Prediction and Analysis of Bitcoin Prices Using Diverse Regression Models

Tianze Li[a]
College of Science and Engineering, The University of Edinburgh, Edinburgh, U.K.

Keywords: Bitcoin, Prediction, Regression Models.

Abstract: Accurate prediction of cryptocurrency prices is crucial for investors and analysts due to the instability and complexity of the trading market. This study explores the effectiveness of diverse predictive models - Long Short-Term Memory (LSTM), eXtreme Gradient Boosting (XGBoost), and Random Forest (RF) - in forecasting Bitcoin prices. The inclusion of these diverse models, each representing different approaches to regression and machine learning, allows for a more comprehensive analysis of predictive accuracy. Simulations are conducted using historical Bitcoin price data from Yahoo Finance, evaluating the models based on their performance metrics: R-squared (R²) score, Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE). The LSTM model demonstrated superior performance with an R² score of 0.955, an MAE of 0.04, and an RMSE of 0.0508, showing its ability to capture complex temporal dependencies. RF also performed well, achieving an R² of 0.936, MAE of 0.0459, and RMSE of 0.0604. In contrast, XGBoost lagged with an R² of 0.654, MAE of 0.1009, and RMSE of 0.1406. This study highlights the strengths of using diverse regression models for predicting Bitcoin prices, with LSTM emerging as the most effective model, while also providing insights into real-market transactions.

1 INTRODUCTION

Bitcoin has been one of the most prevalent cryptocurrencies in the volatile trading market since the 21st century (Wątorek et al., 2021). It is an aggregation of concepts and technologies that form the basis of the digital economic system, and users can transfer Bitcoin over the internet as they handle the conventional currencies (Wątorek et al., 2021). As a result, the emergence of Bitcoin symbolizes not only a technological innovation but also a revolutionary shift in the financial world.

Before Bitcoin, the double spending problem - a challenge where a token could be spent more than one time - bothers the financial market to a large degree (John et al., 2022). This problem is notable in a setting without a centralized entity since a single transaction should transfer ownership of the currency, preventing the original owner from making additional transactions with the same currency (John et al., 2022). However, Bitcoin solved the problem by defining a new mechanism for obtaining consensus in a decentralized background (John et al., 2022).

Given the unique characteristics and the growing role of Bitcoin, it is critical to comprehend its significance and predict its price movements. However, due to various factors in the turbulent trading market, whether economic, social, or technical - accurately forecasting the price of Bitcoin is still a challenge. Previous studies on stock market predictions mainly utilize daily and high-frequency data (Madan et al., 2015). These studies are divided into two categories: analysis of empirical studies and analysis of machine learning algorithms (Chen et al., 2020). According to historical results, the latter performs the prediction task better in accuracy, consistency, and the ability to produce the underlying pattern of the data (Chen et al., 2020). Therefore, by exploring a set of diverse regression models (Long Short-Term Memory (LSTM), eXtreme Gradient Boosting (XGBoost), and Random Forest(RF)), this paper aims to provide several insightful opinions and contributive understandings to the prediction of Bitcoin prices in the cryptocurrency trading market.

This paper is structured with a brief exploratory data analysis in the first place, delivering a

[a] https://orcid.org/0009-0007-5479-8029

Li, T.
Prediction and Analysis of Bitcoin Prices Using Diverse Regression Models.
DOI: 10.5220/0013215000004568
In *Proceedings of the 1st International Conference on E-commerce and Artificial Intelligence (ECAI 2024)*, pages 286-291
ISBN: 978-989-758-726-9

comprehensive view of the original data and basic information. Then, three machine learning models are introduced with quick elaborations, and the results and analysis of this study are presented with detailed illustrations and explanations. After that, this paper discusses the real-life significance of conducting this experiment and provides several suggestions for increasing the accuracy of the results. Finally, it concludes with a summary of key findings and potential directions for future research.

2 EXPERIMENTAL DATA

The dataset used in this simulation was sourced from Yahoo Finance, a popular platform that provides financial news and stock information. The dataset includes historical data on Bitcoin (BTC), with daily information on the opening, closing, high, and low prices, and trading volumes. It spans from September 17th, 2014, to February 19th, 2022, providing an exhaustive view of Bitcoin's market behavior over time.

Normalization is applied before the simulation to scale the features of the dataset to a standard range (typically between 0 and 1). This step is crucial for any machine learning algorithm to ensure that all the features contribute equally to the performance of the model. Also, the missing values/null values are checked during the preprocessing of the data to ensure data integrity and completeness, allowing for accurate and reliable analysis.

The predicted target (output) of the experiment is the close price of Bitcoin, while the input contains more complex information like RSI (relative strength index) transformed from the basic features of the Bitcoin data frame.

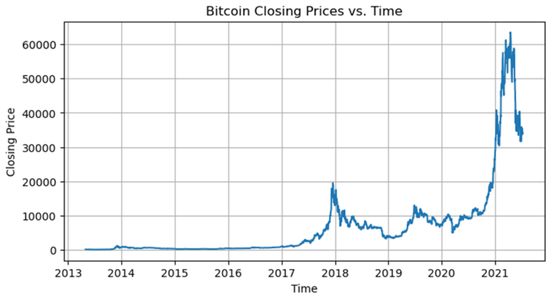

Figure 1: Time series visualization of the Bitcoin close prices (Photo/Picture credit: Original).

Figure 1 offers the time series visualization of Bitcoin prices from September 17th, 2014, to February 19th, 2022. The sudden boom of the close prices around the year 2021 indicates a significant surge in market investment in Bitcoin, reflecting substantial volatility in the cryptocurrency trading market. To reduce the computations of the models and leverage the performances, the study was conducted using data in one year from February 19th, 2021, to February 19th, 2022.

3 REGRESSION MODELS

There are various regression models in the machine learning area, specifically, 3 models are chosen to handle this relatively complex dataset of Bitcoin which are LSTM, XGBoost, and RF. Each model possesses unique strengths but also holds its flaws.

3.1 Long Short-Term Memory

LSTM is a type of recurrent neural network (RNN). Recurrent neural networks are widely utilized in sequential data, but they cannot learn pertinent information when the disparity of the series is large (Yu et al., 2019). By adding the gate function into the cell structure, LSTM is effective in capturing the time series pattern of complicated datasets like cryptocurrencies well (Yu et al., 2019). LSTM also acquires the ability to maintain information over a long time series while limiting the effect of vanishing gradient problems (Yu et al., 2019). This advantage helps to gain a more stable performance during the simulation than normal RNNs do.

3.2 Extreme Gradient Boosting

XGBoost regression model is primarily based on the gradient boosting framework, and it works by combining the predictions of multiple weak learners to form a strong predictive model (Chen & Guestrin, 2016). It is extensively utilized by data scientists to achieve cutting-edge results in numerous machine learning challenges and is particularly appropriate for handling structured data based on its scalability, which can effectively capture non-linear relationships in the dataset (Chen & Guestrin, 2016). The ability to reduce overfitting is highly appreciated, so it is selected to be applied in the experiment to predict Bitcoin prices.

3.3 Random Forest

RF is a powerful prediction tool that combines multiple tree predictors, each relying on a randomly sampled vector (Breiman, 2001). By incorporating the right kind of randomness and adhering to the Law of Large Numbers, they avoid overfitting and become highly accurate classifiers and regressors (Breiman, 2001). Forests produce results that are as effective as boosting and adaptive bagging, but they don't modify the training set over time, which reduces the volatility and fluctuations during the simulation (Breiman, 2001). Previous studies have shown that random inputs and random features yield strong results in classification but act less effectively in regression, which may provide some preconceptions about the upcoming results of the study on the price prediction of Bitcoin.

4 RESULTS AND ANALASIS

4.1 Evaluation Metrics

There are several important metrics to help assess the performance of this experiment using different models, with n representing the total number of the sample and y representing the value.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (1)$$

Equation (1) measures the average value of the prediction error and provides an intuitive sense of the gap between the true values and the prediction values.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (2)$$

Equation (2) shares the same units with the target variable compared to mean squared error (MSE). It is generally used to evaluate and report the performance of the model rather than train the model as MSE does.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (3)$$

Equation (3), the coefficient of determination, measures the proportion of the variation in the dependent variable that is explained by the independent variables in the model. It is indeed an explanatory indicator of the overall effectiveness of the model.

4.2 Performance Evaluation

It is obvious from Figure 3 that XGBoost predictions deviate significantly from the main trend of the original close prices, especially in the fluctuations around November 2021; According to Figure 2 and Figure 4, LSTM and RF models fit relatively well with only a few lines being misaligned with the original stock trend. On one hand, even though the prediction of LSTM looks a bit ahead of time when compared to the original close price, it still stands out of the pack with its extraordinary capability by simulating each tiny movement of the original trend. RF, on the other hand, did not capture the price movement as accurately as LSTM did around November 2021, which symbolizes its weakness in simulating extremely complicated fluctuations.

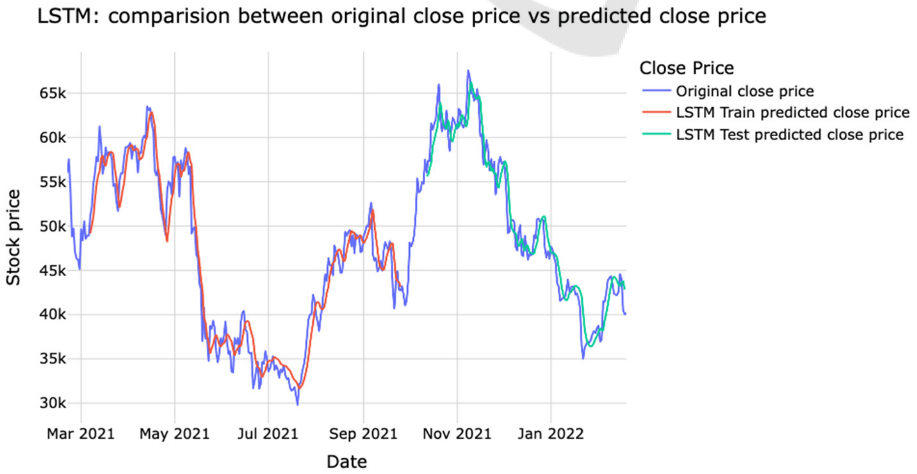

Figure 2: Comparison of prediction close prices with LSTM model vs. original close prices (Photo/Picture credit: Original).

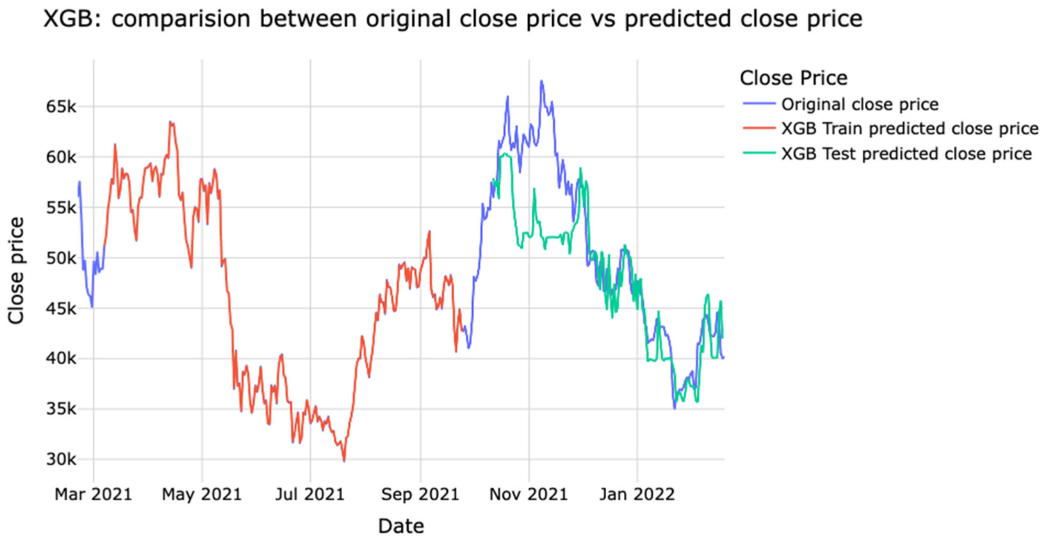

Figure 3: Comparison of prediction close prices with XGB model vs. original close prices (Photo/Picture credit: Original).

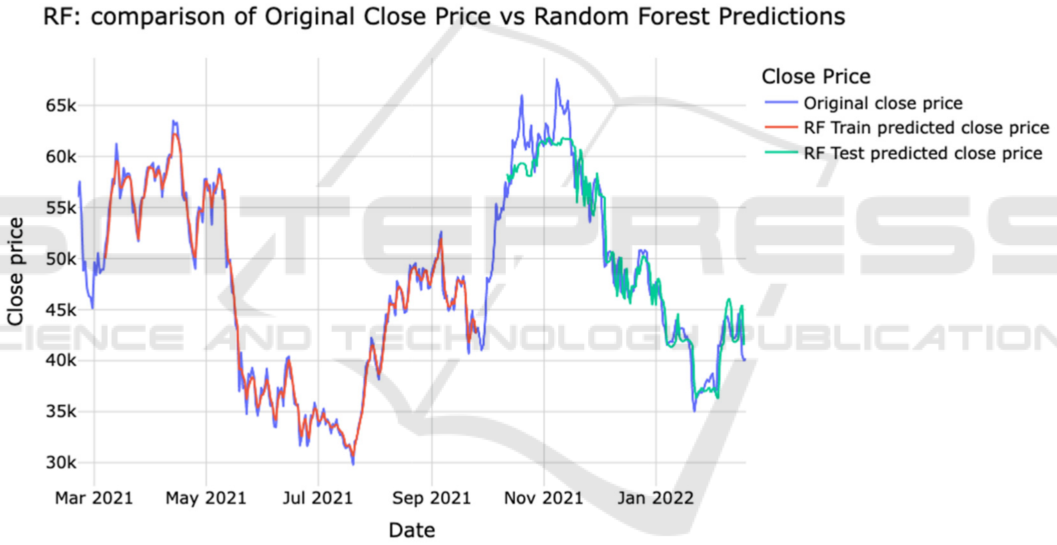

Figure 4: Comparison of prediction close prices with RF model vs. original close prices (Photo/Picture credit: Original).

4.3 Experimental Results and Comparisons

From the results in Table 1, all three models showcase their ability to predict the Bitcoin prices in different ways: LSTM outperforms the other two methods with the highest R-squared (R^2) value and the lowest Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), which proves that it is a relatively effective method in predicting the Bitcoin prices. XGBoost, conversely, fell behind the other two methods by a large degree with nearly 30 percent loss in the R2 value and twice the error generated during the experiment. This phenomenon reveals the incompatibility of this model and the prediction of complicated Bitcoin prices, and it may need further tuning or additional features to help increase its performance. RF is still reliable with a high R2 value and comparatively low errors, rejecting the presumption that it acts less effectively in regression. However, it may not be able to capture the tiny information with high accuracy in the Bitcoin close prices as LSTM does.

Table 1: Metrics of the three models.

	LSTM	XGBoost	RF
R^2	0.9548	0.6541	0.9362
MAE	0.0400	0.1009	0.0459
RMSE	0.0509	0.1407	0.0604

Since normalization has been applied, the numbers in the table shown above are in the range between 0 and 1 to enhance direct comparison between different metrics and models. However, this excellent performance of the LSTM model may be attributed to overfitting due to the large noise under the Bitcoin close prices. This is nearly unavoidable when fitting the price of cryptocurrencies since the market is volatile. Cross-validation could be utilized to mitigate this problem by splitting the dataset into multiple folds and evaluating it.

The data quality itself also determines the reliability of the procedure data preprocessing and eliminates null values. Hence, it is crucial to have a relatively clear dataset to perform the experiment, which will make the whole experimental process more stable and reduce noise to a certain level.

These results provide valuable insights for cryptocurrency investors and market analysts, with LSTM shown to be a preferred model experimentally. However, it is imperative to note the fact that there are significant differences between the simulation and the trading market in real life: real transactions in the market are unexpected and may involve various unknown complexities that are not simulated by the models. As a result, people should be more rational when facing the near-perfect alignment of the model simulation and considering firm orders. Also, risk management and legal problems in the financial market should be realized. Apart from those sides, the results still possess significant implications in analyzing the cryptocurrency trading market.

5 DISCUSSIONS

According to the results presented in the study, researchers should continue to develop new techniques to improve the accuracy of the predictions. One of the most prevalent methods for this aim is called feature engineering, which is to include more relevant features of the Bitcoin prices and weight them in different ratios. This includes adding technical indicators or economic factors to increase the level of fitting and reduce the error. However, it is hard to identify the level of priority of the features in the dataset, and the dynamic trading market may change the importance of each feature, which requires consistent updates of peoples' own feature sets.

Another method of improving the precision is to incorporate deep learning algorithms such as Convolutional Neural Networks (CNNs) or Generative Adversarial Networks (GANs) to simulate the data precisely (Kattenborn et al., 2021). Because of the introduction of non-linearity, the activation functions inherited in CNNs enhance the versatility and capability of the networks to model a broad range of complicated tasks exhibited in reality (Krichen, 2023).

Cross-sectional predictions are also an alternative approach to predicting the price of cryptocurrencies: instead of predicting the close price of the target currency directly, they focus on analyzing the market variables of the currency at a specific moment (Hanauer & Kalsbach, 2023). This method could limit the effect of the outliers and address the impact of differences in the characteristics of the target (Hanauer & Kalsbach, 2023). Thus, prediction accuracy and profitability can be enhanced by applying non-linear combinations through deep learning techniques, rather than relying merely on linear regression to combine various factors (Abe & Nakagawa, 2020).

6 CONCLUSIONS

In this study, the Bitcoin price is predicted by LSTM, XGBoost, and RF models. The paper initially selected the close price as the target variable and chose a specific period from the data time. Then the three models carried out the task with different performances shown above. Finally, they are assessed by multiple metrics and graphs, which demonstrate the best comprehensive quality of the LSTM model. Generally, people have been dedicated to reforming various methods of predicting cryptocurrency prices these years, which implies the importance of accurate forecasting in both commercial and scientific areas. Other methods or refinements should be explored to optimize the actual capability of the models and to achieve more reliable predictions in the dynamic trading market.

REFERENCES

Abe, M., Nakagawa, K., 2020. Cross-sectional stock price prediction using deep learning for actual investment management. In Proceedings of the 2020 Asia Service Sciences and Software Engineering Conference (pp. 9-15).

Breiman, L., 2001. Random forests. Machine learning, 45, 5-32.

Chen, T., Guestrin, C., (2016) Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).

Chen, Z., Li, C., Sun, W., 2020. Bitcoin price prediction using machine learning: An approach to sample dimension engineering. Journal of Computational and Applied Mathematics, 365, 112395.

Hanauer, M. X., Kalsbach, T., 2023. Machine learning and the cross-section of emerging market stock returns. Emerging Markets Review, 55, 101022.

John, K., O'Hara, M., Saleh, F., 2022. Bitcoin and beyond. Annual Review of Financial Economics, 14(1), 95-115.

Kattenborn, T., Leitloff, J., Schiefer, F., Hinz, S., 2021. Review on Convolutional Neural Networks (CNN) in vegetation remote sensing. ISPRS journal of photogrammetry and remote sensing, 173, 24-49.

Krichen, M., 2023. Convolutional neural networks: A survey. Computers, 12(8), 151.

Madan, I., Saluja, S., Zhao, A., 2015. Automated bitcoin trading via machine learning algorithms. URL: http://cs229. stanford. edu/proj2014/Isaac% 20Madan, 20.

Wątorek, M., Drożdż, S., Kwapień, J., Minati, L., Oświęcimka, P., Stanuszek, M., 2021. Multiscale characteristics of the emerging global cryptocurrency market. Physics Reports, 901, 1-82.

Yu, Y., Si, X., Hu, C., Zhang, J., 2019. A review of recurrent neural networks: LSTM cells and network architectures. Neural computation, 31(7), 1235-1270.