Forecasting the movements of Bitcoin prices: an application of Machine Learning Algorithms

Mainack Paul (MD2111)

May 13, 2023

1 Introduction and Objectives

Cryptocurrencies, such as Bitcoin, are one of the most controversial and complex technological innovations in today's financial system. Bitcoin is a decentralized currency, which means it is not controlled by any government or financial institution. Instead, it is based on a peer-to-peer network that allows users to send and receive payments directly without the need for intermediaries such as banks. Bitcoin, similar to other financial assets, exhibit chaotic fluctuations.Because of asymmetric information problems in financial markets and increasing economic-political uncertainties make the prices of bitcoin not easily predictable for investors. But accurately forecasting their prices may minimize potential losses-risks for users. The aim of this study is to forecast the movements of Bitcoin prices at high degree of accuracy.

2 Data

In case of continous data, for the output, we considered the changes of up and down movements of closing prices from previous days. We coded these as $+1$ and -1 for ups and downs, respectively. We used the same output discrete datasets also. For creating the discrete dataset, the continuous dataset was converted to -1 or $+1$ by applying the discretization process. $+1$ and -1 indicate upward and downward movements, respectively. Closing, high and low prices were used for computing technical indicators and output as reported in Table 1.

Table 1: Selected technical indicators

Indicators	Formula
Simple 14 days moving average (MA)	$\frac{(C_t+C_{t-1}+.....+C_{t-14})}{14}$
Simple 14 days weighted moving average (WMA)	$\frac{n*C_t+(n-1)*C_{t-1}+.....+C_{t-14}}{n+(n-1)+...+1}$
Momentum (Mom)	$C_t - C_{t-n}$
Stochastic K% (K%)	$\frac{C_t-LL_{t-n}}{HH_{t-n}-LL_{t-n}}*100$
Stochastic D% (D%)	$\sum_{i=0}^{n-1} K_{t-i}/n$
Relative strength index (RSI)	$100 - \frac{100}{1+(\sum_{i=0}^{n-1} Up_{t-i}/n)/(\sum_{i=0}^{n-1} Dw_{t-i}/n)}$
Moving average convergence/divergence (MACD)	$MACD(n)_{t-1} + \frac{2}{n+1} * (DIFF_t - MACD(n)_{t-1})$
Larry William's R% (LW)	$\frac{H_n-C_t}{H_n-L_n} * 100$
Accumulation/distribution oscillator (A/D)	$\frac{H_t-C_{t-1}}{H_t-L_t}$

C_t:closing price, L_t:low price,H_t:High price. $DIFF_t : EMA(12)_t - EMA(26)_t$. EMA is exponential moving average, $EMA(k)_t : EMA(k)_{t-1} + \alpha * (C_t - EMA(k)_{t-1})$, α is correction factor. LL_t is the lowest low, HH_t is the highest high for the last t days. $M_t = (H_t + L_t + C_t)/3$, $SM_t = \sum_{i=0}^{n} M_{t-i+1}/n$, $D_t = (\sum_{i=0}^{n} |M_{t-i+1} - SM_t|/n$, Up_t and Dw_t are upward and downward price change at time t respectively.

3 Machine Learning Models

Following the calculations of nine technical input parameters, four different Machine Learning (ML) algorithms are applied, namely, the Support Vector Machines (SVM), the Artificial Neural Network (ANN), the Naive Bayes (NB) and the Random Forest (RF) besides the logistic regression (LR) as a benchmark model. Continuous (existing) dataset, between 2008–2019(on a daily basis) was normalized for all ML models and divided into two parts as training (75%) and testing (25%). We used model validation to compare the performances and significances of the models with benchmark model. The validation dataset consists of Bitcoin series between June 2020–October 2020. This validation dataset was divided into 10 sub-datasets. For each sample, the movement estimations of each estimated model and accuracy statistics were calculated. The average accuracies of these 10 sub-datasets were bilaterally compared with the LR statistics with t test. The accuracy statistics for both continuous and discrete datasets were calculated. In order to test algorithms mentioned above and compare their performance, F statistics are calculated by using the true/false positive (TP-FP) and true/false negative (TN-FN), following the equations below:

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

$$F = \frac{2 * Precision * Recall}{Precision + Recall}$$

Machine learning algorithms do not require stationary tests differently from econometric models. In order to test the performances of selected algorithms, besides F statistics, mean absolute error (MAE), root mean square error (RMSE) and root absolute error (RAE) are also used.

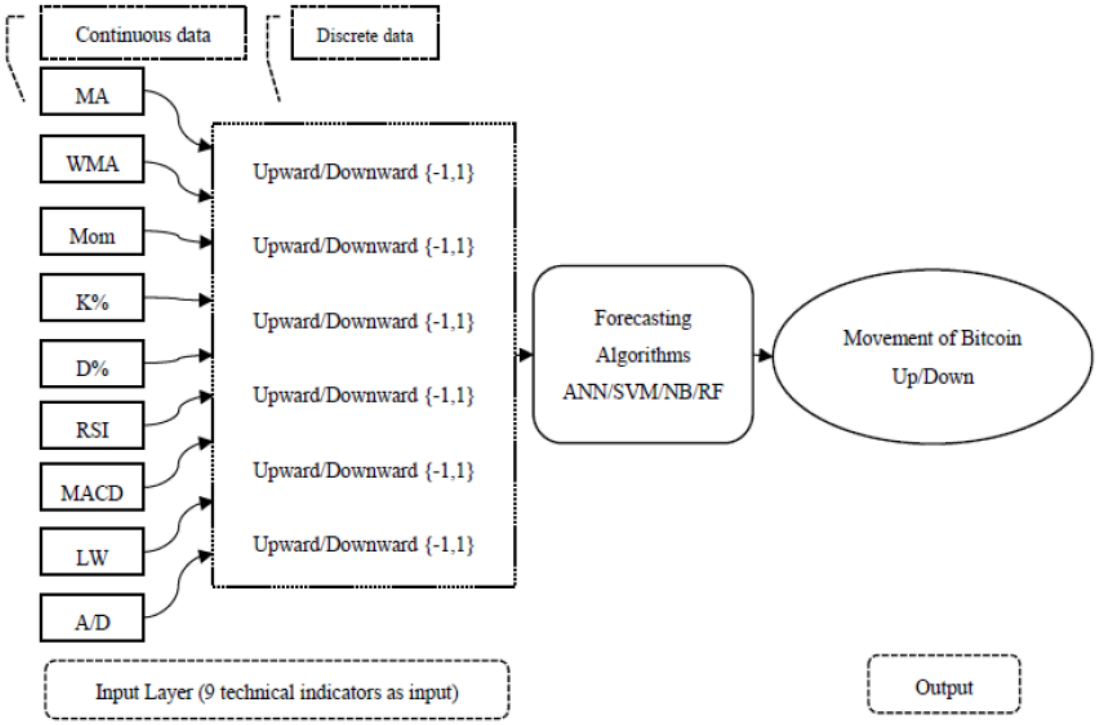

Figure 1. Forecasting mechanisms.

4 Findings

Empirical findings reveal that, while the RF has the highest forecasting performance, the NB has the lowest in continuous dataset. On the other hand, while the ANN has the highest performance, the NB has the lowest in discrete dataset. Furthermore, discrete dataset improves the overall forecasting performance in all models estimated. Each of the nine technical parameters used in this study can also be considered as an estimator. However, these parameters were used after considering their trend characteristics rather than their direct usage as estimators. This transformation done

in the study has increased forecasting performances.

5 Comments

In this study, algorithms has classified Bitcoin prices as up-down. However, instead of only two categories, it is suggested that multi-categories using different algorithms may be used for future forecasts. Besides the nine technical parameters used in this study, it is suggested to use some other macroeconomic parameters, such as exchange rate, interest rate, government policy implementations, for use in these models as new inputs, because all these variables may easily affect the financial markets involving cryptocurrencies.