Supervised Learning

Classification of Bitcoin Price Development

Lukas Hermans

Università degli Studi di Milano
lukas.hermans@studenti.unimi.it

June 23, 2021

Abstract

- rising importance of Bitcoin (market cap comparable to large multinational companies such as Facebook and Tesla) - volatility higher than stockmarket, different asset with different underlying mechanisms, relativily new, Blockchain - current object of research: Bitcoin price prediction offers trader advantage over non-rational traders - here: focus on classification of Bitcoin price (up or down) - techniques: majority predictor, logistic regression, KNN, neural network - state test accuracies for best predictors - subtle advantages of different predictors

Contents

1 Introduction

In recent years, cryptocurrencies such as Bitcoin, Ether, and Dogecoin have gained increasing attention in the financial world. In particular, the first cryptocurrency called Bitcoin - developed by an unknown individual or group under the pseudonym Satoshi Nakamoto and first presented in 2009 [1] - has achieved a market cap of about 750 billion United States dollar (USD) as of 14 June 2021. This is comparable to the market cap at the stock market of companies like Facebook and Tesla [2][3].

Bitcoin is mostly used as a form of digital money. Each Bitcoin transaction between two parties is stored on the Bitcoin blockchain. A simplified illustration of the Bitcoin blockchain is depicted in Fig. 1. The Bitcoin blockchain is an ordered chain of blocks, where each block contains data on Bitcoin transactions. New blocks are created by investing computing power in solving a mathematical task. The person or group that solves the task first is rewarded with new Bitcoins as well as with the transaction fees of the transactions in the new block. The new block is appended to the Bitcoin blockchain. This whole process is called Bitcoin mining because it implies the creation of new Bitcoins.

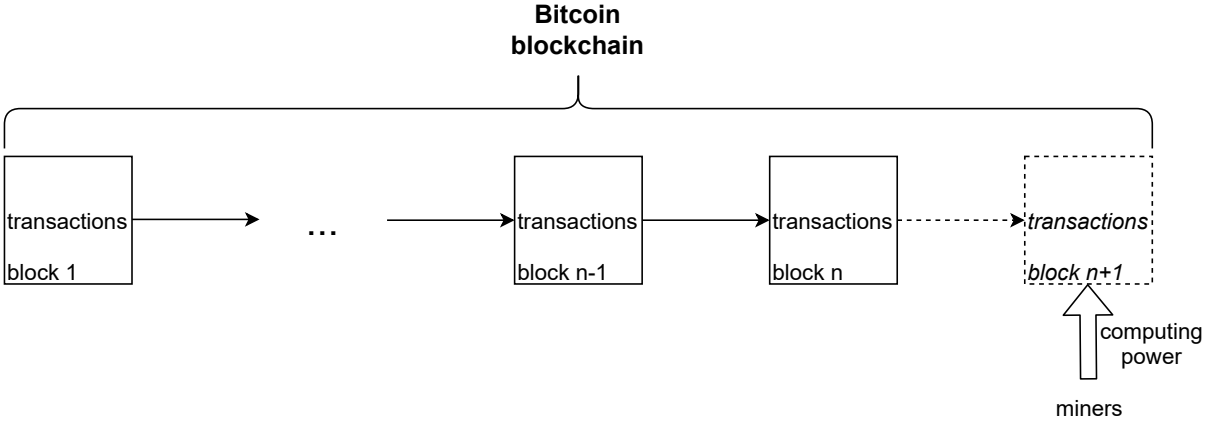


Figure 1: Simplified illustration of the Bitcoin blockchain. The blocks contain all transactions made with the cryptocurrency Bitcoin. A new block $n + 1$ is created by investing computing power to solve a mathematical task. The person or group that solves this task first is rewarded with new Bitcoins as well as with the transactions fees of the transactions in the new block. The new block $n + 1$ is appended to the Bitcoin blockchain. This process is called Bitcoin mining as it implies the creation of new Bitcoins.

Besides its original purpose as a form of digital money, Bitcoins are treated as a financial asset that is exchanged with other currencies, e.g. USD, at cryptocurrency exchanges such as Binance, Kraken, and Coinbase. The market price of Bitcoin - e.g. expressed in USD - is a result of demand and supply as the creation of Bitcoins requires computing power as its underlying value.

For traders, it is of interest to be able to predict the future development of the Bitcoin market price in order to maximize returns. The present work is concerned with the prediction of the direction of the Bitcoin market price one day into the future based on information from the current and past days. The direction is either “up” if the Bitcoin market price is expected to increase, or “down” if it is expected to decrease. Thus, here, the prediction of the Bitcoin market price is treated as a binary classification problem.

In Section 2, a time series dataset of the Bitcoin market price and related measures is developed. A transformation of this dataset to the above described classification task allows an application of supervised machine learning techniques. In the present paper, a simple majority predictor, a logistic regression, the K-nearest neighbors algorithm, as well as a deep neural network are applied. A brief description of these techniques is given in Section 3. In addition, the split of the dataset into training, validation, and test examples as well as cross-validation as evaluation techniques are presented. In Section 4, a list of the applied software is given, and particularities of the implementation are described. The classification results are described in Section 5. A comparison of the different classification techniques is given in Section 8. Finally, in Section 9, the results of the present work are summarized and an outlook on further objects of research is given.

2 Dataset

As described in the last Section, the Bitcoin market price prediction problem in the present work is treated as a binary classification problem. To the best knowledge of the author of the present work, there is no standard dataset that could be applied directly for this purpose. Hence, this Section is concerned with the development of a dataset that can be used for the binary classification of the Bitcoin market price. The basis are day-by-day time series data of the Bitcoin market price as well as related measures that are prepared in the first Subsection. In doing so, features that seem to be usefule for the Bitcoin market price prediction are identified. Then, in the second Subsection, the time series data is transformed into a dataset suitable for binary classification algorithms.

2.1 Bitcoin Time Series Data

The Bitcoin time series data is taken from the Blockchain.com API [4]. In the present work, the time period from 01 September 2011 to 15 June 2021 is considered. Hence, the time series dataset contains the Bitcoin market price as well as related measures visualized in Fig. 2 for a total of 3576 days. The Bitcoin market price is reported in 10^4 USD. It is the average of the Bitcoin trading price of several cryptocurrency exchanges. The market cap is stated in 10^{11} USD and describes the value of all Bitcoins that have ever been mined in USD, similar to the market cap of companies at the stock market. The number of transactions is measured in units of 10^5. The transaction volume - reported in 10^6 Bitcoin (BTC) - is the total number of Bitcoins transferred per day. The average blocksize in megabyte (MB) as well as the relative difficulty to mine blocks measure the complexity of mining new blocks. The hash rate expressed in EH/s is a measure of the computing power that all miners combined are applying. The miner's revenue in 10^7 USD is the reward that miner's obtain for computing a new block. It is the sum of the newly created Bitcoins and the transaction fees of the transactions in the new block.
Fig. 3 shows the correlation matrix of the above described time series dataset. The diagonal - that is the correlation of each time series to itself - is not displayed. The stated values are Pearson's correlation coefficient. The market cap, the difficult, and the miner's revenue reveal a rather high correlation with the market price. These measures are thus not further considered because they do not appear to contain additional information that might be helpful in the prediction of the future Bitcoin market price. In contrast, the transaction volume seems to be uncorrelated to the market price and the other measures. It is also dropped.

In Fig. 4, an overview of the remaining measures that seem to be useful for the prediction of the future Bitcoin market price are displayed.
Finally, Fig. 5 shows the autocorrelation of the Bitcoin market price. It is a measure for how much the Bitcoin market price is related to itself for a given lag.

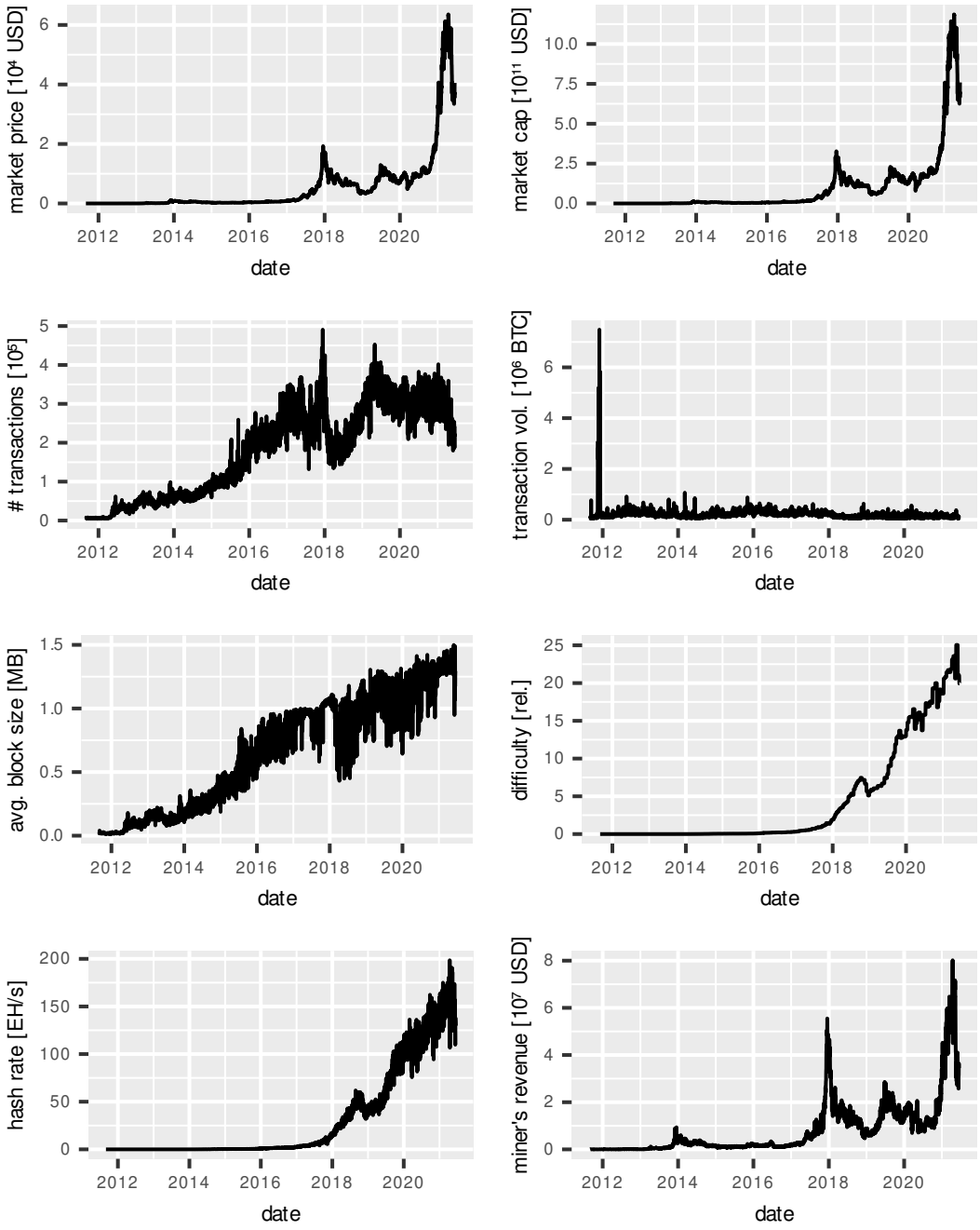


Figure 2: Bitcoin time series data taken from the Blockchain.com API [4]. The time series dataset contains information on the Bitcoin market price and the other displayed measures for the time period from 01 September 2011 to 15 June 2021. This is a total of 3576 days. The different measures are explained in the main text.

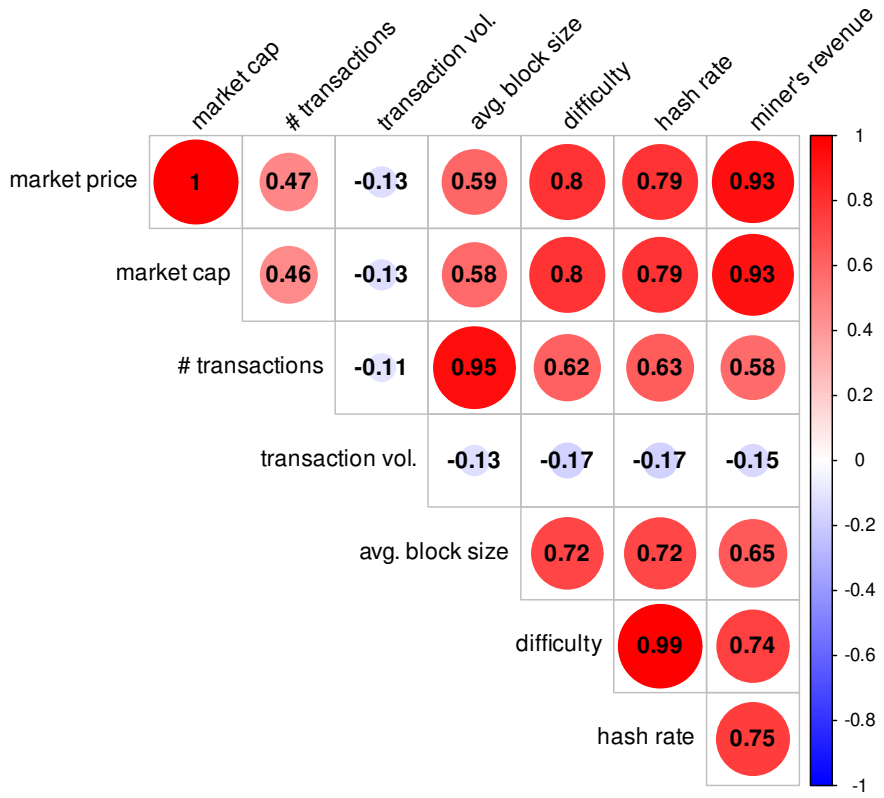


Figure 3: Correlation matrix of the time series from Fig. 2. The displayed values are Pearson's correlation coefficient. Due to their high correlation to the Bitcoin market price, the time series of the market cap, the difficulty, and the miner's revenue are not further considered. In addition, the transaction volume is dropped because it has only a weak correlation to the Bitcoin market price. An overview of the remaining measures can be found in Fig. 4.

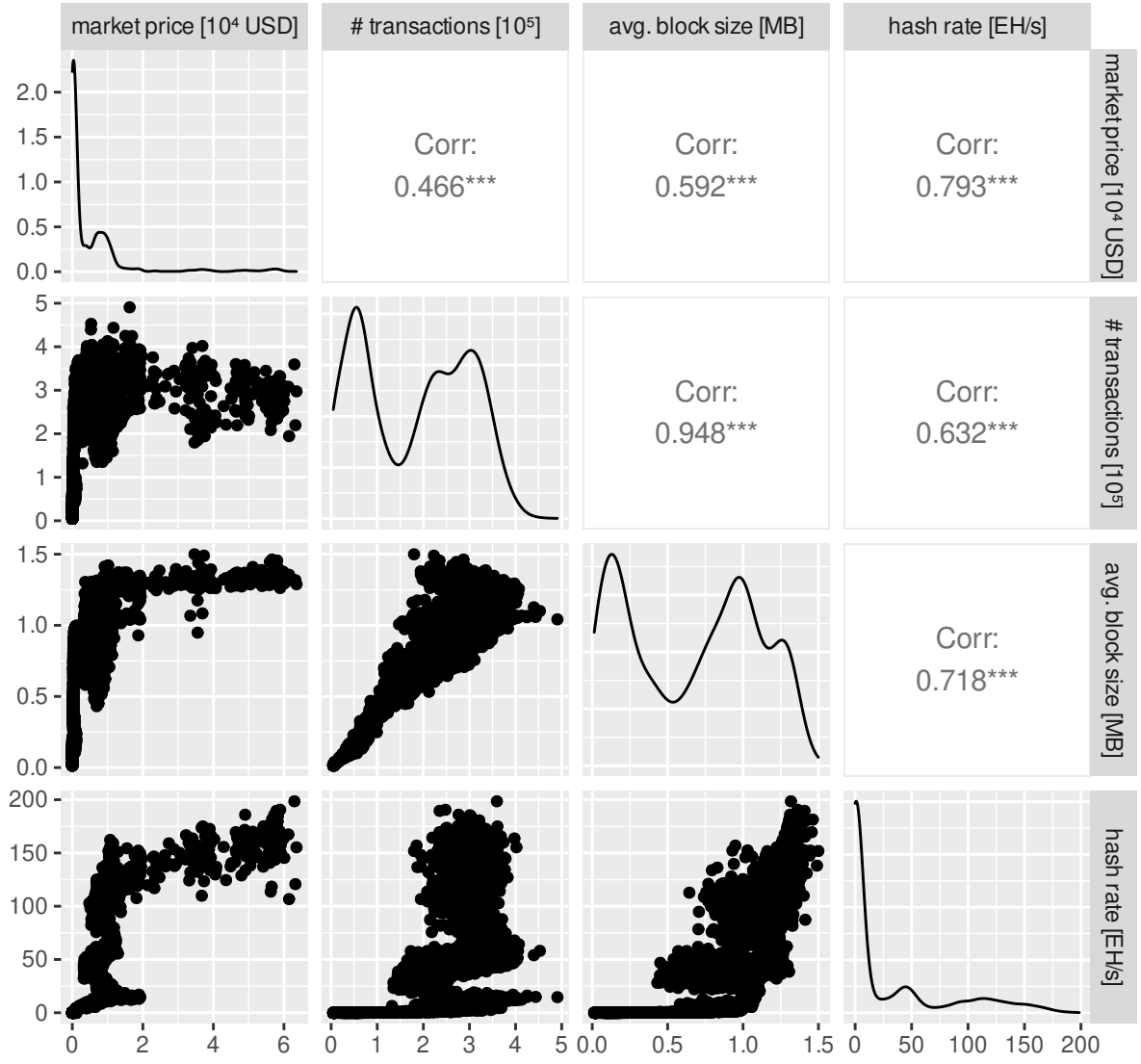


Figure 4: Plot matrix of further considered measures that appear to be useful for the prediction of the future Bitcoin market price. The line plots on the diagonal are the densities of the respective measures. In addition, Pearson's correlation coefficient - the same from Fig. 3 - is shown. It is observed that the number of transactions, the average block size, and the hash rate reveals a significant correlation to the market price.

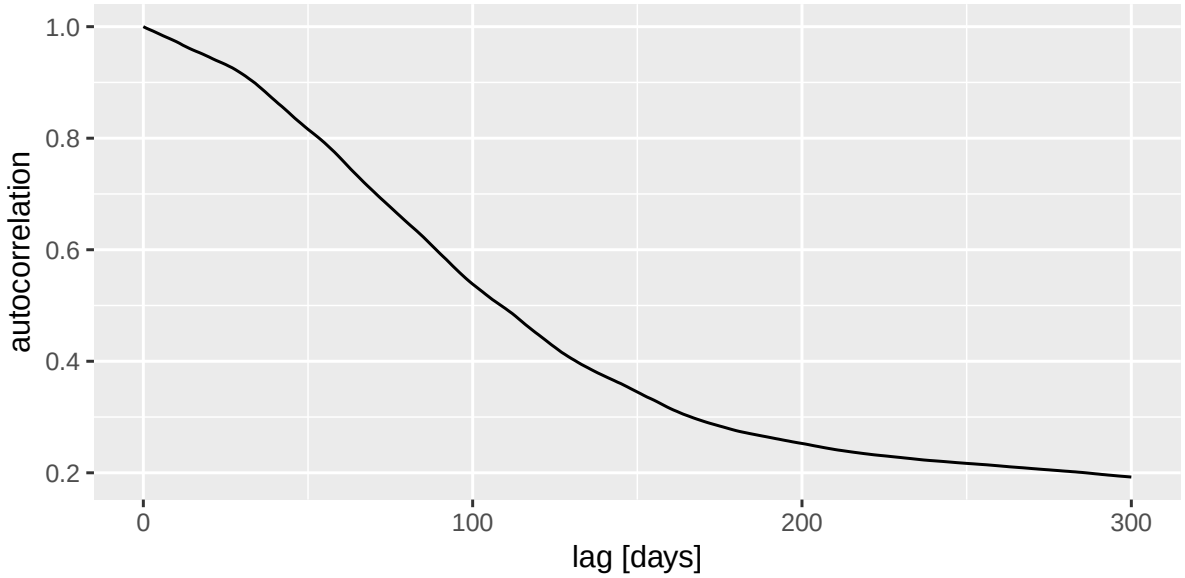


Figure 5: Autocorrelation of the Bitcoin market price. The Bitcoin market price has an autocorrelation larger than 0.5 for a lag of 100 days. It is expected that prediction algorithms make use of this time interval for their predictions.

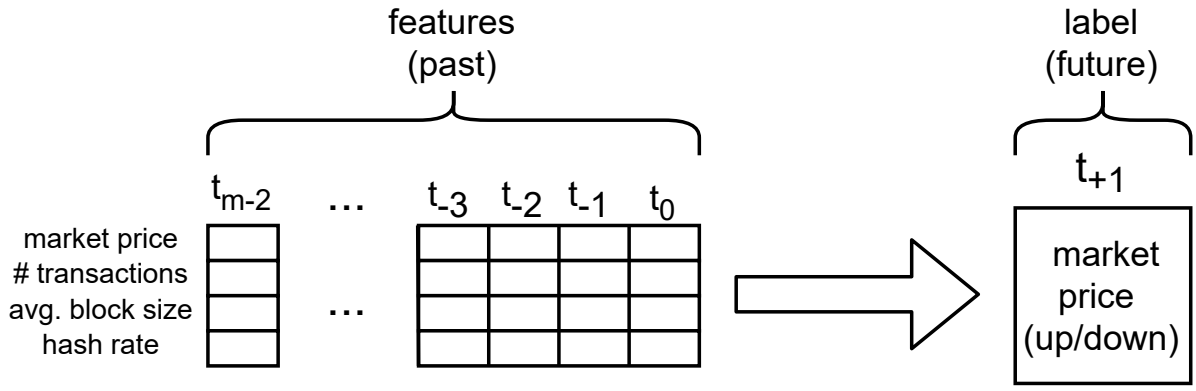


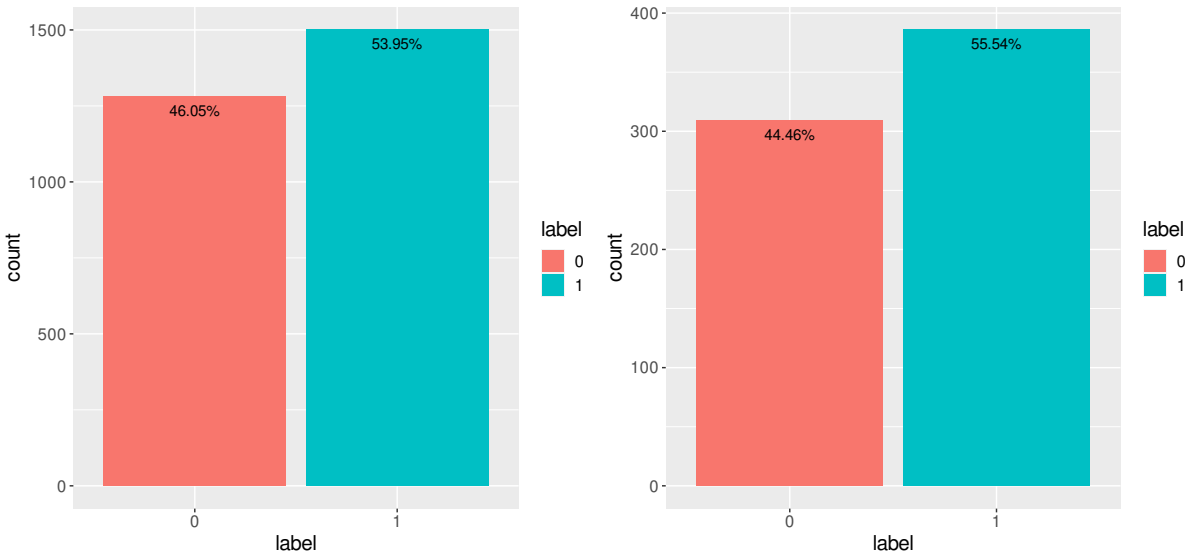
Figure 6: Example from the binary classification dataset on the Bitcoin market price direction. The features include the past Bitcoin market price and the three selected time series. The label is the Bitcoin market price direction one day into the future.

2.2 Transformation to Binary Classification Dataset

To transform the time series dataset developed in the last Subsection into a dataset suitable for binary classification, the so-called moving-window approach (also called sliding-window approach) is used. For this purpose, the first m days of the time series dataset for all four measures are taken. The last day of this window is taken apart as the future. The other days are taken as the features. The future day is labeled with a 1, if the Bitcoin price increases with respect to the last, and with a 0, if it decreases. In the following, the labels 1 and "up", and 0 and "down" are used synonymously. In the present work, $m = 100$ is chosen. All of the four features (that are parts of the overall time series dataset) are standardized using the min-max-transformation

$$\hat{y} = \frac{y - \min(y)}{\max(y) - \min(y)}.$$

The examples in the binary classification dataset have the form displayed in Fig. 6. There is the historical data for $m - 1$ days on the four selected time series including the Bitcoin market price development. The label is the Bitcoin market price direction on the next day. In total, 3477 examples are generated by using the moving window technique.
Finally, the binary classification dataset is split into a training and a test set. Before, all examples in the binary classification dataset are randomly permutated. The split is 80 % training examples and 20 % test examples. For more information on the split of the binary classification dataset into a training and a test part, see Subsection 4 of the next Section that contains the theory part of the present work. An overview of the split of the binary classification dataset into training and test examples is given in Fig. 7.



(a) training examples

(b) test examples

Figure 7: Split of the binary classification dataset into a training and a test part. Both sets are balanced because the contain both labels roughly the same times.

3 Theory

3.1 Majority Predictor

3.2 Logistic Regression

3.3 K-Nearest Neighbors Algorithm

3.4 Deep Neural Network

3.5 Training, Validation, and Test Set & Cross-Validation

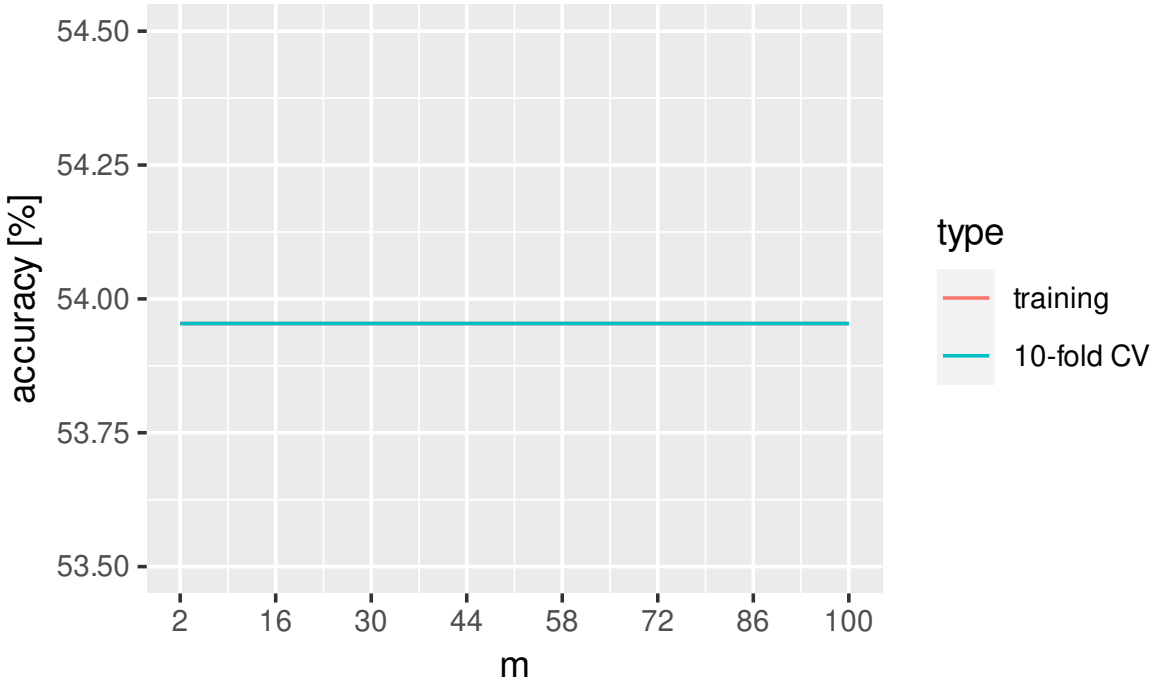


Figure 8: accuracy

predicted \ true	0	1
0	0	0
1	309	386

Table 1: confusion matrix of majority predictor on test set

4 Implementation & Software

- R & several packages - RStudio - Python & Anaconda for package managing - Tensorflow

5 Results

- explain procedure: 1. find most suitable m for all machine learning techniques via cross-validation and find best hyperparameters for best m (if applicable) via cross-validation 2. compute test accuracy, compute confusion matrix, ROC plot

5.1 Majority Predictor

- CV accuracy and test accuracy constant for all m (in general, different) - take majority predictor as a reference for the performance of the other prediction algorithms
- test accuracy: 55.54 %

5.2 Logistic Regression

- test accuracy: 56.70 % at $m = 4$ - training accuracy becomes much larger because model stores training examples , CV accuracy remains roughly constant (decreases slightly)
- model distinguishes, better than majority predictor as also Bitcoin down is predicted

6 K-Nearest Neighbors Algorithm

- test accuracy: 51.80 % at $m = 91$

Figure 9: accuracy

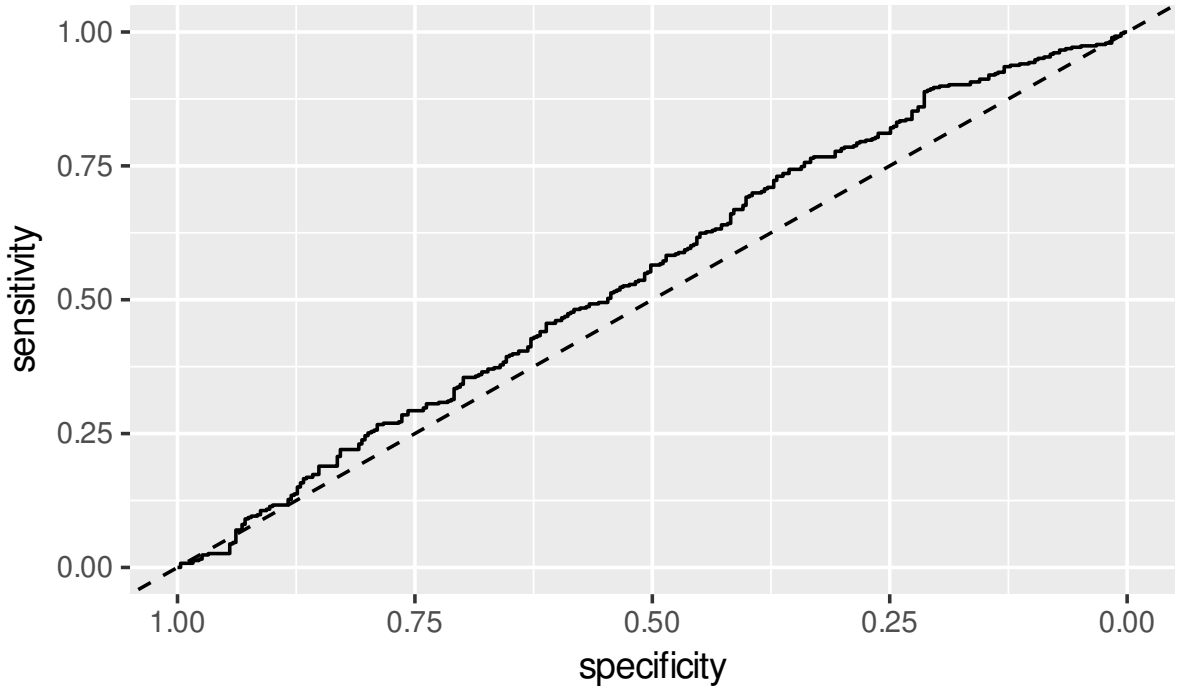


Figure 10: roc curve for logistic predictor

predicted \ true	0	1
0	77	73
1	232	313

Table 2: confusion matrix of logistic regression predictor on test set

coefficient	estimate	std.	z score	$P(>\|z\|)$ [%]
(intercept)	−0.15	0.22	−0.688	49.13
market price (t_{-2})	0.45	−0.60	0.764	44.51
market price (t_{-1})	0.71	0.82	0.874	38.20
market price (t_0)	−0.65	0.57	−1.121	26.22
# transactions (t_{-2})	−0.28	0.27	−1.040	29.81
# transactions (t_{-1})	0.20	0.32	0.624	53.28
# transactions (t_0)	0.11	0.27	0.394	69.32
avg. block size (t_{-2})	−0.10	0.28	−0.252	80.10
avg. block size (t_{-1})	−0.54	0.32	−1.682	9.25
avg. block size (t_0)	0.69	0.28	2.501	1.24
hash rate (t_{-2})	−0.10	0.26	−0.171	86.44
hash rate (t_{-1})	−0.10	0.28	−0.043	96.56
hash rate (t_0)	−0.10	0.26	−0.029	97.67

Table 3: confusion matrix of logistic regression predictor on test set

coefficient	$\sqrt{VIF}$
market price (t_{-2})	4.69
market price (t_{-1})	6.64
market price (t_0)	4.82
# transactions (t_{-2})	1.68
# transactions (t_{-1})	2.01
# transactions (t_0)	1.70
avg. block size (t_{-2})	1.78
avg. block size (t_{-1})	2.08
avg. block size (t_0)	1.79
hash rate (t_{-2})	1.45
hash rate (t_{-1})	1.55
hash rate (t_0)	1.46

Table 4: sqrt(VIF) for logistic regression

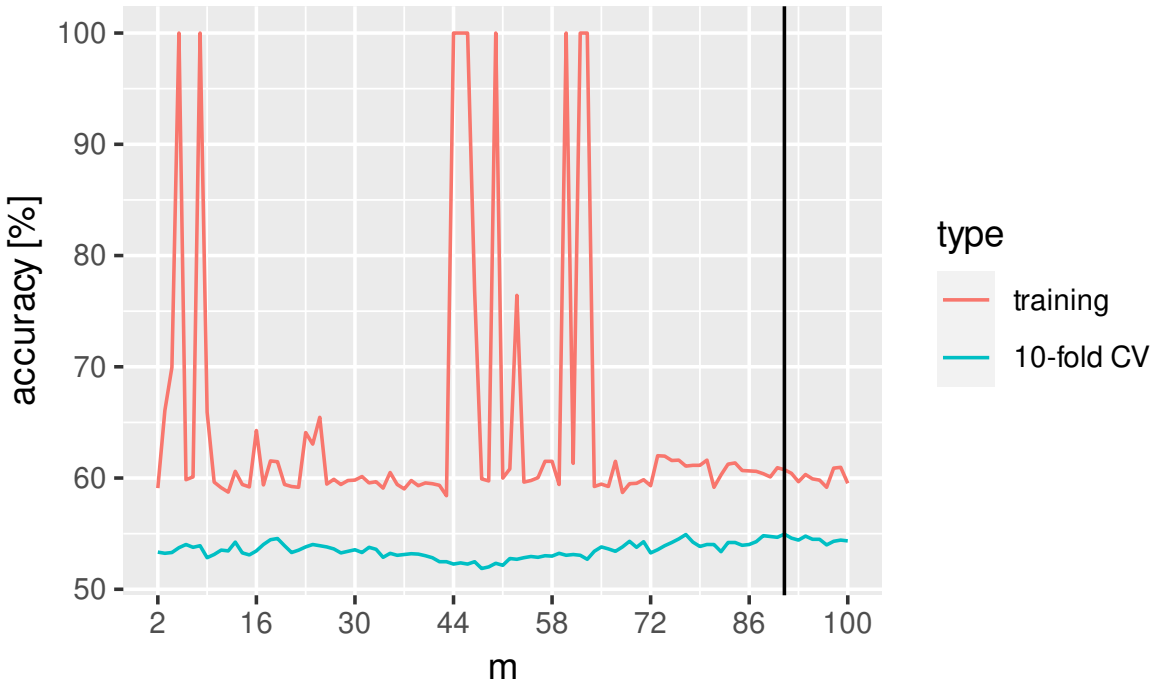


Figure 11: accuracy knn

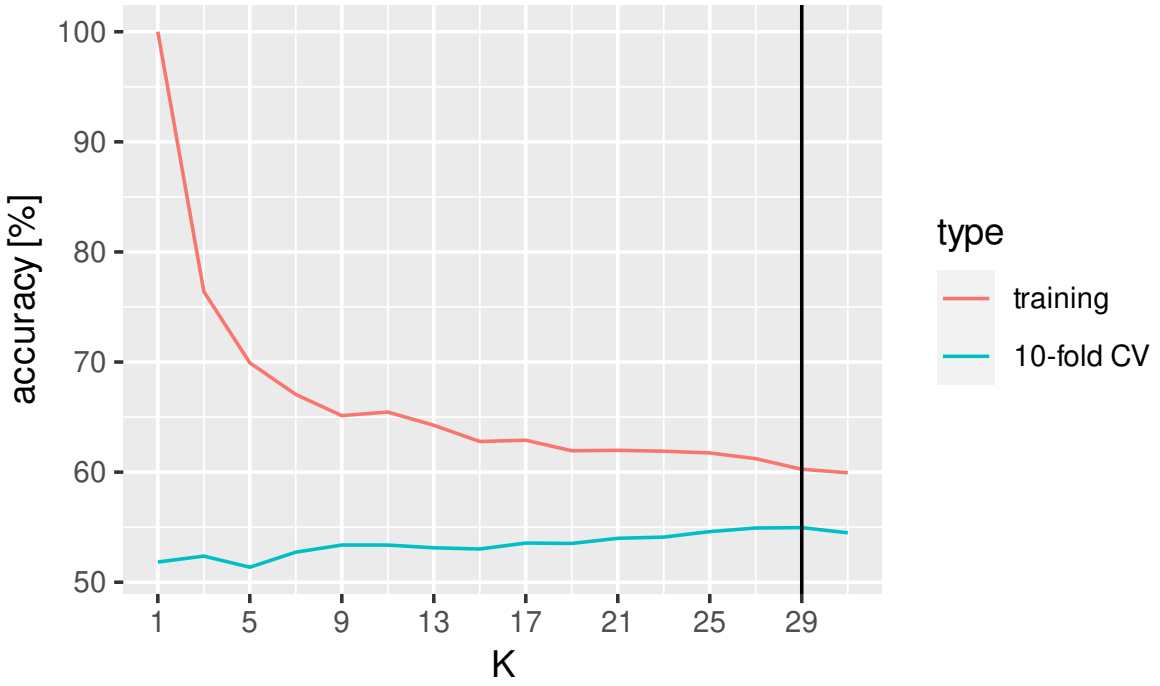


Figure 12: accuracy knn vs. K

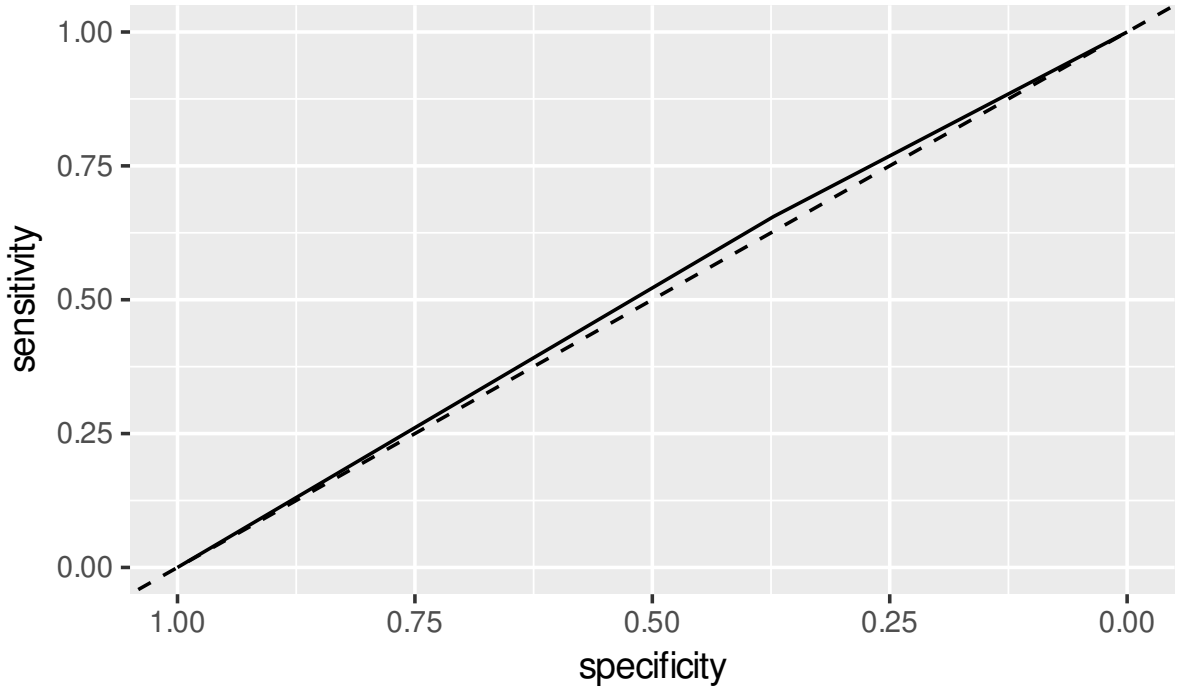


Figure 13: knn ROC

predicted \ true	0	1
0	115	133
1	194	253

Table 5: confusion matrix of knn on test set

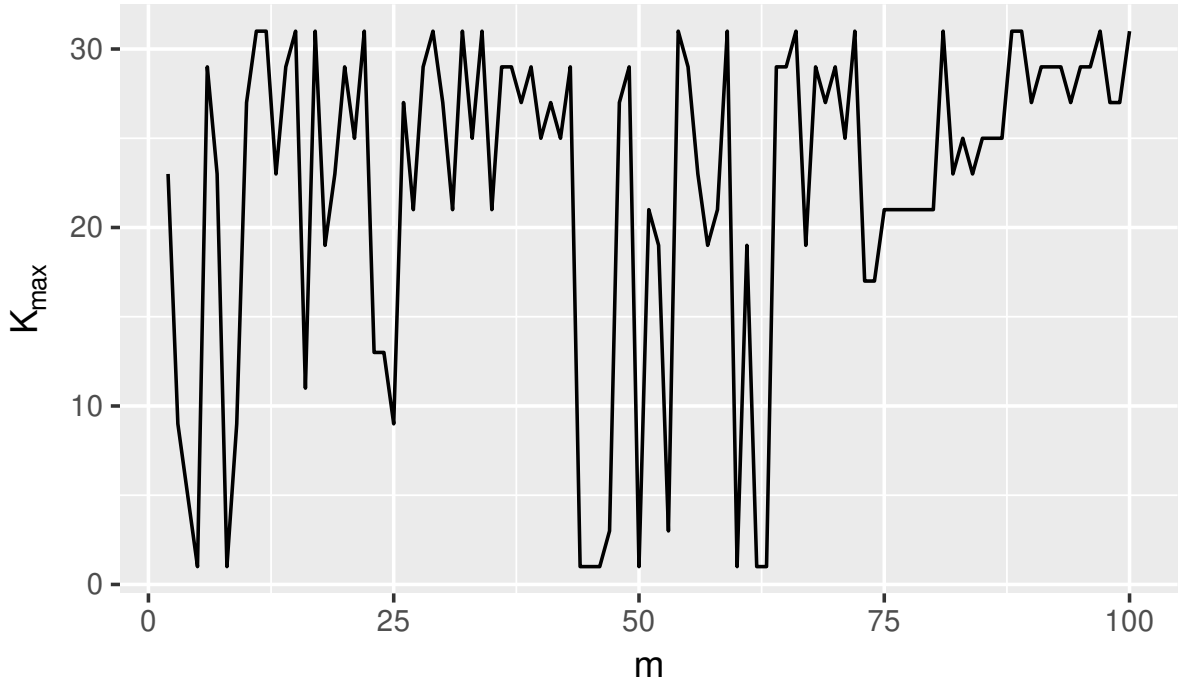


Figure 14: knn K max

7 Deep Neural Network

8 Discussion

9 Conclusion & Outlook

Outlook: - Google searches [5], Twitter [6] - consider individual transactions [7] - predict major crashes [8] [9] [10]

References

[1] Satoshi Nakamoto. "Bitcoin: A Peer-to-Peer Electronic Cash System." In: *Cryptography Mailing list at https://metzdowd.com* (Mar. 2009).

[2] Statista. *The 100 largest companies in the world by market capitalization in 2021*. 2021. URL: https://www.statista.com/statistics/263264/top-companies-in-the-world-by-market-capitalization/ (visited on 06/14/2021).

[3] CoinMarketCap. *Today's Cryptocurrency Prices by Market Cap*. 2021. URL: https://coinmarketcap.com/ (visited on 06/14/2021).

[4] Blockchain.com. *Blockchain Charts & Statistics API*. 2021. URL: https://www.blockchain.com/api/charts_api (visited on 06/16/2021).

[5] Martina Matta, Maria Ilaria Lunesu, and Michele Marchesi. "Bitcoin Spread Prediction Using Social And Web Search Media." In: June 2015.

[6] Germán Cheuque and Juan Reutter. "Bitcoin Price Prediction Through Opinion Mining." In: May 2019, pp. 755–762. ISBN: 978-1-4503-6675-5. DOI: 10.1145/3308560.3316454.

[7] A. Greaves and Benjamin Au. "Using the Bitcoin Transaction Graph to Predict the Price of Bitcoin." In: 2015.

[8] Elon Musk. *Tweet*. 2021. URL: https://twitter.com/elonmusk/status/1392602041025843203 (visited on 06/15/2021).

[9] CNBC. *Bitcoin price falls after China calls for crackdown on bitcoin mining and trading behavior*. 2021. URL: https://www.cnbc.com/2021/05/21/bitcoin-falls-after-china-calls-for-crackdown-on-bitcoin-mining-and-trading-behavior.html (visited on 06/15/2021).

[10] Bloomberg. *Binance Faces Probe by U.S. Money-Laundering and Tax Sleuths*. 2021. URL: https://www.bloomberg.com/news/articles/2021-05-13/binance-probed-by-u-s-as-money-laundering-tax-sleuths-bore-in (visited on 06/15/2021).