Forecasting and trading cryptocurrencies with machine learning under changing market conditions

School of Computer Science
and Electronic Engineering
University of Essex
CF969-7-SU-CO : Big Data
for Computational Finance
name: Supasun Khumpraphan
registration number:2110366

ABSTRACT

In this paper, a study was conducted on cryptocurrency stock predictions. Three of them are Bitcoin, Ethereum, and Litecoin with three machine learning models: Random Forests, and Support Vector Machines. According to the authors of this paper, research shows that Ethereum is making a 44.65 percent annual profit, Litecoin was 34.86 percent, and Bitcoin 5.868 percent. From the results of this paper, it can be seen that we can use machine learning how to make profits in cryptocurrency stocks.

Author Keywords

Machine learning; Support Vector Machine; Random Forest; Forecasting; Trading; Bitcoin; Ethereum; Litecoin

CCS Concepts

•**Cryptocurrency** → *Bitcoin; Ethereum; Litecoin;* •**Model** → *Support Vector Machine; Random Forest Classifier;*

1 INTRODUCTION

Well-known investor Warren Buffett once said in stock investing: “The Great Enemy of Stock Investors is expenses and emotion [3].” That is because investing without skills and knowledge. by relying on news sources from friends or unreliable news sources Or even our own feelings is the greatest risk that makes us lose money. So this is the main problem and we can use AI that doesn’t have feelings involved. Solve this problem by using past and present stock price data to predict future stocks. The reason this research is so important is that the investment in cryptocurrencies is rapidly growing and attracting a wide audience. It can be clearly seen from the question most people searched on google in 2018 “what is bitcoin” [6]. It is also one of the largest unregulated markets in the world today [4]. That makes it very difficult to control the cost of stocks with only one person or one group, unlike stocks. A large proportion of shareholders can control the share price easily [8]. To predict the stock market, including cryptocurrency There is information that can be predicted. such as sales, net profit, economic growth, and unpredictable information such as politics. Terrorism, natural disasters, or even the emotions of greed and fear of the majority. Many of these unpredictable factors make predicting machine learning and deep learning very difficult and cannot be predicted with 100 percent accuracy but are still more effective than human predictors[11]. Most of the previous research has used machine learning to predict digital currency with bitcoin. However, little research has been done to predict multiple cryptocurrencies and in many models. That makes this research very interesting. The study will use Random Forests, and Support Vector Machines as a predictive model. In addition, it predicts 3 digital currencies: Bitcoin, Ethereum, and Litecoin. All three were chosen because they dominated the cryptocurrency market up to 75 percent of the value of all cryptocurrencies.

2 METHODOLOGY

2.1 Import Data and Data Pre-processing

In the first step, this research imports 1305 days of data and makes data pre-processing from two sources, CoinMarketCap and Coin Metrics. which have data on Bitcoin, Ethereum, and Litecoin from August 7, 2015, to March 3, 2019. The data is cleaned and used by The authors of the research used data from CoinMarketCap to use variables closing prices, highest, lowest, daily trading volume, and capitalization. All of these are in dollars and create an additional variable by doing Returns. A variable that uses the closing price to calculate Logarithmic returns of bitcoin, Ethereum, and Litecoin. Volume variable is the daily exchange price. Capitalization variable is the cryptocurrency market capitalization. Relative price change variable is daily price fluctuation. That takes the high and low prices of that day (calculate in Equation1)[9]. Parkinson’s volatility variable. That takes the high price and the low price of that day (to be calculated in Equation 2)[7]. Data from Coin Metrics was then used for the 12 blockchain data variables, including the On-chain volume variable, which is based on the total value on the blockchain. The adjusted on-chain volume variable is the total value of the blockchain used in economic transactions. The median value variable is the median of the transaction. The number of transactions variable is the number of transactions on the blockchain. The new coins variable is calculated from the coin which created daily new coins of each currency. The total fees variable is the fee for all coins in every transaction per day in each currency. The median fees variable is the median in each currency per day. The active addresses variable is The unique number of addresses used in the network. The average difficulty variable is the average

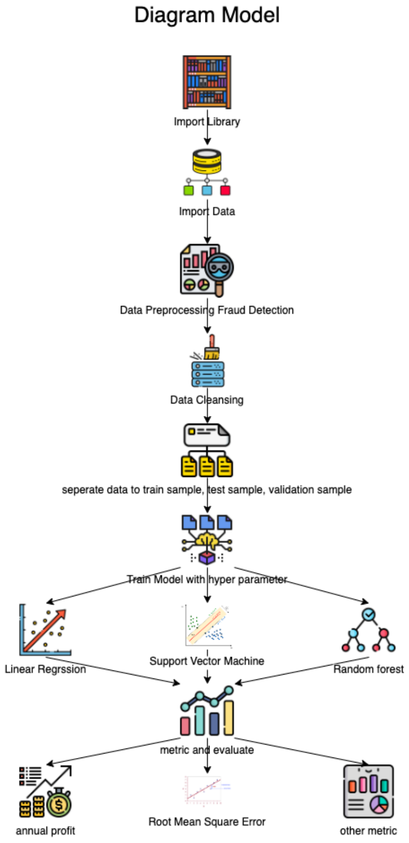


Figure 1. Diagram in this research

time to find a new block. The number of blocks variable is The number of new blocks in the blockchain system. The block size variable is The average size of the blockchain in bytes. The number of payments variable is the number of buyer's transactions. Then create the Daily dummies variable to convert from day to number ready to use for the prediction model.

$$RR_t = \frac{2(H_t - L_t)}{H_t + L_t} \tag{1}$$

where $H_t =$ is highest price in one day. $L_t =$ is lowest price in one day. $RR_t =$ is relative price range.

$$\sigma_t = \sqrt{\frac{(\ln(H_t/L_t))^2)}{4\ln(2)}} \tag{2}$$

where $H_t =$ is highest price in one day. $L_t =$ is lowest price in one day. $\sigma_t =$ is range volatility estimator of Parkinson.

2.2 Separate data in train, test, and validation sample

After that we have to split the data to use in the prediction because the model can work accurately with data that has never been seen before is called Generalization. it is an important concept in developing machine learning systems because if we have a system that works precisely only with the data it has seen before It's like being a student who remembers the exam, enters the exam, gets it right, only the questions that are exactly the same. Can't deal with even slightly different problems when it comes to actual use When the actual data is found, the model will have unacceptably low accuracy performance, known as overfitting [10]. Thus, in this study, Bitcoin, Ethereum, Litecoin were divided into three parts, with the first 648 days being a training sample which accounted for 50 percent of the total data. Used to train the model. The second part is the validation sample we will use data From 649 to 972, a total of 324 days, which is 25 percent approximately Used to test the performance of the model after training to see how well the model performs. after each tuning, which model performs better? finally, we bring day 973 to day 1297. Approximately 325 days or 25 percent is the test sample used to test after the best model has been obtained. on how well the model will perform with never-before-seen data.

2.3 Statistic

In this paper, many data statistics were examined and the results were obtained for 3 periods of time. The first period is the period of training where the price tends to continue rising. Although there is some volatility, it is not much in Bitcoin and Litecoin. While Ethereum is more volatile in the early days than in the latter. The second period is the period of validation, it can be seen that sometimes the prices of the three currencies will fall drastically but after that, the prices tend to rise even more and gain profit. The third period is the duration of the test. There is a tendency for most of the currency's prices to drop

2.4 Model

In this research, we will use the information we have to predict the model, where we will predict the future returns of digital currencies using the regression model and calculate the trend of whether the price will go up or down with the classification model. It will be used in R language and the model that will be used in this research is Support Vector Machine and Random Forest which can do both Regression and Classification model. In Random Forest, multiple trees are trained together, where each tree gets its features and data is a random subset of all features and data. When making a prediction, let each tree make a prediction of each and choose a final prediction result from the value. The most accurate prediction technique is called bagging or boostrapping(Fig. 2) [1]. Additionally, this research used Support Vector Machines, or SVMs, which are both flexible and functional. Especially when there are many of feature information but the number of samples is small. (less than a hundred thousand) It creates a straight line that divides the data (Hyperplane) and finds the best one. We will choose the line with the margin If the margin is narrow, giving a high chance of overfitting. Therefore, we will Choose a lot of Margins, causing less Overfit or called Soft Margin. Adjusting the parameters by using the value C will change the size of the dividing line. The more C will make the area narrower, but if too much will lead to less chance of overfitting C. will make the area wider but less accurate (Fig. 3)[5], but the SVM algorithm we have presented The limitation is that only linear decision boundaries can be created. This may not work well if the data relationships are so complex that they cannot be

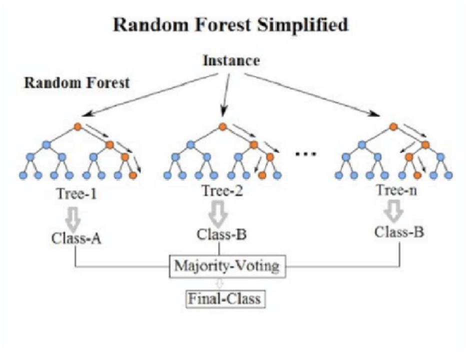


Figure 2. Architecture of Random Forest[1].

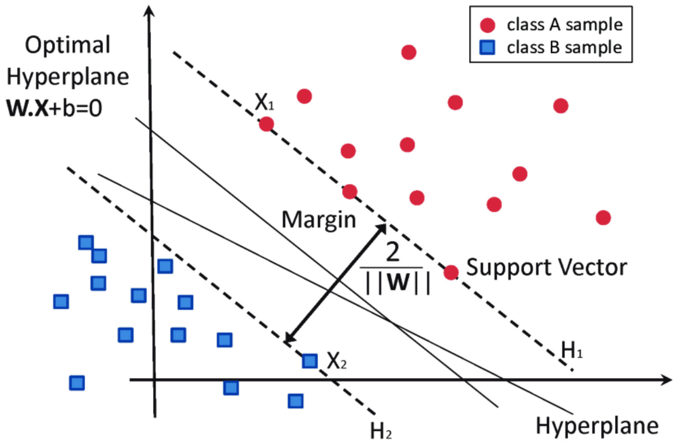


Figure 3. Architecture of Support Vector Machine[5].

divided by straight lines. This solution is called a non-linear kernel to address the problem. by the way, Create a dimension from the original 2D to 3D and draw a line across the middle to allow the data to be divided into groups (Fig. 4)[2].

2.5 Hyperparameter

To find the best model, the authors of this paper altered several variables and defined hyperparameters to determine the best-performing parameters of each model and cryptocurrencies and to predict the data.

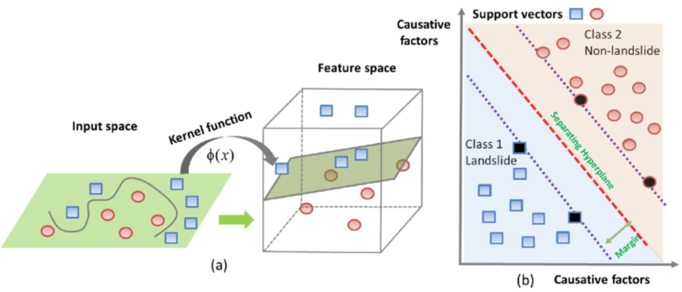


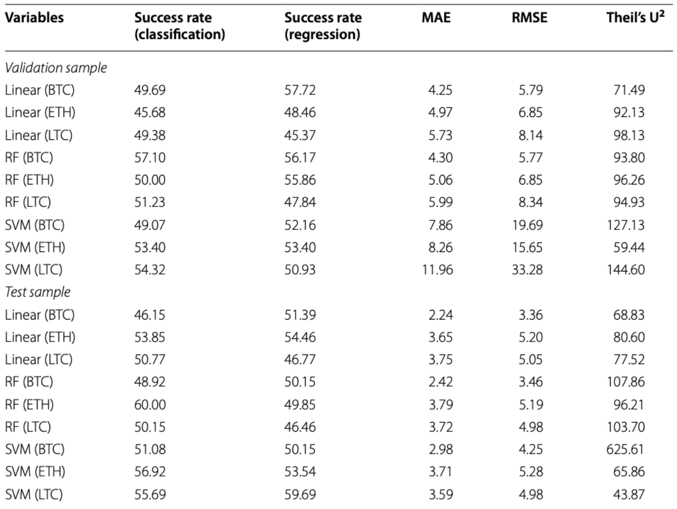
Figure 4. Architecture of Kernel Support Vector Machine[2].

2.6 Metric and evaluate

This research uses the metric Multiple models to test the accuracy and discrepancy of the model using a success rate predicted 1 day in advance which will return the value later in the day whether the prediction is correct or wrong and measure the number of times the model is correct. This can be calculated for both regression and classification models. It also uses, root means square error (RMSE), mean absolute error (MAE), Teil's U2, and mean square error (MSE), but does not use accuracy metrics because the goal of This research is not minimizing model errors, but maximizing average return. Therefore, accuracy is not necessary. In addition, perform statistical and profitable analysis with ensembles. The ensemble is the ratio of the correct number of days to predict the next day's trend and the total number of days calculated. It is expressed as mean and standard deviation. The annualized return is calculated based on the cumulative daily return and the annualized Sharpe ratio is calculated by multiplying the daily yield and standard deviation by square root 365(19.1049731745) There is also Several metrics used including the BH strategy, which measures daily return probability. tail risk measures the likelihood of unforeseen occurrences. former measures that measure expected damage.

3 RESULT

To measure the accuracy of this paper, the success rate is used. in validation sample with the highest success rate in the classification model is a random forest that predicts Bitcoin with a success rate is 57.10 percent while the success rate with the lowest is Linear predicted Ethereum coin was 45.68 percent. While the regression model's highest success rate was linear predicted Bitcoin at 57.72 percent, while the success rate with the lowest is Linear which predicts Litcoin at 45.37 percent, and looking at MAE. the lowest value is model Linear which predicts Bitcoin at 4.25 and the highest MAE value is the predicted model support vector machine in Litecoin is at 11.96 and looking at the RMSE, the lowest value is the model random forest predicting bitcoin at 5.77 and the highest RMSE value is the model support vector machine that predicts Litecoin at 33.28. The smallest value of Theil's U2 is the model support vector machine that predicts Ethereum at 59.44 and Theil's U2 highest is the predicted model support vector machine in Litecoin at 144.60. All 3 currencies have a success rate classification average was 51.10 percent and the average regression was 51.99 percent. In the test sample, the model with the highest success rate in the classification model was a random forest predicting the Ethereum coin with a success rate of 60 percent, while The lowest success rate was Linear, predicting bitcoin at 46.15 percent. While the highest success rate in the regression model is the support vector machine predicted in Litecoin at 59.69 percent, while the lowest success rate is the random forest that predicts Lite coin at 46.46 percent and when looking at the MAE, the lowest value is the Linear model predicting Bitcoin at 2.24 and the highest MAE is the model random forest that predicts Ethereum at 3.79. In terms of RMSE, the lowest value is the model Linear predicting bitcoin at 3.36 and the highest RMSE is the model support vector machine predicting Ethereum at 5.28 and if you look at Theil's The smallest value U2 is the model support



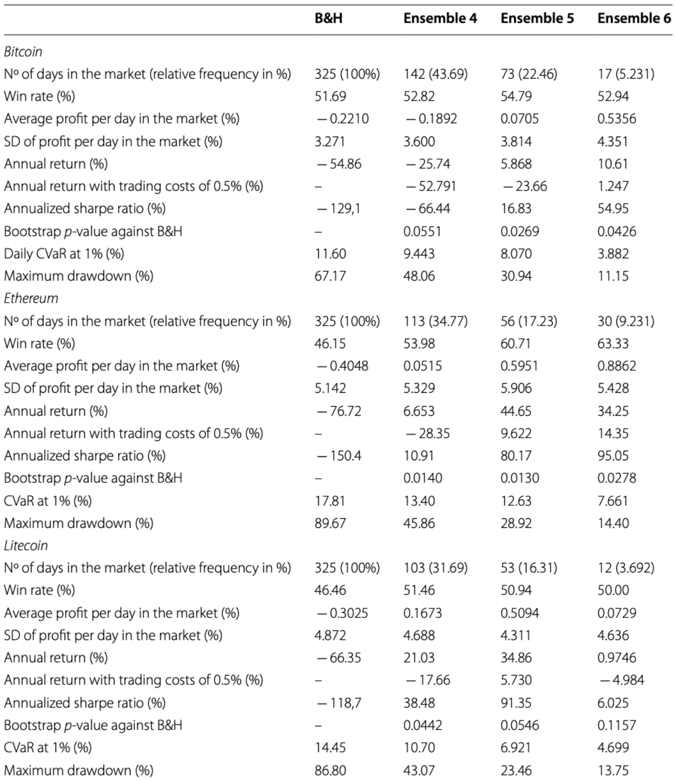
Variables	Success rate (classification)	Success rate (regression)	MAE	RMSE	Theil's U^2
Validation sample					
Linear (BTC)	49.69	57.72	4.25	5.79	71.49
Linear (ETH)	45.68	48.46	4.97	6.85	92.13
Linear (LTC)	49.38	45.37	5.73	8.14	98.13
RF (BTC)	57.10	56.17	4.30	5.77	93.80
RF (ETH)	50.00	55.86	5.06	6.85	96.26
RF (LTC)	51.23	47.84	5.99	8.34	94.93
SVM (BTC)	49.07	52.16	7.86	19.69	127.13
SVM (ETH)	53.40	53.40	8.26	15.65	59.44
SVM (LTC)	54.32	50.93	11.96	33.28	144.60
Test sample					
Linear (BTC)	46.15	51.39	2.24	3.36	68.83
Linear (ETH)	53.85	54.46	3.65	5.20	80.60
Linear (LTC)	50.77	46.77	3.75	5.05	77.52
RF (BTC)	48.92	50.15	2.42	3.46	107.86
RF (ETH)	60.00	49.85	3.79	5.19	96.21
RF (LTC)	50.15	46.46	3.72	4.98	103.70
SVM (BTC)	51.08	50.15	2.98	4.25	625.61
SVM (ETH)	56.92	53.54	3.71	5.28	65.86
SVM (LTC)	55.69	59.69	3.59	4.98	43.87

Figure 5. performance model in cryptocurrency[12].

vector machine predicting Litecoin at 43.87 and The highest Theil's U2 is the predicted model support vector machine in Bitcoin is at 625.61. All 3 currencies have a success rate classification average was 52.61 percent and the average regression was 51.38 percent. From the above-mentioned results, The model's accuracy results were unimpressive, with success rate classification below 50 percent in both test and validation samples in 6 cases, while in regression in 6 cases. As we can see from (Fig. 5). However, the strategy based on the Annual return in Ensemble 5 of this research is very profitable for Ethereum. The earned annual profit of 44.65percent. Meanwhile, Litecoin posted an annual profit of 34.86percent and bitcoin 5.868 percent. As we can see from (Fig. 6).

4 CONCLUSION

This study examined three cryptocurrencies: Bitcoin, Ethereum, and Litecoin. which, although the model's accuracy is not very accurate, can be hugely profitable. The highlight of this research that makes it possible to make huge profits is because of good data management. The data is converted clean. and have complete information What sets it apart from other research is the fact that 3 cryptocurrencies are compared and used. and separate train samples, test samples, and validation sample data for use in the model. Three models were used: linear regression, support vector machine, and random forest so that the capabilities of each model could be compared. There is also tuning hyperparameter to get parameter the best in each model improves the predictive efficiency of the model. In addition, there are many metrics and evaluations that allow for a wide range of results such as accuracy, tolerances, and annual profit. and clearly measure the performance and evaluation of the model. However, although this research is of great quality, it can add some things that make this research and model even more effective: Firstly, this research researches cryptocurrencies with very high value and relatively low volatility. But it doesn't compare to coins with high volatility and lower value such as Dogecoin. we are not sure. it can use this research in real situations. Secondly, there are too few models. we should add the model decision tree, XGBoost, and multi-layer perceptron may produce more results and profits than a model support vector machine or random forest and although this research focuses on the use of machine learning. However, this research doesn't have deep learning used, although recurrent neural networks or Long Short-Term Memory had excellent regression prediction performance, and convolution neural networks had the best predictive performance of classification. in addition, when it comes to predicting stocks, the use of Q learning in reinforcement learning is the best. However, it was not present in this study. which the model mentioned above would be more profitable and had a more efficient model. Thirdly, although hyperparameter is used But it is considered to be used very little. I think that if the tuning parameters are more diverse, it will give a more efficient model. for example in a random forest with n estimators only. if we use criterion and max features parameters may make the model more efficient. In addition, despite some statistical investigations, Including some variable adjustments. However, this research doesn't use EDA, Feature engineering, Feature Importance, which makes we can't know which variable is important? Some variables may cause the model's performance to decrease as well.



	B&H	Ensemble 4	Ensemble 5	Ensemble 6
Bitcoin				
Nº of days in the market (relative frequency in %)	325 (100%)	142 (43.69)	73 (22.46)	17 (5.231)
Win rate (%)	51.69	52.82	54.79	52.94
Average profit per day in the market (%)	− 0.2210	− 0.1892	0.0705	0.5356
SD of profit per day in the market (%)	3.271	3.600	3.814	4.351
Annual return (%)	− 54.86	− 25.74	5.868	10.61
Annual return with trading costs of 0.5% (%)	–	− 52.791	− 23.66	1.247
Annualized sharpe ratio (%)	− 129,1	− 66.44	16.83	54.95
Bootstrap *p*-value against B&H	–	0.0551	0.0269	0.0426
Daily CVaR at 1% (%)	11.60	9.443	8.070	3.882
Maximum drawdown (%)	67.17	48.06	30.94	11.15
Ethereum				
Nº of days in the market (relative frequency in %)	325 (100%)	113 (34.77)	56 (17.23)	30 (9.231)
Win rate (%)	46.15	53.98	60.71	63.33
Average profit per day in the market (%)	− 0.4048	0.0515	0.5951	0.8862
SD of profit per day in the market (%)	5.142	5.329	5.906	5.428
Annual return (%)	− 76.72	6.653	44.65	34.25
Annual return with trading costs of 0.5% (%)	–	− 28.35	9.622	14.35
Annualized sharpe ratio (%)	− 150.4	10.91	80.17	95.05
Bootstrap *p*-value against B&H	–	0.0140	0.0130	0.0278
CVaR at 1% (%)	17.81	13.40	12.63	7.661
Maximum drawdown (%)	89.67	45.86	28.92	14.40
Litecoin				
Nº of days in the market (relative frequency in %)	325 (100%)	103 (31.69)	53 (16.31)	12 (3.692)
Win rate (%)	46.46	51.46	50.94	50.00
Average profit per day in the market (%)	− 0.3025	0.1673	0.5094	0.0729
SD of profit per day in the market (%)	4.872	4.688	4.311	4.636
Annual return (%)	− 66.35	21.03	34.86	0.9746
Annual return with trading costs of 0.5% (%)	–	− 17.66	5.730	− 4.984
Annualized sharpe ratio (%)	− 118,7	38.48	91.35	6.025
Bootstrap *p*-value against B&H	–	0.0442	0.0546	0.1157
CVaR at 1% (%)	14.45	10.70	6.921	4.699
Maximum drawdown (%)	86.80	43.07	23.46	13.75

Figure 6. multiple metric in trading strategies[12].

REFERENCES

[1] Mourad Azhari, Altaf Alaoui, Zakia Acharoui, Badia Ettaki, and Jamal Zerouaoui. 2019. Adaptation of the random forest method: solving the problem of pulsar search. *SCA '19: Proceedings of the 4th International Conference on Smart City Applications* (10 2019), 1–6. DOI:`http://dx.doi.org/10.1145/3368756.3369004`

[2] Jie Dou, Ali P. Yunus, Dieu Bui, Mehebub Sahana, Chiwen Chen, Zhongfan Zhu, Weidong Wang, and Binh Pham. 2019. Evaluating GIS-Based Multiple Statistical Models and Data Mining for Earthquake and Rainfall-Induced Landslide Susceptibility Using the LiDAR DEM. *Remote Sensing* 11 (03 2019), 638. DOI: http://dx.doi.org/10.3390/rs11060638

[3] Gregg S Fisher. 2014. Advising the behavioral investor: lessons from the real world. *Investor Behavior–The Psychology of Financial Planning and Investing, USA* (2014).

[4] Sean Foley, Jonathan R Karlsen, and Tālis J Putniņš. 2019. Sex, drugs, and bitcoin: How much illegal activity is financed through cryptocurrencies? *The Review of Financial Studies* 32, 5 (2019), 1798–1853.

[5] Esperanza García-Gonzalo, Zulima Fernández-Muñiz, Paulino Jose Garcia Nieto, Antonio Sánchez, and Marta Menéndez. 2016. Hard-Rock Stability Analysis for Span Design in Entry-Type Excavations with Learning Classifiers. *Materials* 9 (06 2016), 531. DOI: http://dx.doi.org/10.3390/ma9070531

[6] Ying-Ying Hsieh, Jean-Philippe Vergne, Philip Anderson, Karim Lakhani, and Markus Reitzig. 2018. Bitcoin and the rise of decentralized autonomous organizations. *Journal of Organization Design* 7, 1 (2018), 1–16.

[7] Zahra Lashgari and Mousa Ahmadi. 2014. The impact of dividend policy on stock price volatility in the Tehran stock exchange. *Kuwait Chapter of the Arabian Journal of Business and Management Review* 3, 10 (2014), 273.

[8] Ernst Maug. 1998. Large shareholders as monitors: Is there a trade-off between liquidity and control? *The journal of finance* 53, 1 (1998), 65–98.

[9] Richard W Parks. 1978. Inflation and relative price variability. *Journal of Political Economy* 86, 1 (1978), 79–95.

[10] Huy Nguyen Anh Pham and Evangelos Triantaphyllou. 2008. The impact of overfitting and overgeneralization on the classification accuracy in data mining. In *Soft computing for knowledge discovery and data mining*. Springer, 391–431.

[11] Ashwini Saini and Anoop Sharma. 2019. Predicting the unpredictable: an application of machine learning algorithms in Indian stock market. *Annals of Data Science* (2019), 1–9.

[12] Helder Sebastião and Pedro Godinho. 2021. Forecasting and trading cryptocurrencies with machine learning under changing market conditions. *Financial Innovation* 7, 1 (2021), 1–30.