Proceedings of the 3rd International Conference on Business and Policy Studies
DOI: 10.54254/2754-1169/79/20241747

Bitcoin Price Prediction Based on Multiple Machine Learning Algorithms

Yaowen Hu[1,a,*]

[1]*College of Liberal Arts, University of Minnesota, Minneapolis, Minnesota, 55414, USA*
a. yaowen_hu01@163.com
[*]*Corresponding author*

Abstract: In this paper, we performed bitcoin price prediction based on bitcoin price dataset using Support Vector Machine model, Random Forest model, Neural Network model, XGBoost model and LightGBM model and evaluated the performance of these models. We divided the Bitcoin price dataset into training and test sets in a ratio of 7:3, where 70 were used as the training set and 30 as the test set. The models were trained with the training set and tested with the test set using the stock price change (yield) as the target variable and other variables as input variables. By comparing the MSE, RMSE, MAE, MAPE and R^2 of the different models were evaluated and it was found that XGBoost has the best performance and the best prediction. The performance of the other four models ranged from good to poor, including LightGBM, Random Forest, Support Vector Machine and Neural Network. Among them, the neural network, whose MSE is tens of times higher than the other four models, performs the worst. The research results in this paper can provide reference value for future Bitcoin price prediction, and also provide some reference for choosing appropriate machine learning models.

Keywords: XGBoost, LightGBM, Bp neural network

1. Introduction

The Bitcoin Price Prediction dataset from the UCL database is a set of time-series data used to predict the price of Bitcoin, including several metrics such as the price of Bitcoin, market capitalization, and trading volume [1]. The purpose of this dataset is to help researchers and data scientists predict the bitcoin price in order to better understand the trends and sentiment of the bitcoin market [2].

Research in machine learning has focused on how to use this dataset to predict the Bitcoin price [3]. Bitcoin price prediction is a very challenging problem because Bitcoin price is very volatile and is affected by many factors such as government policies, media reports, and market sentiment [4]. Therefore, researchers need to use various machine learning algorithms and techniques to deal with this problem [5].

When using this dataset for bitcoin price prediction, researchers usually use some common machine learning algorithms such as linear regression, support vector machine, random forest, etc [6,7]. These algorithms can help researchers build models to predict future bitcoin prices [8]. In addition, there are some emerging machine learning techniques such as deep learning and reinforcement learning that are also used in bitcoin price prediction [9].

DOI: 10.54254/2754-1169/79/20241747

Deep learning is a neural network-based machine learning technique that can process large amounts of data and extract useful features from it. In bitcoin price prediction, researchers can use deep learning algorithms to build complex neural network models to better capture trends and changes in bitcoin prices [10]. Reinforcement learning is a reward-based machine learning technique that helps machine learning systems learn and optimize strategies autonomously. In bitcoin price prediction, researchers can use reinforcement learning algorithms to train machine learning systems to better adapt to changes in the bitcoin market.

Overall, the Bitcoin price prediction dataset from the UCL database provides researchers with a very useful tool that can help them understand the trends and quotations of the Bitcoin market. By using various machine learning algorithms and techniques, researchers can build predictive models in order to better predict the future movements of the Bitcoin price. The applications of this dataset are very promising and can help investors, traders, policy makers, and others to better understand the Bitcoin market and make more informed decisions.

2. Data set overview and feature calculation

2.1. Data set overview

The Bitcoin Price Forecast dataset includes a number of metrics such as Bitcoin price, market capitalization, and trading volume. The purpose of this dataset is to help researchers and data scientists predict the Bitcoin price in order to better understand the trends and sentiment of the Bitcoin market. The source of this dataset is CoinDesk, a specialized digital currency news site. The applications of this dataset are very promising and can help investors, traders, policy makers and others to better understand the bitcoin market and make more informed decisions. The following is the first 15 lines of the Bitcoin Price dataset.

	Open Time	Open	High	Low	Close	Volume	Close Time	Quote asset volume	Number of trades	Taker buy base asset volume	Taker buy quote asset volume
0	2021-01-01 00:13:20	28923.63	28961.66	28913.12	28961.66	27.457032	2021-01-01 00:13:20	7.943820e+05	1292	16.777195	4.853908e+05
1	2021-01-01 00:13:20	28961.67	29017.50	28961.01	29009.91	58.477501	2021-01-01 00:13:20	1.695803e+06	1651	33.733818	9.781765e+05
2	2021-01-01 00:13:20	29009.54	29016.71	28973.58	28989.30	42.470329	2021-01-01 00:13:20	1.231359e+06	986	13.247444	3.840769e+05
3	2021-01-01 00:13:20	28989.68	28999.85	28972.33	28982.69	30.360677	2021-01-01 00:13:20	8.800168e+05	959	9.456028	2.740831e+05
4	2021-01-01 00:13:20	28982.67	28995.93	28971.80	28975.65	24.124339	2021-01-01 00:13:20	6.992262e+05	726	6.814644	1.975194e+05
5	2021-01-01 00:13:20	28975.65	28979.53	28933.16	28937.11	22.396014	2021-01-01 00:13:20	6.483227e+05	952	9.127550	2.642179e+05
6	2021-01-01 00:13:20	28937.11	28963.25	28937.10	28943.87	20.480294	2021-01-01 00:13:20	5.929263e+05	750	5.444172	1.576094e+05
7	2021-01-01 00:13:20	28943.88	28954.48	28930.00	28934.84	20.962343	2021-01-01 00:13:20	6.065811e+05	782	13.154737	3.806512e+05
8	2021-01-01 00:13:20	28934.84	28936.15	28889.24	28900.00	52.645478	2021-01-01 00:13:20	1.521739e+06	886	28.440008	8.218062e+05
9	2021-01-01 00:13:20	28900.00	28920.06	28846.28	28858.94	98.083975	2021-01-01 00:13:20	2.831962e+06	1558	57.594864	1.662928e+06
10	2021-01-01 00:13:20	28858.94	28883.20	28848.06	28848.68	42.665900	2021-01-01 00:13:20	1.231421e+06	825	15.510990	4.477374e+05
11	2021-01-01 00:13:20	28848.69	28862.12	28782.88	28824.35	96.600376	2021-01-01 00:13:20	2.783336e+06	1962	53.027598	1.527954e+06
12	2021-01-01 00:13:20	28824.36	28858.28	28818.75	28838.68	42.540963	2021-01-01 00:13:20	1.226967e+06	1052	18.728299	5.401093e+05
13	2021-01-01 00:13:20	28838.69	28839.08	28706.16	28706.64	104.225054	2021-01-01 00:13:20	2.998588e+06	2635	32.119426	9.243218e+05
14	2021-01-01 00:13:20	28716.85	28764.23	28690.17	28752.80	156.587294	2021-01-01 00:13:20	4.497094e+06	2302	80.527451	2.313473e+06

Figure 1: Data overview.
(Photo credit : Original)

DOI: 10.54254/2754-1169/79/20241747

2.2. Feature calculation

In the feature engineering step, we try to create new features that might help improve the performance of the model. This article creates a feature - return. The return is calculated as the percentage change between the current closing price and the previous closing price. This is a common way to represent changes in stock prices.

A Return is a common way of expressing a change in the price of a stock or other asset. In time series data, returns usually refer to the percentage change between the current price and the price at a previous point in time.

In the Bitcoin price data set, we can use returns as a new feature for predicting future price trends. The formula for calculating returns is as follows:

$$B=(C-C0)/C0 \quad (1)$$

Where B represents the return, C represents the current closing price, and C0 represents the previous closing price.

3. Data set overview and feature calculation

3.1. Support vector machine model

Support Vector Machine (SVM) is a commonly used machine learning algorithm for classification and regression problems. The core idea of SVM is to find an optimal hyperplane that separates data points of different categories. In classification problems, the goal of SVM is to find an optimal hyperplane that maximizes the distance to the nearest hyperplane for different categories of data points. This optimal hyperplane is called the "maximally spaced hyperplane". An important feature of SVM is that it can handle data in high dimensional spaces. In high-dimensional spaces, data points are more easily separated because there are more choices of hyperplanes. SVMs can also use kernel functions to map data into high-dimensional spaces in order to better separate different classes of data points.

The training process of SVM can be divided into two steps. First, a suitable kernel function and hyperparameters need to be selected in order to construct a suitable model. Secondly, the model needs to be trained using the training data and adjusted according to the performance of the model. During the training process, SVM uses optimization algorithms such as gradient descent to minimize the loss function in order to find the optimal hyperplane.

Overall, SVM is a commonly used machine learning algorithm for classification and regression problems. The core idea of SVM is to find an optimal hyperplane that separates data points of different classes. The advantage of SVM is that it can deal with data in high-dimensional spaces and has good generalization ability. However, SVM takes a long time to train, requires the selection of appropriate kernel functions and hyperparameters, and is sensitive to outliers.

3.2. Random forest model

Random Forest is an integrated learning algorithm that performs prediction and classification by constructing multiple decision trees. Random Forest reduces the risk of overfitting and improves the accuracy of the model compared to a single decision tree. The core idea of Random Forest is to construct multiple decision trees by randomly selecting features and samples, and finally voting or averaging to arrive at the final prediction.

One of the advantages of random forest is that it can handle high dimensional, large scale data and does not require feature selection. In random forests, each decision tree uses only a portion of the samples and features of the dataset, which helps reduce the risk of overfitting. Additionally, random

DOI: 10.54254/2754-1169/79/20241747

forests provide an importance ranking of features, which helps us understand which features in the dataset have the greatest impact on the prediction results.

3.3. Neural network model

A neural network is a computational model similar to the human nervous system, which consists of multiple neurons and carries out information transfer and processing through the connections between these neurons. A neural network can learn to automatically adjust its own weights and biases to achieve tasks such as classification, prediction and recognition of input data.

One of the advantages of neural networks is that they can handle nonlinear, high-dimensional data and do not require feature selection. In a neural network, each neuron can adaptively extract features by learning to better fit the input data. In addition, neural networks can increase the complexity and accuracy of the model by increasing the number of hidden layers and neurons.

Neural networks are powerful machine learning algorithms that can be used to solve a variety of classification, prediction, and recognition problems. Although neural networks have some drawbacks, their advantages far outweigh the disadvantages, and thus they are widely used in various fields. With the continuous progress of computational resources and technology, the application prospect of neural networks will be even broader.

3.4. XGBoost model

XGBoost is an integrated learning algorithm based on decision trees, which performs prediction and classification by constructing multiple decision trees. Compared with traditional decision trees, XGBoost can reduce the risk of overfitting and improve the accuracy of the model.The core idea of XGBoost is to train multiple decision trees through a weighted loss function and arrive at the final prediction by weighted averaging.

XGBoost can handle high-dimensional, large-scale data and excels in handling sparse data. In XGBoost, each decision tree is trained with a weighted loss function, which helps reduce the risk of overfitting. Additionally, XGBoost can provide an importance ranking of features, which helps us understand which features in the dataset have the greatest impact on the prediction results.

Since XGBoost is an integrated learning algorithm, it is computationally more expensive than a single decision tree. Secondly, XGBoost requires parameter tuning, which may lead to large biases in the model if the parameters are not set properly. XGBoost is sensitive to the quality and quantity of data, which may affect the performance of the model if the input data is noisy or missing.

XGBoost is a powerful machine learning algorithm that can be used to solve a variety of classification and regression problems. Although XGBoost has some drawbacks, its advantages far outweigh its disadvantages, and thus it is widely used in various fields. With the continuous progress of computational resources and technology, the application prospect of XGBoost will be even broader.

DOI: 10.54254/2754-1169/79/20241747

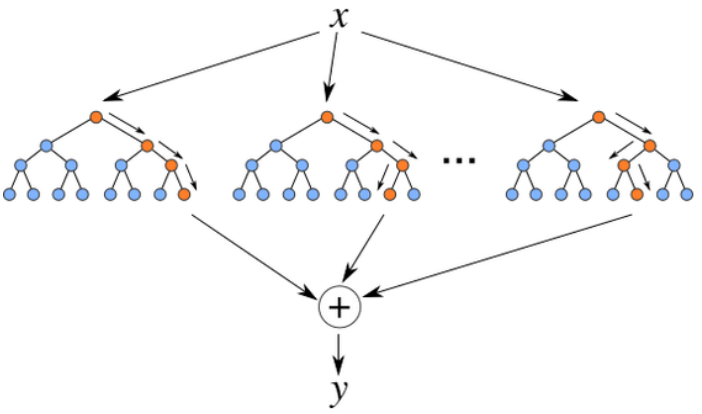

Figure 2: XGBoost model.
(Photo credit : Original)

XGBoost has high accuracy and generalization capabilities, and can handle large datasets and high-dimensional data.

3.5. LightGBM model

LightGBM is also a gradient-lift tree based model that uses a histogram based decision tree algorithm and mutually exclusive feature bunding to speed up the model's training and prediction process. LightGBM has high accuracy and generalization ability, and can handle large data sets and high-dimensional data.

4. Data set overview and feature calculation

The Bitcoin Price dataset was divided according to the ratio of 7:3, with 70% as the training set, 30% as the test set, stock price change (return) as the target variable, and other variables as the input variable. The model was trained with the training set and tested with the test set. The MSE, RMSE, MAE, MAPE and R^2 of each model were calculated for subsequent model evaluation.

Table 1: Model evaluation parameter.

Support vector machine	MSE	RMSE	MAE	MAPE	R^2
Training set	1.938	1.392	1.353	103.974	-133.917
Test set	1.944	1.394	1.36	104.208	-145.967

Table 2: Model evaluation parameter.

Random forest	MSE	RMSE	MAE	MAPE	R^2
Training set	0.009	0.097	0.071	1551.626	0.333
Test set	0.012	0.109	0.074	1269.72	0.155

Table 3: Model evaluation parameter.

Neural network	MSE	RMSE	MAE	MAPE	R^2
Training set	77.623	8.81	6.203	101.018	-5481.082
Test set	76.059	8.721	6.165	100.499	-5545.096

DOI: 10.54254/2754-1169/79/20241747

Table 4: Model evaluation parameter.

XGBoost	MSE	RMSE	MAE	MAPE	R^2
Training set	0.003	0.057	0.042	708.643	0.773
Test set	0.007	0.084	0.056	707.592	0.463

Table 5: Model evaluation parameter.

LightGBM	MSE	RMSE	MAE	MAPE	R^2
Training set	0.006	0.079	0.056	1094.247	0.582
Test set	0.008	0.089	0.067	2056.323	0.34

In the prediction of Bitcoin stock price change, from the evaluation parameters of each model, it can be seen that MSE, RMSE, MAE, MAPE and R^2 of XGBoost are the best, and the prediction effect is also the best. The other four models range from good to differential LightGBM, random forest and support vector machine, and the performance of neural network is the worst. MSE is dozens of times larger than the other four models.

5. Conclusion

Bitcoin is a digital currency with high price volatility, so the prediction of Bitcoin stock price changes has been one of the hot topics in research. Various machine learning models are widely used in bitcoin stock price prediction, including XGBoost, LightGBM, Random Forest, Support Vector Machines, and Neural Networks. Each of these models has its own advantages and disadvantages, so they need to be evaluated and compared to determine which model is best suited for Bitcoin stock price prediction.

First, let's look at the XGBoost and LightGBM models. Both models are integrated learning algorithms based on decision trees with strong prediction capabilities. XGBoost is a gradient boosting decision tree algorithm that integrates multiple weak classifiers into a single strong classifier through continuous iteration, thereby improving prediction accuracy. LightGBM is a gradient one-sided sampling-based decision tree algorithm that accelerates the sample features by sampling the the training process of the decision tree, thus improving the efficiency of the model. In principle, both XGBoost and LightGBM are integrated learning algorithms based on decision trees, but they use different techniques in the training process, which makes their predictions different. In Bitcoin stock price prediction, both XGBoost and LightGBM have good prediction effect with MSE of 0.007 and 0.008 respectively, which shows that they have strong prediction ability in Bitcoin stock price prediction.

Next, let's look at the random forest model. Random forest is an integrated learning algorithm based on decision trees, which constructs multiple decision trees by random sampling and feature selection on the dataset to improve the prediction accuracy. Compared to XGBoost and LightGBM, Random Forest has a simpler training process, but it also has good prediction results. In Bitcoin stock price prediction, the prediction effect of Random Forest is MSE of 0.012, which is better than the prediction effect of Support Vector Machine.

Next, let's look at the support vector machine model. Support Vector Machine is a machine learning algorithm based on Maximum Spaced Classification, which achieves classification or regression by mapping data to a high-dimensional space and constructing an optimal hyperplane. In Bitcoin stock price prediction, the support vector machine has a poor prediction with an MSE of 1.944. This may be due to the fact that the support vector machine model is more sensitive to noise and outliers in the dataset, which makes it prone to overfitting.

DOI: 10.54254/2754-1169/79/20241747

Finally, we look at the neural network model. A neural network is a machine learning algorithm based on artificial neurons, which realizes classification or regression of data by building a multi-layer neuron network. In Bitcoin stock price prediction, the neural network model has the worst prediction effect, with an MSE of 76.059. This may be due to the fact that the neural network model is more sensitive to the noise and outliers in the dataset, and is prone to the phenomenon of overfitting.

In summary, XGBoost and LightGBM have the best prediction effect, which may be due to their strong prediction ability by using the integrated learning algorithm based on decision tree. The prediction effect of Random Forest is also better, which may be due to the fact that it adopts the method of random sampling and feature selection to construct multiple decision trees, which improves the prediction accuracy. Support Vector Machines have poorer prediction results, which may be due to the fact that it is more sensitive to noise and outliers in the dataset, and is prone to overfitting. The neural network model has the worst prediction effect, which may be due to the fact that it is more sensitive to noise and outliers in the dataset and is prone to overfitting. Therefore, in Bitcoin stock price prediction, we should choose models such as XGBoost, LightGBM or Random Forest to get better prediction results.

References

[1] *Jiyang C ,Sunil T ,Djebbouri K , et al.Forecasting Bitcoin prices using artificial intelligence: Combination of ML, SARIMA, and Facebook Prophet models[J].Technological Forecasting Social Change,2024,198122938-.*

[2] *Harish K ,Sudhir S ,P. N , et al.A two level ensemble classification approach to forecast bitcoin prices[J].Kybernetes,2023,52(11):5041-5067.*

[3] *Ruchi G ,E. J N .Metaheuristic Assisted Hybrid Classifier for Bitcoin Price Prediction[J].Cybernetics and Systems,2023,54(7):1037-1061.*

[4] *Tyson M .Use TensorFlow to predict Bitcoin prices[J].InfoWorld.com,2023.*

[5] *Sina F .Designing a forecasting assistant of the Bitcoin price based on deep learning using market sentiment analysis and multiple feature extraction[J].Soft Computing,2023,27(24):18803-18827.*

[6] *Moinak M ,B. D V ,Michael F .Quantifying the asymmetric information flow between Bitcoin prices and electricity consumption[J].Finance Research Letters,2023,57.*

[7] *Xiangling W ,Shusheng D .The impact of the Bitcoin price on carbon neutrality: Evidence from futures markets[J].Finance Research Letters,2023,56.*

[8] *W. J G ,Sami J B ,Foued S , et al.Explainable artificial intelligence modeling to forecast bitcoin prices[J].International Review of Financial Analysis,2023,88.*

[9] *Brahim G ,Sahbi M N ,Jean-Michel S , et al.Interactions between investors' fear and greed sentiment and Bitcoin prices[J].North American Journal of Economics and Finance,2023,67.*

[10] *Zaman S ,Yaqub U ,Saleem T .Analysis of Bitcoin's price spike in context of Elon Musk's Twitter activity[J].Global Knowledge Memory and Communication,2023,72(4/5):341-355.*