CRYPTOCURRENCY PRICE PREDICTION MODEL DEVELOPMENT USING MACHINE LEARNING ALGORITHMS

Gita Maliukaitė
Vilnius University, Kaunas Faculty,
Muitinės St. 8, Kaunas
gita.maliukaite@knf.stud.vu.lt

Mantas Vaitonis
Vilnius University, Kaunas Faculty,
Muitinės St. 8, Kaunas
Mantas.Vaitonis@knf.vu.lt

INTRODUCTION

Cryptocurrencies, starting with Satoshi Nakamoto's distribution of the first BTC in 2009, have created a unique financial shift. This digital currency, which exists in virtual space, has transformed the world of finance by allowing people to pay for goods and services without central banks.

Cryptocurrencies offer advantages such as confidentiality, fast settlements, cheaper financial services and investment opportunities. However, due to highly volatile prices, investors face a higher risk of losing money. Nevertheless, cryptocurrencies are attractive to traders because of the possibility of higher returns. The complexity of these price fluctuations makes price forecasting a challenge, prompting the search for models to help investors predict cryptocurrency price movements.

THE OBJECT OF RESEARCH

The focus of this study is to apply machine learning techniques to predict the close prices of BTC and ETH cryptocurrencies. The main objective is to evaluate the accuracy of these methods using RMSE, MAE and R2 criteria. Data collected on Gemini and Bitstamp exchanges for BTC and ETH from 2021-01-01 to 2023-05-30 will be used to develop the model in order to obtain reliable results. Different researchers use different machine learning algorithms to predict cryptocurrency prices. For example, Roy, Nanjiba and Chakrabarty (2018) used an autoregressive integrated moving average (ARIMA) model to predict BTC prices. Shin, Mohaisen and Kim (2021) used a long short-term memory (LSTM) approach to forecast BTC prices over 3 minutes, an hour and a day. Their results showed that the best accuracy was achieved when predicting 4-hour prices using LSTM. Most researchers often use RMSE, MAPE and MAE criteria to assess price accuracy.

THE ALGORITHMS OF MACHINE LEARNING

Linear regression is a technique that allows you to analyse and predict the relationship between one or more independent variables and a dependent variable. The multiple linear regression used in this study includes one dependent variable and more than one independent variable, allowing analysis of the complex effects of the interaction between variables. The multiple linear regression used in this study includes one dependent variable and more than one independent variable. Multiple linear regression allows you to identify and assess the impact of each independent variable on the dependent variable, determine their statistical significance and construct predictions based on the resulting model parameters. Linear regression equation:

$y = b_0 + b_1x_1 + \cdots + b_px_p + \varepsilon$;

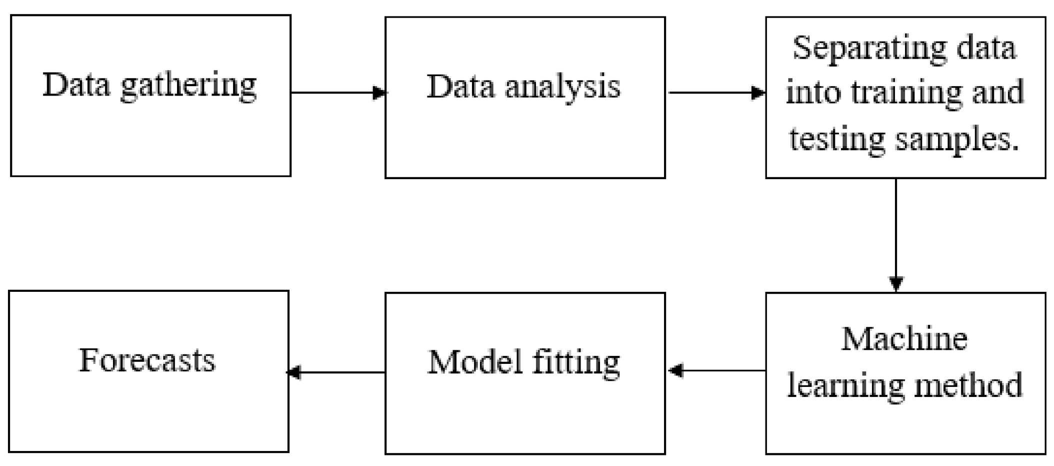
where y is the dependent variable;

$b_0, b_1, \ldots, b_p$- regression coefficients;

$x_1, x_2, \ldots, x_n$ – independent variables.

Gaussian Process Regression (GPR) is a powerful method for performing non-parametric regression using Gaussian processes. The reason this method is non-parametric is that the Gaussian process seeks to build a model that does not restrict the form of the data dependence, but tries to flexibly adapt the regression function to the available data. GPR uses data correlation rather than a linear relationship as linear regression does.

DEVELOPING A MODEL FOR FORECASTING CLOSING PRICES



EXPERIMENTAL SETUP

The cryptocurrency price prediction model uses data collected from the cryptodatadownload.com platform. For this phase of the study, it was chosen to use two years of BTC and ETH data from the Gemini and Bitstamp cryptocurrency exchanges. The data samples include hourly price indices from 29 May 2021 to 29 May 2023. Both datasets consist of 17544 rows of data, which include 9 different variables in each currency sample.

unix	date	symbol	open	high	low	close	Volume.BTC	Volume.USD
1.685401e+12	2023-05-29 23:00:00	BTC/USD	27690.28	27814.40	27690.28	27743.95	3.7827361	104948.04
1.685398e+12	2023-05-29 22:00:00	BTC/USD	27678.17	27692.63	27572.55	27690.28	6.3581394	176058.66
1.685394e+12	2023-05-29 21:00:00	BTC/USD	27696.93	27735.40	27669.17	27678.17	1.8426129	51000.15
1.685390e+12	2023-05-29 20:00:00	BTC/USD	27651.26	27711.11	27647.94	27696.93	2.8303242	78391.29
1.685387e+12	2023-05-29 19:00:00	BTC/USD	27646.99	27708.00	27644.08	27651.26	3.1561031	87270.23
1.685383e+12	2023-05-29 18:00:00	BTC/USD	27700.51	27730.94	27622.44	27646.99	9.2761794	256458.44
1.685380e+12	2023-05-29 17:00:00	BTC/USD	27604.43	27745.11	27545.78	27700.51	11.1361118	308475.98
1.685376e+12	2023-05-29 16:00:00	BTC/USD	27632.44	27677.66	27555.00	27604.43	9.7004488	267775.36
1.685372e+12	2023-05-29 15:00:00	BTC/USD	27774.57	27844.69	27603.91	27632.44	8.6564882	239199.89
1.685369e+12	2023-05-29 14:00:00	BTC/USD	27915.97	27936.51	27701.02	27774.57	9.3219495	258913.14
1.685365e+12	2023-05-29 13:00:00	BTC/USD	27901.01	27954.60	27885.72	27915.97	4.0025033	111733.76
1.685362e+12	2023-05-29 12:00:00	BTC/USD	27926.47	27954.92	27880.46	27901.01	5.0992034	142272.92
1.685358e+12	2023-05-29 11:00:00	BTC/USD	27841.11	27949.37	27841.11	27926.47	3.5801817	99981.84
1.685354e+12	2023-05-29 10:00:00	BTC/USD	27942.15	27970.07	27762.46	27841.11	3.6404657	101354.61
1.685351e+12	2023-05-29 09:00:00	BTC/USD	27905.69	27942.15	27871.74	27942.15	0.7322099	20459.52

Data processing takes place in the RStudio environment. First, the date variables are decomposed into three levels, Years, Months, Days and Hours, with corresponding columns 3, 12, 31 and 24. These columns are identified by factors, allowing them to be included in the analysis. A linear regression is then carried out and statistically significant variables are selected in each data set.

RESULTS

The data were split into training and testing samples with a 70%/30% split. Linear regression was run in RStudio, while Gaussian process regression was implemented using MATLAB.

Data	Features	
BTC from Gemini	High, open, low, Volume.BTC, Volume.USD, 07 hour	**P value < 0,05**
ETH from Gemini	High, open, low, Volume.BTC, Volume.USD, 05 month, 07 hour	
BTC from Bitstamp	High, open, low, Volume.BTC, Volume.USD, 10 month, 11 month, 18 hour	
ETH from Bitstamp	High, open, low, Volume.BTC, Volume.USD, 05 month	

	Linear regression			**Gaussian Process Regression**		
Data	**MAE**	**RMSE**	**R-Squered**	**MAE**	**RMSE**	**R-Squered**
BTC from Gemini	32,791	56,836	0,999	32,621	55,373	0,999
ETH from Gemini	2,753	4,283	0,999	648,335	730,481	-6,968
BTC from Bitstamp	33,645	56,615	0,999	32,426	57,166	0,999
ETH from Bitstamp	2,725	4,094	0,999	735,716	831,499	-9,295

Gaussian Process Regression for BTC from Gemini		
Data	**Predicted value**	**True value**
29/05/2023 23:00	27781,07	27743,95
29/05/2023 22:00	27610,08	27690,28
29/05/2023 21:00	27702,24	27678,17
29/05/2023 20:00	27690,99	27696,93
29/05/2023 19:00	27688,19	27651,26
29/05/2023 18:00	27663,61	27646,99
29/05/2023 17:00	27666,22	27700,51

Gaussian Process Regression for ETH from Gemini		
Data	**Predicted value**	**True value**
29/05/2023 23:00	2316,57	1892,89
29/05/2023 22:00	2140,98	1892,46
29/05/2023 21:00	2115,29	1891,34
29/05/2023 20:00	2129,63	1894,43
29/05/2023 19:00	2110,57	1890,24
29/05/2023 18:00	2075,47	1891,65
29/05/2023 17:00	2194,58	1890,76

CONCLUSIONS

1. The potential for higher returns and the increasing consumerisation of cryptocurrencies have led to a need for models that predict prices. This is driving the search for the best model, which can be built on machine learning techniques.
2. Before applying the data to the prediction of the closing price of BTC and ETH cryptocurrencies, it is necessary to organise them and select which variables are statistically significant.
3. A linear regression showed that the model more accurately predicts ETH prices, than Gaussian process regression. The biggest margin of error is around 0,29%. However, Gaussian process regression is not suitable for predicting ETH closing prices, as it has an error of around 22,38%. In contrast, the prediction of BTC closing prices by Gaussian Process Regression showed an error of only about 0,29%.
4. Further work will be carried out by fitting the data with the Long-Short Term Memory method.