocurrency have been launched into the market, including Ethereum, Litecoin, etc. In recent years, cryptocurrency has experienced massive development. Increasing amounts of investment have been made into the cryptocurrency market, despite the high volatility of cryptocurrencies' prices [2]. Up until Sep 2023, there exists over 20000 types of cryptocurrency in the market, with a global market capitalization over 1 trillion US dollars [3]. The huge investment and fast development of the cryptocurrency market has also drawn interests in financial research area. According to a sampled survey by Fang et al., from 2013 to June 2021, over 85% of research papers regarding cryptocurrency trading have been publicized after 2018 [4].

Due to the decentralized nature of cryptocurrency, its price is also much more volatile and susceptible than prices of normal commodities and currencies. For example, Chen stated that, from 2015 to 2022,

DOI: 10.54254/2755-2721/55/20241404

the standard deviation of Bitcoin's daily return rate is about 1.7 times greater than that of gold [5]. Such huge price fluctuation has always been an argument against the value and function of cryptocurrencies. On the other hand, cryptocurrency still occupies a huge market sector despite its high risk. Thus, to help with market investments, predicting the trends of cryptocurrency prices has become a popular topic among researchers and investors. In recent years, with the advancements of machine learning technology, price prediction of cryptocurrencies becomes more compassable. Generally, related works fall into one of the two categories: classification tasks or regression tasks, where each of the two categories has their own rationale and advantages. Classification tasks are used to forecast whether the price will rise or fall in the next period, while regression tasks model the prices directly.

Past studies focusing on cryptocurrency price prediction have implemented and compared numerous models, including support vector machine (SVM), random forest, gradient boosting decision tree (GBDT), deep learning algorithms like long short-term memory (LSTM), etc. Various research has shown different results under various circumstances. In order to predict short-term cryptocurrency price trends, Sun, Liu and Sima employed 42 economic indicators to explain cryptocurrency market, and implemented 3 models: SVM, random forest, and light gradient boosting machine (LightGBM) [6]. LightGBM outperformed the other two models and showed the best accuracy. Chowdhury et al. also studied relative short-period cryptocurrency price prediction. However, they only considered price variables as explanatory variables, and employed 4 models: GBDT, neural network, ensemble learning, and k-Nearest Neighbor algorithm (k-NN) [7]. Under this circumstance, the overall performances of the models are decent except that of k-NN. In another study conducted by Jaquart, Dann and Weinhardt, the researchers employed classification methods in short-term Bitcoin price prediction. Technical, asset-based, blockchain-based, and sentiment features were considered. Neural network models including LSTM, tree-based models, random forest, gradient boosting classifier, and ensemble models were fitted and evaluated [8]. All the models seemed viable, but LSTM showed the best accuracy. In addition to those studies, there are also plenty of other studies investigating Bitcoin prediction with different features, in which LSTM was commonly used as the baseline model. Major of them, like Aggarwal et al. and Chen et al., have shown that LSTM was a preferable model [9, 10]. On contrary, there are also some studies that suggest other models as more accurate. For example, under Chen's study design and feature selection, random forest regression was a more stable and accurate model than LSTM [5].

Past studies normally focused on Bitcoin, the biggest cryptocurrency in the world. Despite there do exist research about minor cryptocurrencies, e.g., Patel et al. on Litecoin and Agarwal et al. on Dogecoin, the majority of them only focused on short-term price predictions [11, 12]. Thus, the focus of this study will be shifted to long-term price prediction of Litecoin, a rather smaller type of cryptocurrency. Regression algorithms are implemented. The proposed models are random forest regression, LightGBM, and LSTM. These models are selected because they have shown to be accurate and efficient models in previous works.

2. Data and method

The data of the study were obtained from yahoo.finance and investing.com. The collected data are in a 5-year range, from September 1, 2018 to September 1, 2023. The dataset after preprocessing includes 1827 rows and 20 columns. Each row represents one day in the 5-year range. Each column represents one variable, including 1 target variable and 19 explanatory variables. The target variable is the closing price of Litecoin in USD on each day. The explanatory variables include 6 categories: price variables for Litecoin, stock prices for Litecoin, prices of other popular cryptocurrencies, major foreign exchange prices, market indices, and future prices for major commodities. The specific variables under each category are listed in Table 1.

The study employs three supervised machine learning algorithms: Random Forest Regression, Light GBM, and LSTM Network. All the three models are trained and tested using Python machine learning libraries. Due to the temporal and sequential characteristic of the data, the first 80% of the data (first 4 years) is used as training data, and the last 20% (last 1 year) is used as testing data to make predictions. Random Forest Regression is a supervised machine learning algorithm that is built upon a collection of

DOI: 10.54254/2755-2721/55/20241404

decision trees. Each decision tree separates the data based on the given features and make predictions at its leaf nodes. The Random Forest algorithm then use an ensemble technique to make predictions (by aggregating the results from all the independent trees) [13]. This way, the final prediction can overcome the problem of overfitting that is inherit to decision tree algorithm. Random Forest is also robust when handling data with multiple continuous features [14]. In this study, random forest regression is implemented using the random forest regressor function from Python's Scikit-learn library. The number of decision trees in the forest is set to 100, and the random state parameter is set to 7 to ensure reproducibility. The rest of the parameters are set to default.

LightGBM is one special implementation of the GBDT algorithm. GBDT is a popular machine learning model which is built on ensembles of decision trees. While the traditional GBDT methods can inefficient or inaccurate when retrieving information from large dataset, LightGBM resolves such problem by two techniques: Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB) [15]. In specifical, GOSS reduces number of instances by removing samples with small gradient, and EFB reduces number of features by combining mutually exclusive ones [15]. LightGBM can thus achieve both efficiency and accuracy in regression tasks of large dataset. In this study, LightGBM is implemented through Python's lightgbm library. The parameters for the LightGBM model are set as follow. Objective is set to regression, metric is set to l2 (mean squared error), boosting type is set to gbdt (EFB technique is automatically employed), number of leaves for each decision tree is set to 30, learning rate is set to 0.05, feature fraction (fraction of features in each boosting round) is set to 0.8, and number of trees is set to 100.

LSTM network is one type of recurrent neural network (RNN). The RNN algorithm is created to handle sequential problems [16], but it has a deficit (gradient vanishing) in retaining long time-sequential data as input increase [17]. LSTM resolves this problem by employing input gate, forget gate, and output gate. With the help of these three gates, LSTM can automatically preserve significant features and remove uncorrelated ones [17]. Thus, LSTM preforms better in long-term time-series data. In the study, LSTM is implemented through Python's Keras library. Specifically, LSTM layer with 50 neurons is added to the sequential initializer with activation function Rectified Linear Unit (ReLU). ReLU is a common activation function and performs better than tanh and sigmoid in such price prediction task [5]. Then a dense layer with 1 neuron is added. Adam is used as the optimizer, and mean squared error is used as the loss function. When fitting the model, the two hyperparameters, epochs (number of times the model learn the train data) and batch size (number of training samples used to update model weights in each update), are set to 50 and 32 respectively.

Table 1. Specific variables.

Category	Variable	Description
Price Variables for Litecoin	Open	Opening price of Litecoin
	High	Highest price of Litecoin on that day
	Low	Lowest price of Litecoin on that day
	Volume	Litecoin daily transaction volume
Stock Price for Litecoin	ltc stock	Litecoin's stock price
Other Popular Cryptocurrencies	btc	Bitcoin's closing price
	doge	Dogecoin's closing price
	eth	Ethereum's closing price
	usdt	Tether's closing price
	xrp	Ripple's closing price
Major Foreign Exchange Prices	cad	Canadian dollar price
	cny	Chinese yuan price
	eur	Euros price
	gbp	British pound price
	jpy	Japanese yen price
Market Indices	nasdaq	NASDAQ composite index

DOI: 10.54254/2755-2721/55/20241404

Table 1. (continued).

	sp500	The standard and poor's 500 index
Commodity Future Prices	gold	Gold future price
	oil	Crude oil future price

3. Results and discussion

There are totally 19 features (explanatory variables) in the dataset, including 6 categories: price variables for Litecoin, stock prices for Litecoin, prices of other popular cryptocurrencies, major foreign exchange prices, market indices, and future prices for major commodities. The specific features in each category are listed in Sec. 2. These features are selected as they are either direct or indirect influential factors to Litecoin close price. Among these features, Litecoin price variables and other popular cryptocurrency prices do not have missing values. However, all the other features contain missing values on weekends and holidays. It is exchange rates, stock prices, commodity future prices, and market indices are unavailable on these days. Since the data is temporal-sequential, deleting instances with missing values will break the time continuity of the data and thus is not applicable. Therefore, the study chooses to fill in the missing values using values of the previous day for each feature. Feature scaling or normalization can improve model performances based on selected algorithm. However, the first two models of the study, random forest and LightGBM, do not require this process, because these two models are built upon decision trees. The partitions that decision trees make are not influenced by feature scaling or normalization. On the other hand, LSTM requires feature scaling since unscaled input can cause gradient problems, which adversely affect model performances. In the study, min-max scaling is implemented for LSTM model. The three models, random forest regression, LightGBM, and LSTM, are trained using first 1461 days (first 80%) of the data, and tested using data from Sep 1, 2022 to Sep 1, 2023 (last 20% of the data). To evaluate the models, metrics including root mean squared error (RMSE), mean absolute percentage error (MAPE), and R-squared score are used (RMSE and R-squared are retrieved through scikit-learn library, MAPE is calculated manually through NumPy). The three models' performances within the testing period are as follow. Fig. 1 shows the true Litecoin prices against the random forest models' prediction. The blue line represents the true Litecoin prices and the red line represents the random forest model predicted prices. It shows that the random forest predicted price strictly follow the moving trend of the true Litecoin price. To be more specific, when the true price increases or decreases, the model generates similar moving direction and speed and largely has miniscule differences with the real price. On the other hand, the model's predictions on local maximum or local minimum are not as well. On Sep 6, 2022, the true Litecoin price came to a local floor of 54.31, however, the prediction only declined to 57.07. Such discrepancies happen at most of the local price floors and ceilings.

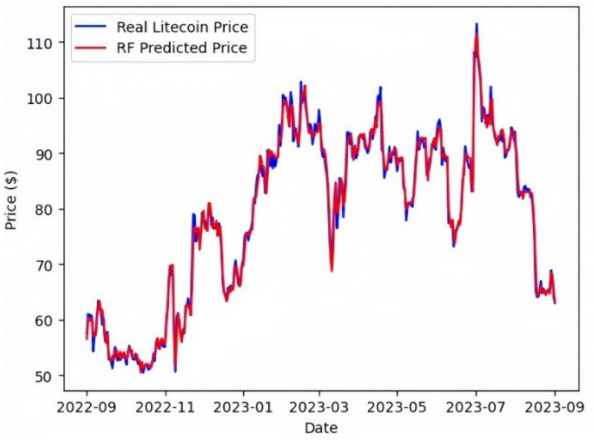

Figure 1. Random Forest Prediction (Photo/Picture credit: Original).

DOI: 10.54254/2755-2721/55/20241404

Fig. 2 shows the true Litecoin prices against the LightGBM models' prediction. The blue line represents the true Litecoin prices and the red line represents the LightGBM model predicted prices. The general trend of LightGBM prediction fit well with the actual Litecoin price. However, there are multiple short-term periods where the model fails. For instances, the model fails to predict the sudden price increase and the following moving trend at the ends of Feb 2023 and Apr 2023.

Figure 2. LightGBM Prediction (Photo/Picture credit: Original).

Fig. 3 shows the true Litecoin prices against the LSTM models' prediction. The blue line represents the true Litecoin prices and the red line represents the LSTM model predicted prices. Comparing with the predictions that random forest and LightGBM models made, LSTM model's prediction is much less accurate. Despite that LSTM model detected the correct moving trends of the data, it normally failed in predicting the amount of increase of decrease of the actual Litecoin price. This has resulted in the large discrepancy between the model's prediction and the real price curve.

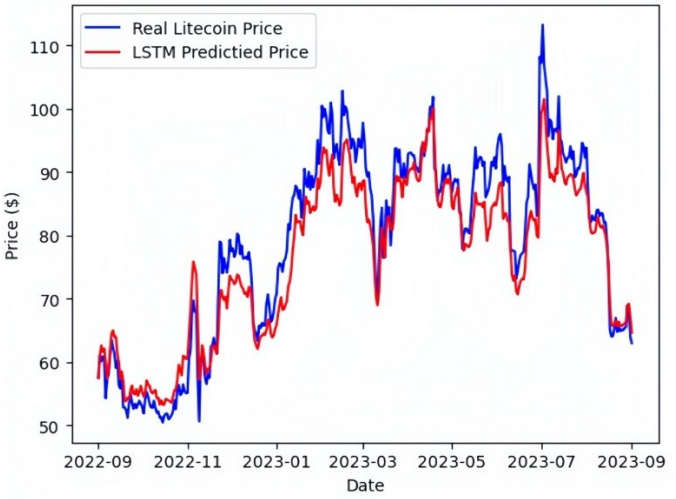

Figure 3. LSTM Prediction (Photo/Picture credit: Original).

Table 2. Evaluation metrics.

	RMSE	MAPE	R-squared
Random Forest	1.459918	1.377161%	0.990782
LightGBM	2.010264	1.896285%	0.982521
LSTM	4.351041	4.342786%	0.918118

DOI: 10.54254/2755-2721/55/20241404

The model's performances can also be reflected from the error metrics. Table 2 shows the evaluation metrics (RMSE, MAPE, and R-squared score) of the three models. Among the three models, random forest has the least RMSE (1.459918) and MAPE (1.377161%) values, and the greatest R-squared score (0.990782). On the contrary, LSTM has the most values of RMSE (4.351041) and MAPE (4.342786%), and the least R-squared score (0.918118). Such results are consistent with the figures in the previous sections. Comparing the metrics of the three models, values of RMSE and MAPE suggest that the random forest prediction has the least error, while LSTM prediction has the most error. Additionally, R-squared score represents the proportion of variance in Litecoin price that can be predicted using the 19 features, and from the values, random forest regression employed the 19 variables best. However, such results also suggest the overfitting problem in random forest model and LightGBM model. In conclusion, among random forest, LightGBM, and LSTM models, random forest model achieved the most accurate prediction with least error.

4. Limitations and prospects

There are limitations to the study. First and foremost, the 19 explanatory features employed to train the models are not the only influencing factors to Litecoin price variations. There are huge amounts of other variables that can affect the price trend, including prices of other commodities or other market indices. Among them, market sentiment can be an influential one. Market sentiment reflects people's willingness to purchase the cryptocurrency, and is closely related to its prices [18]. Without containing these variables, models can fail to reflect certain price fluctuations. Additionally, given that the dataset contains merely daily information in a 5-year range, the results are only applicable for long-period prediction. In other words, the models cannot predict subtle price fluctuations within a day since the data contain only Litecoin close price for each day. In addition, beside the 3 models implemented in this study, there are other algorithms that are effective and efficient under the context. For example, Akila et al. proposed an adjusted LSTM model using Pruned Exact Linear Time (PELT) algorithm, and achieved much higher accuracy than the normal LSTM model [19]. Without implementing those models under the same context, the study cannot make an absolute conclusion that the random forest model is the most accurate one. For future research on long-term cryptocurrency price prediction, more data points can be added. For instance, instead of using daily close price as the target variable, hourly data can be used. This way, the model can be trained to predict short-term fluctuations as well as long-term trends. Subtle price changes can better help investors analyse the risks of the market. Then, more influential explanatory feature should also be considered in the future. Since the price of cryptocurrency is hugely affected by its demand, people's willingness to purchase the cryptocurrency is considered as a significant price influencing factor. Fear and greed index, and other market sentiments can reflect people's willingness, so adding them into the features can result in more accurate model. On the other hand, more input features can also lead to other problems like overfitting, so more elaborate feature engineering process need to be considered in future studies. Finally, to evaluate the models' performances, merely studying one single type of cryptocurrency is not enough. In future works, the more proposed models can be implemented to other cryptocurrency data, so that the most accurate models under different settings, or a potentially most preferable model in general can be concluded.

5. Conclusion

To sum up, this study predicted Litecoin prices through three models: random forest regression, light gradient boosting machine, and long short-term memory network. Data from Sep 1, 2018 to Sep 1, 2023 were employed, where the first four years' data were used to train the models and the last one year's data were used to test the prediction. In the study, 6 different categories of variables including 19 features were used to make the predictions. Among the three models, random forest regression gave the least error scores (RMSE and MAPE) and can explain the most proportion of Litecoin price's variance (most R-squared score). However, the study failed to consider price fluctuations within a day and the effects of market sentiments, and is limited to long-term price prediction of Litecoin. Such limitations can provide insights for future works, in which more types of cryptocurrencies can also be taken into

DOI: 10.54254/2755-2721/55/20241404

consideration. In conclusion, the study investigated long-term Litecoin price prediction using random forest regression, LightGBM, and LSTM. Among them, the random forest model showed the best performance. Such result can provide insights for investors to prevent risks in investments, and for researchers to development more accurate models.

References

[1] Nakamoto S 2008 Decentralized business review vol 11 p 1.
[2] Catania L, Grassi S and Ravazzolo F 2018 Mathematical and statistical methods for actuarial sciences and finance: MAF 2018 pp 203-207) Springer
[3] CoinMarketCap 2023 Cryptocurrency Market Capitalizations | CoinMarketCap CoinMarketCap Retrieved from: https://coinmarketcapcom/
[4] Fang F, Ventre C, Basios M, Kanthan L, Martinez-Rego D, Wu F and Li L 2022 Financial Innovation vol 8(1) pp 1-59.
[5] Chen J 2023 Journal of Risk and Financial Management vol 16(1) p 51.
[6] Sun X, Liu M and Sima Z 2018 Finance Research Letters vol 32 p 101084.
[7] Chowdhury R, Rahman M A, Rahman M S and Mahdy M R C 2020 Physica A: Statistical Mechanics and Its Applications vol 551 p 124569.
[8] Jaquart P, Dann D and Weinhardt C 2021 The Journal of Finance and Data Science vol 7 pp 45–66.
[9] Aggarwal A, Gupta I, Garg N and Goel A 2019 Twelfth International Conference on Contemporary Computing (IC3) p 177.
[10] Chen W, Xu H, Jia L and Gao Y 2021 International Journal of Forecasting vol 37(1) pp 28–43.
[11] Patel M M Tanwar S Gupta R and Kumar N 2020 Journal of Information Security and Applications vol 55 p 102583.
[12] Agarwal B, Harjule P, Chouhan L, Saraswat U, Airan H and Agarwal P 2021 EAI Endorsed Transactions on Industrial Networks and Intelligent Systems vol 8(29) p 171188.
[13] Flach P A 2012 Machine Learning: *The Art and Science of Algorithms that Make Sense of Data*, (Cambridge university press, London).
[14] Liu Y, Wang Y and Zhang J 2012 Information Computing and Applications vol 7 pp 246-252.
[15] Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q and Liu T Y 2017 Advances in neural information processing systems vol 30.
[16] Medsker L R and Jain L C 2000 *Recurrent Neural Networks: Design and Applications* (CRC Press, Berlin)
[17] Hung C L 2023 Intelligent Nanotechnology: Merging Nanoscience and Artificial Intelligence pp 307-329.
[18] Lamon C, Nielsen E and Redondo E 2017 SMU Data Sci. Rev vol 1(3) pp 1-22.
[19] Akila V, Nitin M V S, Prasanth I, Sandeep R M and Akash Kumar G 2023 4th International Conference on Design and Manufacturing Aspects for Sustainable Energy (ICMED-ICMPC 2023) p 17.