Forecasting Bitcoin Returns: A Data Mining Approach

Review of Integrative Business & Economics Research

Jen-Peng Huang
Department of Information Management, Southern Taiwan University of Science and Technology, Taiwan

Genesis Sembiring Depari*
Department of Business and Management, Southern Taiwan University of Science and Technology, Taiwan

ABSTRACT

The purpose of this study is to obtain a robust algorithm for predicting Bitcoin returns. To reduce dimensionality, we ran a weighting strategy using the information gain ratio technique. Methodologies including support vector machine, deep learning, and random forest algorithms were examined and compared in terms of processing time, receiver operator characteristic curve, and accuracy. A selected algorithm was used to weight the relative importance of input variables for predicting the Bitcoin return trend. The insights can help traders make decisions regarding selling or buying Bitcoin. This study contributes practical knowledge and insight to the literature. First, we introduce a simple methodological framework for predicting Bitcoin returns with 60%–70% accuracy, which assists decision making by weighting the relative importance of input variables. Second, we present managerial implications for how and when a trader should sell or buy Bitcoin according to the relative importance of input variables from a robust algorithm.

Keywords: Bitcoin, deep learning, support vector machine, random forest, data mining.

Received 6 August 2019 | Revised 19 November 2019 | Accepted 15 January 2020.

1. INTRODUCTION

As cryptocurrencies are becoming more popular and are considered a profitable and speculative commodity worldwide, many investors are attempting to profit by buying and selling them. According to Ciaian *et al.* (2016), Bitcoin is one of the most prominent cryptocurrencies in terms of extraordinary price development and volatility. However, predicting the trend of Bitcoin price is not an easy task because Bitcoin price is affected by many related factors, which are in turn affected by uncertain economic conditions. Therefore, researchers have tried to examine the Bitcoin price trend using various methods. For instance, Aalborg *et al*. (2018) studied what variables could explain the return, volatility, and trading volume of Bitcoin. However, the variables used by them covered only return, volatility, trading volume, transaction volume, change in the number of Bitcoin addresses, VIX index, and Google search. Since many other factors are potentially affecting Bitcoin returns, analyzing additional input variables might contribute to the literature. McNally *et al.* (2018), for instance, employed deep learning and ARIMA models to predict the trend of Bitcoin returns. In the present study, closing and opening

Copyright © 2021 GMP Press and Printing
ISSN: 2304-1013 (Online); 2304-1269 (CDROM); 2414-6722 (Print)

prices, daily high and low, mining difficulty, and hash rate are considered as independent variables. As such, using more input variables can potentially generate more complete information for predicting the Bitcoin price trend.

Bitcoin traders have realized that high fluctuations and uncertainties characterize Bitcoin trading, and thus the conditions of internal and external factors are essential. Moreover, accurate prediction of Bitcoin returns requires robust methodologies and input predictors. According to Turban *et al*. (2010), data mining is a useful research method for analyzing various datasets. Larose and Larose (2014) argued that machine learning, data collection, statistical method, data storage, and data visualization are parts of the data mining method. In addtion, machine learning algorithms have been widely applied particularly for prediction. Such algorithms were used by Patel *et al*. (2015) to predict stock price, by Gabralla *et al.* (2013) to predict oil price, by Sivalingam *et al*. (2016) to predict gold price, and by Park and Bae (2015) to predict house price. Therefore, using machine learning in predicting Bitcoin returns is promising.

Support vector machines (SVMs) are well-known machine learning algorithms. Georgoula *et al*. (2015) used an SVM to identify the determinants of Bitcoin price and found that Twitter sentiments, hash rates, and Wikipedia searches are positively related to Bitcoin price. Moreover, McNally *et al.* (2018) found that a nonlinear deep learning model outperformed an ARIMA model in forecasting the time series of Bitcoin price. Moreover, Pichl and Kaizoji (2017) found that an artificial neural network can predict the actual log distribution to analyze Bitcoin volatility. According to Cao and Tay (2003), a neural network is noise-tolerant and can learn with complex, unstructured, flexible, and incomplete data. As in Kumar and Thenmozhi's (2006) research, SVM and random forest are promising algorithms for predicting stock price movements and may be applicable to predicting the price trend of Bitcoin. Taken together, SVM, deep learning, and random forest are promising machine learning algorithms for cryptocurrency price prediction.

Poyser (2017) argued that internal and external factors affect the price of a cryptocurrency. Therefore, in this study, we explored 27 input variables consisting of internal and external factors related to Bitcoin price. We employed and compared the performance across three machine learning algorithms, namely deep learning, SVM, and random forest, in predicting the price trend of Bitcoin on a daily basis data. The comparison was made in terms of accuracy, receiver operating characteristic (ROC) curve, and processing time. The selected algorithm was then used to weight the relative importance of input variables that will help traders predict Bitcoin price and make investment decisions.

2. RELATED WORK

This section is divided into two parts: The first focuses on the econometric literature on Bitcoin, and the second on the data mining literature on Bitcoin.

2.1 Econometric literature on Bitcoin

Copyright © 2021 GMP Press and Printing
ISSN: 2304-1013 (Online); 2304-1269 (CDROM); 2414-6722 (Print)

Panagiotidis *et al*. (2018) used lasso regression to identify the determinants of Bitcoin returns. Search intensity, gold returns, and regulation uncertainty were disclosed as the most relevant factors for Bitcoin returns. Demir *et al*. (2018) revealed that economic policy uncertainty is negatively associated with Bitcoin returns and concluded that Bitcoin could be used as a hedging tool against uncertainties. Aalborg *et al*. (2018) studied which variables could explain the volatility, trading volume, and returns to Bitcoin. Their results indicated that Google trends could predict trading volume, and trading volume could in turn improve the model of volatility. By contrast, Baur *et al*. (2018) revealed that Bitcoin is not correlated with traditional commodities, stocks, and bonds.

2.2 Datamining literature on Bitcoin

A smaller number of past studies have predicted Bitcoin returns with external and internal factors as input variables, and a few of them have used a data mining method as a measurement tool. In this study, we compared three machine learning algorithms, namely deep learning, SVM, and random forest, along with dynamic parameters in terms of accuracy, ROC curve, and processing time. This study will help traders make decisions based on weighted internal and external factors using the selected algorithm. Aalborg *et al*. (2018) and McNally *et al*. (2018) used data mining techniques with a small number of related factors. Nakano *et al*. (2018), Karasu *et al*. (2018), and Georgoula *et al*. (2015) each used only a single machine learning algorithm and therefore lacked model comparison. Therefore, in this study, we aim to close this gap in the literature by comparing the three aforementioned algorithms using adaptive parameters for each in them.

3. DATA AND METHODOLOGY

3.1 Data

According to Poyser (2017), there are two major factors affecting a cryptocurrency's price: internal and external factors. Internal-related factors include supply/demand and the crypto market of cryptocurrency. Macrofinancial and political factors are considered as external factors. This study explored 26 input variables covering both internal and external factors of cryptocurrency price determinant. This dataset consists of daily data from February 16, 2017 to February 14, 2019. The characteristics of this dataset are provided in Table 1.

Table 1. Input and output variables in predicting bitcoin price return

Features	Region	Source	Role
Day	Worldwide	blockchain	Input variables
Daily economic policy uncertainty index (US)	US	EPU indices	Input variables
Daily Economic Policy Uncertainty (EPU) Index (UK)	UK	EPU indices	Input variables
Type of day (weekday/weekend)	Worldwide	blockchain	Input variables

Copyright © 2021 GMP Press and Printing
ISSN: 2304-1013 (Online); 2304-1269 (CDROM); 2414-6722 (Print)

Return direction (1/-1)	Worldwide	Calculated	Label
Hash Rate	Worldwide	Blockchain	Input variables
Difficulty	Worldwide	blockchain	Input variables
Confirmed Transactions Per Day	Worldwide	blockchain	Input variables
Estimated Transaction Value	Worldwide	blockchain	Input variables
Total Transaction Fees	Worldwide	Quandl	Input variables
my wallet number of transaction per day	Worldwide	Quandl	Input variables
my wallet transaction volume	Worldwide	Quandl	Input variables
average block size	Worldwide	Quandl	Input variables
API Blockchain size	Worldwide	Quandl	Input variables
Cost per transaction	Worldwide	Quandl	Input variables
Cost % of transaction volume	Worldwide	Quandl	Input variables
estimated transaction volume USD	Worldwide	Quandl	Input variables
estimated transaction volume	Worldwide	Quandl	Input variables
total output volume	Worldwide	Quandl	Input variables
number of transaction per block	Worldwide	Quandl	Input variables
Number of unique bitcoin addresses used	Worldwide	Quandl	Input variables
Number of transaction excluding popular addresses	Worldwide	Quandl	Input variables
Total transaction fee USD	Worldwide	Quandl	Input variables
Number of transaction	Worldwide	Quandl	Input variables
Market capitalization	Worldwide	Quandl	Input variables
Total Bitcoin	Worldwide	Quandl	Input variables
Wikipedia trend	Worldwide	Wikipedia	Input variables

The price of Bitcoin was converted to −1 for negative returns and 1 for positive returns on the next day. To have a normal distribution of the dataset, we normalized all the input variables by Z-transformation as follows:

$$Z_i = \frac{Xi - \overline{X}}{S}, \quad (1)$$

where Z_i is the transformed data, X_i the real dataset, $\overline{X}$ is the sample mean, and S is the standard deviation of the sample data.

3.2 Methodology

SVM, which deals with classification and regression tasks, was first proposed by Cortes and Vapnik (1995). It has become one of the most powerful machine learning algorithms for many applications. SVM delivers a robust hyperplane to separate a dataset into several different classes using a kernel. The classification function of SVM can be described as follows:

$$f(X) = \sum_{i=1}^{N} \propto_i y_i K(X, X_i) + b, \quad (2)$$

where $K(X,X_i)$ is a kernel function; $\propto_i$ and b are the controllable parameters used to tune model performance for greater accuracy in training and testing data processes. The

Copyright © 2021 GMP Press and Printing
ISSN: 2304-1013 (Online); 2304-1269 (CDROM); 2414-6722 (Print)

prevalently used kernel functions include dot, polynomial, radial, epachnenikov, multiquadric, and ANOVA. These kernel functions are defined as follows:

$$\text{Dot} = k(x,y) = x.y; \quad (3)$$

$$\text{Polynomial} = k(x,y) = (x^T.y + 1)^d, \quad (4)$$

where d is the polynomial's degree;

$$\text{Radial} = \exp(-g||x - y||^2), \quad (5)$$

where g is gamma that is adjustable and should be carefully determined;

$$\text{Epachnenikov} = ¾ (1 - u^2), \quad (6)$$

where u is between −1 and 1, where 0 is not included in the range; and

$$\text{Multiquadric} = ||x - y||^2 + c^2, \quad (7)$$

$$\text{Anova} = \exp(-g(x - y)) \quad (8)$$

where g is gamma and d is the degree.

SVM uses a structural risk minimization principle to reduce risks in the training process. This principle helps SVM construct the margin of separation to achieve maximum accuracy. However, SVM also demonstrates the quadratic programming problem, particularly when dealing with large datasets. Because quadratic programming contains quadratic complexity, long processing time and large memory are required when analyzing large datasets. For a broader picture regarding how SVM performs, the process is described in Figure 1.

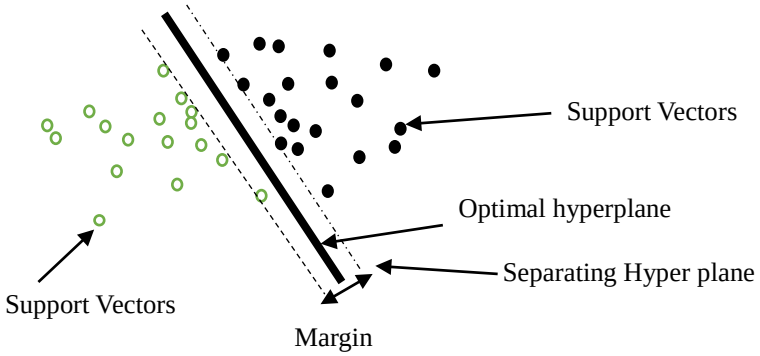


Fig.1 Support Vector Machine

The dots around the dashed line are called support vectors, and the thick black line between the dashed line is called the hyperplane. The hyperplane is an optimal line used to separate a dataset into several classes obtained using a kernel function. At the beginning, SVMs were used for binary classification and then was developed by Fung and Mangasarian (2005), Weston and Watkins (1998), and Nemmour and Chibani (1993) to work more efficiently with multiclass tasks. Furthermore, since 2000, Chang and Lin have worked on developing LIBSVM. In 2011, Chang and Lin published a library, LIBSVM, to support SVMs. LIBSVM has been used in many applications, such as that

Copyright © 2021 GMP Press and Printing
ISSN: 2304-1013 (Online); 2304-1269 (CDROM); 2414-6722 (Print)

in the studies of Lin *et al*. (2008) on anomaly detection in an intruded system, Lin *et al*. (2011) on predicting business failure, Chen and Hsiao (2008) on diagnosing a business crisis, Kim (2003) on financial time series forecasting, Lee and To (2010) on evaluating enterprise financial distress, Han and Chen (2007) on analyzing financial statements to predict stock movements, and Huang and Depari (2019) on analyzing the performance of paid ads in Facebook. For accurate model predictability, some parameters must need to be adjusted carefully, such as type of kernel function used, gamma (0.1, 1, 10, 100), and the regularization parameter, usually called C (0.1, 1, 10, 100, 1000), as in the study by Duan *et al*. (2003). C, a principal parameter in SVM, is used as a tool to control and determine the tradeoff between margin maximization and error minimization, according to Chapelle *et al*. (2002).

Deep learning is established on a multilayer feed-forward artificial neural network that is trained by gradient descend using backpropagation and representative learning studied by a neural network model, according to Chollet (2017). To understand how the calculation process works, the sequence is presented in Figure 2. Deep learning is a proven, powerful algorithm for dealing with a large and complex dataset, according to Najafabadi *et al*. (2015). Moreover, in the field of finance, Heaton *et al*. (2016) revealed that it has the ability to produce a large variety of important results than existing methodology.

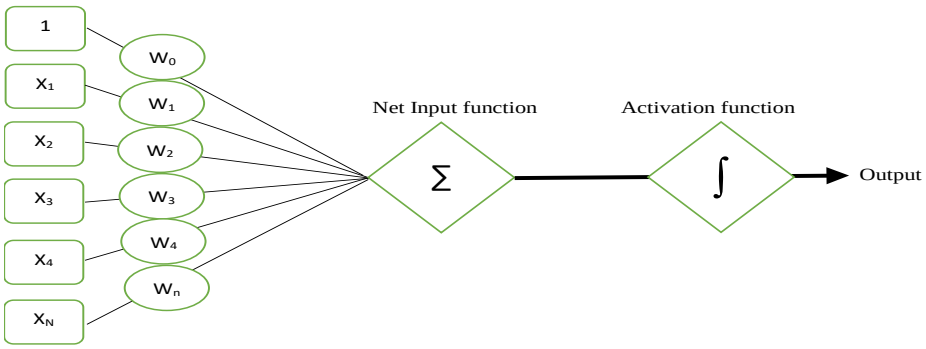


Figure. 1 Sequence of Deep Learning Calculation Process

The architecture of deep learning consists of input, hidden, and output layers. The four famous deep learning activation functions are Than, Rectifier, Maxout, and ExpRectifier. Therefore, we ran a grid search to identify the optimal activation function for dealing with Bitcoin data. Figure 2 illustrates the working of deep learning; the thickness of each line describes the quality of the relationship among nodes – the thicker the line, the closer is the relationship to a particular node.

Copyright © 2021 GMP Press and Printing
ISSN: 2304-1013 (Online); 2304-1269 (CDROM); 2414-6722 (Print)

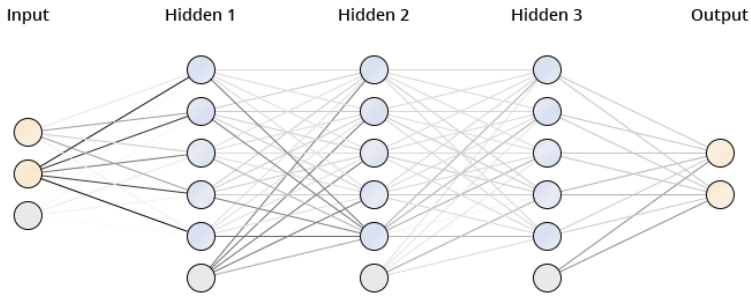


Fig. 3 Deep Learning

The decision tree is a famous type of machine learning algorithm. However, the concept of using only a single tree leads to risks of error and bias, according to Bhattacharyya *et al*. (2011). Therefore, Breiman (2001) proposed the concept of random forest. Random forests are collaborations of tree predictors in which each tree depends on the values of random samples in a forest. The error can be minimized by using a voting strategy in generating accuracy of prediction. Liaw and Wiener (2002) argued that, because it uses a few parameters, random forest is a simple and user-friendly algorithm. Another feature of random forest is its ability to determine the importance of input variables by calculating how much error occurs when data for a certain variable are changed while the other data remain unchanged.

We tested the ability of SVM, deep learning, and random forest in predicting the trend of Bitcoin returns and examined them by using the data mining software RapidMiner. The sequence and process of using RapidMiner are illustrated in Figure 4. For optimal parameters, we applied a grid search method for each algorithm. To reduce dimensionality, we also optimized the number of input variables to be used by performing the "weight by information gain ratio" analysis for each algorithm. We then assessed the performance of the algorithms by comparing them in terms of accuracy, ROC curve, and processing time. The complete sequence can be seen in Figure 3. The first step is feeding the data into the system then preprocessing the data such as by converting the string variables into numerals, normalizing the input variables, and modifying the Bitcoin price into 1 and −1 (classification); 1 denotes a positive return and −1 denotes a negative return.

The second step was performing a dimensionality reduction by weighting and selecting the input variables using the information gain ratio technique. The information gain ratio was first proposed by Quinlan (1986) to reduce error and bias by selecting the most appropriate attributes. The information gain ratio is defined as follows:

$$Information\ gain\ ratio = \frac{Information\ gain}{Split\ information} \tag{9}$$

Information gain determines and collects the information using training data results. The information gain ratio attempts to reduce bias by using split information to correct the calculation of information gain. The information gain ratio was proven by Yoshida (2001) an effective summarization method in terms of weighting. Therefore, we applied the technique to reduce the number of attributes selected.

Copyright © 2021 GMP Press and Printing
ISSN: 2304-1013 (Online); 2304-1269 (CDROM); 2414-6722 (Print)

For an optimal number of input variables with respect to prediction accuracy, we ran a grid search method. Furthermore, selected input variables were used as the next input variables for the algorithms. We also performed a grid search method to discover the optimal parameters for the algorithms. The optimized parameters of deep learning were the number of hidden layers (1–10), the activation function, learning rate (0.1–1.00), local random seed (true/false), and reproducible (true/false). The optimized parameters of SVM were kernel functions, shrinking (true/false), and confidence for multiclass (true/false). The optimized parameters of a random forest were the number of trees (1–100), maximal depth (1–100), criterion (gain ratio/information gain/gini index/accuracy), voting strategy (confidence vote/majority vote), and apply prepruning (true/false).

The third step was to split the dataset using the split validation technique into training (80%) and test (20%). Split validation was performed to avoid the overfitting problem.

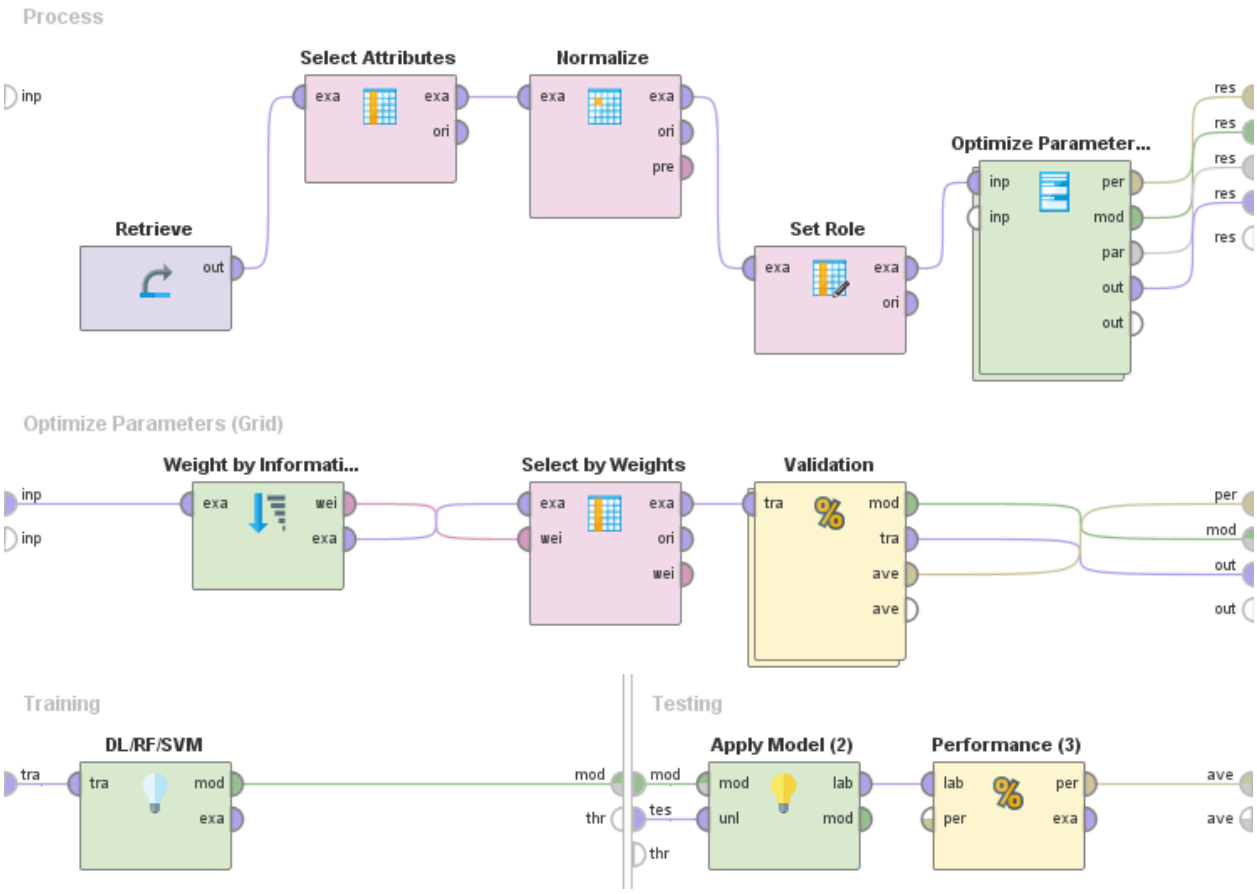


Fig. 4 Data Mining Sequence using Rapidminer

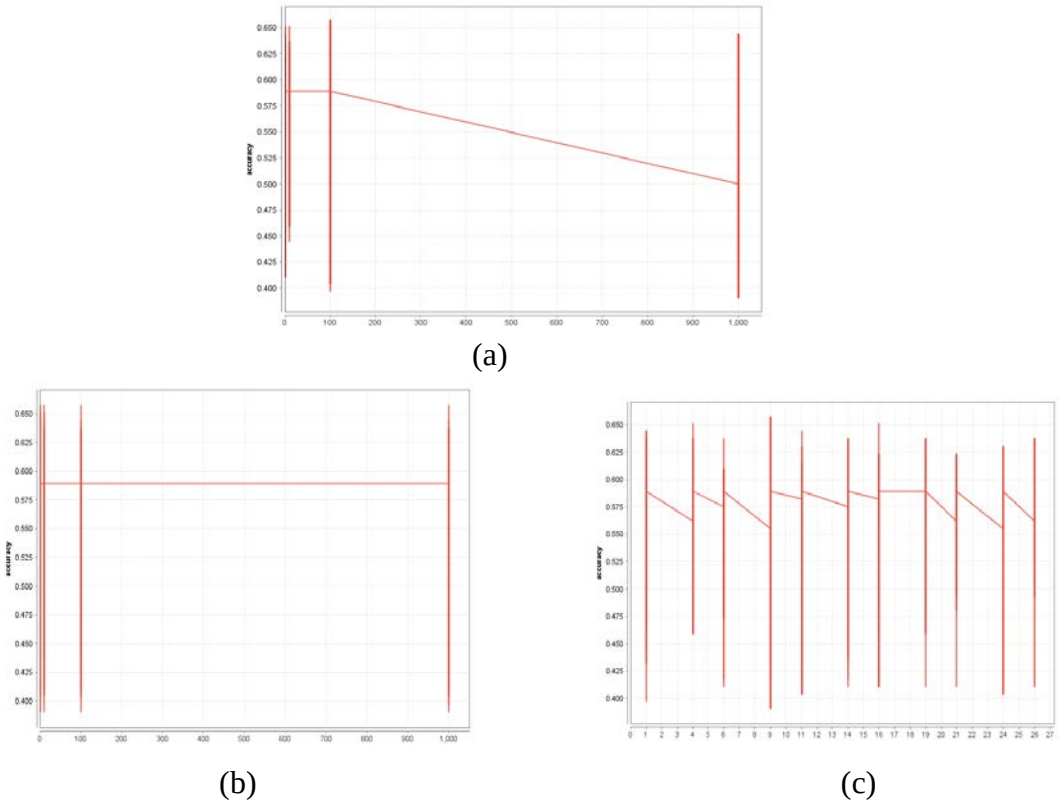
4. DISCUSSION AND RESULTS

4.1 Algorithm comparison

Copyright © 2021 GMP Press and Printing
ISSN: 2304-1013 (Online); 2304-1269 (CDROM); 2414-6722 (Print)

Table 1. The results of grid search of Support Vector Machine

Gamma	Kernel Function	Number of Variables	C	Accuracy
0.1	epachnenikov	9	100.0	0.657
1	epachnenikov	9	100.0	0.657
10	epachnenikov	9	100.0	0.657
100	epachnenikov	9	100.0	0.657
.	.	.		.
.	.	.		.
.	.	.		.
10	dot	9	1000	0.390
100	dot	9	1000	0.390
1000	dot	9	1000	0.390



(a)

(b) (c)

Figure 5. The result of *C* (a), kernel gamma (b), and (c) number of selected variables Grid search

In applying the grid search method, three SVM parameters were combined and compared to achieve better prediction accuracy. Those parameters were kernel function (dot, polynomial, radial, Epanechnikov, multiquadric, and ANOVA), the value of C, and kernel gamma. We also optimized the number of selected variables for high accuracy (number of k). We discovered that the optimal kernel was Epanechnikov, and the optimal number of selected input variables was nine; with kernel gamma 100 and the value of C

Copyright © 2021 GMP Press and Printing
ISSN: 2304-1013 (Online); 2304-1269 (CDROM); 2414-6722 (Print)

0.1, it achieved 65.7% accuracy (Table 1). We also discovered that the difference value of kernel gamma afforded the same impact on our model (Table 1). These phenomena emphasize that the kernel gamma did not much affect the model in reaching high accuracy. When predicting a positive return, the model was able to achieve 69.15% accuracy, but when predicting a negative return it was only 51.67% accurate. These results can be seen in the SVM confusion matrix.

Table 2. The confusion matrix of SVM

	true 1	true -1	class precision
pred. 1	65	29	69.15%
pred. -1	21	31	59.62%
class recall	75.58%	51.67%	

The same method was conducted to optimize the deep learning parameters. The parameters were k (number of variables weighted by information gain ratio), activation function (maxout/Rectifier/Tanh/ExpRectifier), reproducible (False/True), local random seed (True/False), and learning rate (0.1–1.0). We also attempted to optimize the number of hidden layers used. Therefore, we tested 1–10 hidden layers to achieve the highest accuracy. We discovered that applying two hidden layers achieved the highest accuracy. The results are described in Table 3.

Table 3. Optimization of Hidden Layers

Number of Hidden layer	Accuracy	Time processing
1	60.27%	14:59
2	64.38%	51:54
3	59.59%	32:46
4	59.59%	1:13:17
5	57.53%	1:32:30
6	55.48%	1:52:32
7	54.11%	2:12:26
8	55.48%	2:33:52
9	54.79%	2:51:01
10	52.05%	3:11:23

We therefore applied two hidden layers to the deep learning model. Eventually, 26 variables were selected as the number of optimum input variables; maxout activation was found as the most accurate activation, with a 10% learning rate and both reproducible and local random seed set as false (Table 4). These grid results led the model to achieve 64.4% accuracy in predicting the trend of Bitcoin.

Copyright © 2021 GMP Press and Printing
ISSN: 2304-1013 (Online); 2304-1269 (CDROM); 2414-6722 (Print)

Table 4. Grid search result of Deep Learning

Iterations	Number of variables	Activation function	Reproducible	Local random seed	Learning Rate	Accuracy
341	26	Maxout	FALSE	FALSE	0.1	0.643836
198	26	Rectifier	TRUE	TRUE	0.1	0.623288
1331	26	Tanh	TRUE	FALSE	0.7	0.623288
1583	24	ExpRectifier	FALSE	FALSE	0.8	0.609589
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
1906	6	Rectifier	FALSE	FALSE	1	0.356164
793	1	Tanh	TRUE	FALSE	0.4	0.335616
1618	1	ExpRectifier	TRUE	TRUE	0.9	0.335616

For a broader picture, we also provide a confusion matrix of the deep learning prediction model. The overall prediction accuracy of the model was 64.4%, which is slightly better than that of the SVM prediction model (64%). However, in predicting the positive return for the next day, deep learning was proven better than SVM. Therefore, when dealing with this dataset, deep learning performs slightly better than SVM.

Table 5. The confusion matrix of Deep Learning

	true 1	true -1	class precision
pred. 1	40	14	74.07%
pred. -1	38	54	58.70%
class recall	51.28%	79.41%	

The final candidate algorithm was random forest. Random forest is a supervised learning algorithm that is an ensemble of a decision tree and a bagging method to train data. In this study, we optimized the following parameters: the number of trees, maximal depth, applied pre pruning, criterion, voting strategy, and the number of selected variables. The results are reported in Table 6. Using the optimal parameters, the model reached 70.55% accuracy, the highest of the three algorithms. To reach this level of accuracy, random forest had to reach 14,666 iterations, higher than the number of iterations used by SVM and deep learning. This means its processing time was also the longest.

Copyright © 2021 GMP Press and Printing
ISSN: 2304-1013 (Online); 2304-1269 (CDROM); 2414-6722 (Print)

Table 6. Grid Search Result of Random Forest

Number of trees	Maximal Depth	Apply pre prunning	Criterion	Voting Strategy	Number of variables	Accuracy
21	21	FALSE	gain_ratio	majority vote	19	0.705479
100	31	FALSE	gini_index	confidence vote	21	0.684932
100	1	TRUE	gain_ratio	majority vote	4	0.678082
41	70	TRUE	gini_index	majority vote	4	0.678082
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
1	31	TRUE	gain_ratio	majority vote	14	0.383562
1	1	FALSE	gini_index	majority vote	16	0.383562
1	1	FALSE	gain_ratio	confidence vote	16	0.356164

To present model performance, a confusion matrix is included in Table 7. From a total of 146 samples tested, 103 were precise predictions, and the remainder was error predictions. In predicting of the positive returns, the model could reach 70.68% accuracy, which is not better than the deep learning model (74% accuracy). This difference shows us that the deep learning model was robust in predicting the positive returns but not in predicting the negative returns. Therefore, deep learning is a potential model to predict the returns to Bitcoin. However, since the optimization of the number of neurons and hidden layers requires powerful computational power, the optimization was limited in this research.

Table 7. Confusion matrix of Random Forest

	true 1	true -1	class precision
pred. 1	94	39	70.68%
pred. -1	4	9	69.23%
class recall	95.92%	18.75%	

We also assessed algorithm performances by comparing their ROC curves. A ROC curve is a quantitative method commonly used for binary classification. ROC curves were first used in aircraft detection to distinguish enemy aircraft based on noise; it refers

Copyright © 2021 GMP Press and Printing
ISSN: 2304-1013 (Online); 2304-1269 (CDROM); 2414-6722 (Print)

to a "Chain Home," according to Galati (2016). We compared the performance of SVM, deep learning, and random forest by comparing the true- and false-positive rates. The blue, red, and green lines represent the SVM, random forest, and deep learning performance, respectively. Therefore, when using this dataset, random forest slightly outperformed the other algorithms (Figure 6).

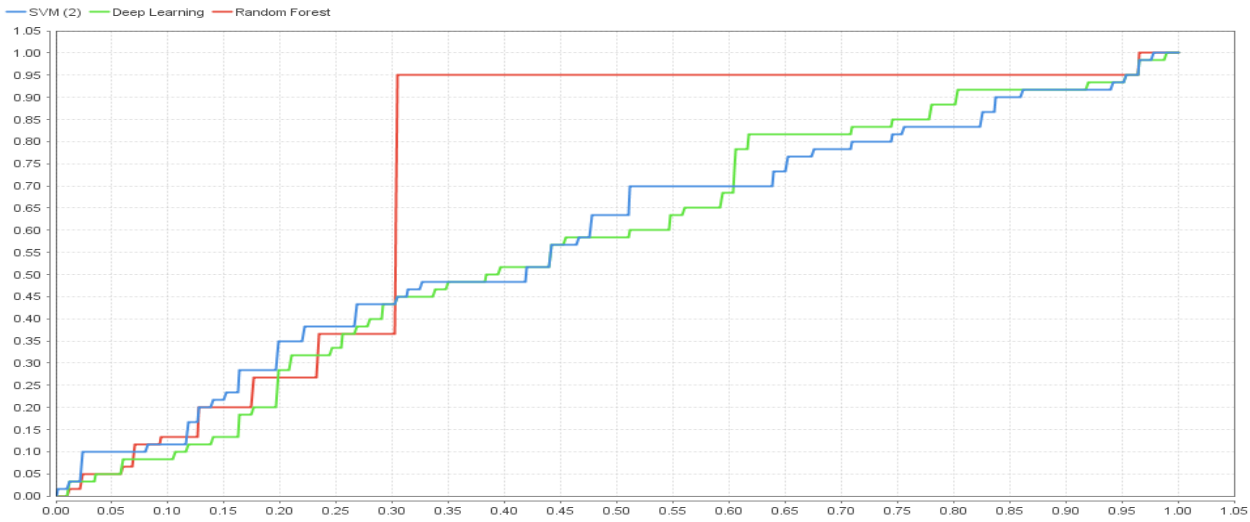


Fig 6. Algorithm Comparison by ROC curve

The summary of the performance comparison among the machine learning algorithms is presented in Table 8. Random forest achieved the highest accuracy. However, it also required more time for processing. SVM had the lowest processing time but a low prediction accuracy. These results are aligned with Kumar and Thenmozhi's 2006 study, in which random forest outperformed the neural network model in predicting stock index movement. Random forest also excelled over other algorithms used in literature, according to Khaidem *et al.* (2016). Therefore, we employed random forest to weight the importance of 27 input variables and to propose a strategy as a managerial implication to help traders make decisions in the Bitcoin market.

Table 8. Algorithm Comparisons

Algorithms	Accuracy	Time Processing	Precision (Pred -1)	Precision (Pred 1)
Support Vector Machine		3:17:31		
Deep Learning	64.40%	1:38:52	58.70%	74.07%
Random Forest	70.55%	5:47:45	69.23%	70.68%

4.2 Weighting results and managerial implication

As random forest is a robust algorithm, we applied random forest together with the optimum parameters to weigh the input variables. There were 26 input variables and one label variable (i.e., the return trend of Bitcoin). As can be seen in Figure 7, eight ranks or classes of variable importance exist with respect to the return trend of Bitcoin (i.e.,

Copyright © 2021 GMP Press and Printing
ISSN: 2304-1013 (Online); 2304-1269 (CDROM); 2414-6722 (Print)

negative and positive returns). The key factors that traders should be aware of are total Bitcoin traded and API blockchain size. API blockchain size is the total size of all transactions and block headers in the blockchain system. Both of these variables are in the first class. Nevertheless, the weight differences among the other classes are only slightly different, except the type of day (weekend or weekday) and day of the week, which were the least valuable variables. The total volume of Bitcoin traded and the API of Bitcoin are considered internal-related factors that can be monitored through the blockchain system platform and are also provided as free access information to traders and researchers.

Economic policy uncertainties in the U.S. has strong relevance for Bitcoin returns, as proven by the importance results generated by this study. U.S. economic policy uncertainties occupy the second rank of importance. This result is also supported by the study conducted by Demir *et al*. (2018), who discovered that economic policy uncertainties had a strong negative influence on Bitcoin price. Align with this result, Panagiotidis *et al.* (2018) revealed that Bitcoin responds to US economic policy uncertainties in high influence. Surprisingly, Economic policy uncertainties in the U.K. is of little relevant to Bitcoin returns, as can be seen by its fifth rank among all economic policy uncertainties. By this phenomenon, we can see the different levels of influence generated by Economic policy uncertainties in the two nations. Therefore, Economic policy uncertainties in U.S. should be closely watched by traders and researchers. On the other hand, the type of day (weekend or weekday) was not a relevant factor for Bitcoin returns. This result is expected because buying and selling Bitcoin is more affected by external and internal factors.

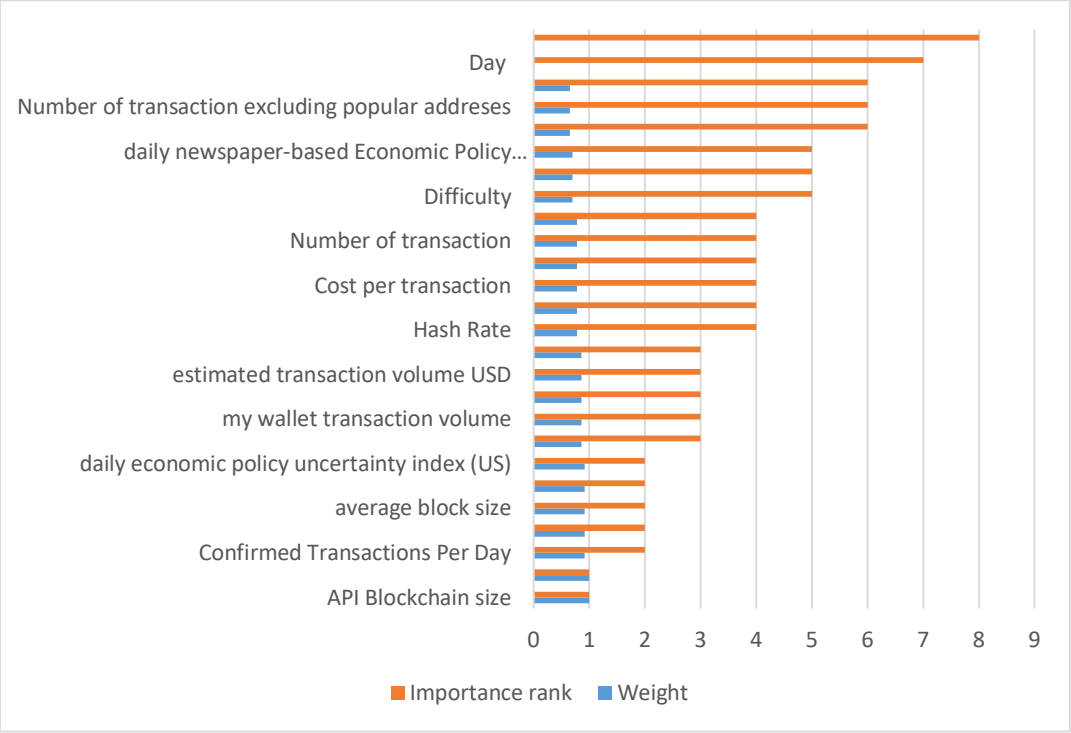


Figure 7. The importance of Input Variables

Copyright © 2021 GMP Press and Printing
ISSN: 2304-1013 (Online); 2304-1269 (CDROM); 2414-6722 (Print)

5. CONCLUSION AND FUTURE RESEARCH

This study employed a data mining approach to predict the daily returns of Bitcoin and revealed the most impactful factors affecting Bitcoin price. There were 26 input variables and one label variable involved, which cover both internal- and external-related factors affecting Bitcoin price. To reduce dimensionality, the information gain ratio technique was applied. Several methodologies including support vector machine, deep learning, and random forest were performed to assess and compare the performance of these algorithms. As a result random forest with 70.5% accuracy was selected as the most accurate algorithm.

The method of random forest was then used to weight the importance of the 26 input variables. The results indicated that the Bitcoin trading volume and API blockchain size were the most important input variables in predicting the returns to Bitcoin on a daily basis. However, among the other variables (except for the day of the week and weekend/weekday), the importance percentage differed only slightly. Therefore, predicting Bitcoin returns using these 26 input variables is not easy. Both internal- and external-related factors can influence the returns to Bitcoin but at a close degree. Day of the week and type of day (weekday or weekend) were the least relevant factors affecting the returns to Bitcoin. The U.S. economic policy uncertainties was ranked the second in terms of input variable importance. The U.K. economic policy uncertainties was relatively less relevant to Bitcoin returns prediction, as this factor was ranked only the fifth in terms of input variable importance.

This study contributes by emphasizing a simple, concise, and robust methodological framework that might help traders and researchers to predict the returns to Bitcoin. This study also provide managerial implications for traders to develop buying and selling strategies in the Bitcoin market based on the relative importance of various input variables. A possible limitation is that, among the 27 attributes considered in this study, the weighting results contained only slight differences in absolute terms. Collecting a larger dataset on daily basis may provide broader and more complete information, which may reveal other useful hidden information for predicting Bitcoin returns.

ACKNOWLEDGEMENT

The authors greatly appreciate all parties for their constructive advice and supervision from the beginning to the end of the research process. Special gratitude goes to Professor Jen-Peng Huang for his kindness and effort in guiding the first author. Moreover, the authors convey many thanks to the academic editors and reviewers of the journal *Review of Integrative Business and Economics Research* for creating a space for collaboration in the international research community.

Copyright © 2021 GMP Press and Printing
ISSN: 2304-1013 (Online); 2304-1269 (CDROM); 2414-6722 (Print)

REFERENCES

[1] Aalborg, H.A., Molnár, P. and de Vries, J.E., 2018. What can explain the price, volatility and trading volume of Bitcoin?. *Finance Research Letters.*

[2] Baur, D.G., Hong, K. and Lee, A.D., 2018. Bitcoin: Medium of exchange or speculative assets?. *Journal of International Financial Markets, Institutions and Money*, *54*, pp.177-189.

[3] Bhattacharyya, S., Jha, S., Tharakunnel, K. and Westland, J.C., 2011. Data mining for credit card fraud: A comparative study. *Decision Support Systems*, *50*(3), pp.602-613.

[4] Breiman, L., 2001. Random forests. *Machine learning*, *45*(1), pp.5-32.

[5] Cao, L.J. and Tay, F.E.H., 2003. Support vector machine with adaptive parameters in financial time series forecasting. *IEEE Transactions on neural networks*, *14*(6), pp.1506-1518.

[6] Chang, C.C. and Lin, C.J., 2011. LIBSVM: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, *2*(3), p.27.

[7] Chapelle, O., Vapnik, V., Bousquet, O. and Mukherjee, S., 2002. Choosing multiple parameters for support vector machines. *Machine learning*, *46*(1-3), pp.131-159.

[8] Chen, L.H. and Hsiao, H.D., 2008. Feature selection to diagnose a business crisis by using a real GA-based support vector machine: An empirical study. *Expert Systems with Applications*, *35*(3), pp.1145-1155.

[9] Chollet, F., 2017. *Deep learning with python*. Manning Publications Co..

[10] Ciaian, P., Rajcaniova, M. and Kancs, D.A., 2016. The economics of BitCoin price formation. *Applied Economics*, *48*(19), pp.1799-1815.

[11] Cortes, C. and Vapnik, V., 1995. Support-vector networks. *Machine learning*, *20*(3), pp.273-297.

[12] Demir, E., Gozgor, G., Lau, C.K.M. and Vigne, S.A., 2018. Does economic policy uncertainty predict the Bitcoin returns? An empirical investigation. *Finance Research Letters*, *26*, pp.145-149.

[13] Duan, K., Keerthi, S.S. and Poo, A.N., 2003. Evaluation of simple performance measures for tuning SVM hyperparameters. *Neurocomputing*, *51*, pp.41-59.

[14] Fung, G.M. and Mangasarian, O.L., 2005. Multicategory proximal support vector machine classifiers. *Machine learning*, *59*(1-2), pp.77-97.

[15] Gabralla, L.A., Jammazi, R. and Abraham, A., 2013, August. Oil price prediction using ensemble machine learning. In *2013 INTERNATIONAL CONFERENCE ON COMPUTING, ELECTRICAL AND ELECTRONIC ENGINEERING (ICCEEE)*(pp. 674-679). IEEE.

[16] Galati, G., 2016. *100 years of radar*. Springer.

[17] Georgoula, I., Pournarakis, D., Bilanakos, C., Sotiropoulos, D. and Giaglis, G.M., 2015. Using time-series and sentiment analysis to detect the determinants of bitcoin prices. *Available at SSRN 2607167*.

[18] Han, S. and Chen, R.C., 2007. Using svm with financial statement analysis for prediction of stocks. *Communications of the IIMA*, *7*(4), p.8.

Copyright © 2021 GMP Press and Printing
ISSN: 2304-1013 (Online); 2304-1269 (CDROM); 2414-6722 (Print)

[19] Huang, J.P. and Depari, G.S., 2019. Paid Advertisement on Facebook: An Evaluation Using a Data Mining Approach. *Review of Integrative Business and Economics Research*, *8*(4), p.1.

[20] Karasu, S., Altan, A., Saraç, Z. and Hacioğlu, R., 2018, May. Prediction of Bitcoin prices with machine learning methods using time series data. In *2018 26th Signal Processing and Communications Applications Conference (SIU)* (pp. 1-4). IEEE.

[21] Khaidem, L., Saha, S. and Dey, S.R., 2016. Predicting the direction of stock market prices using random forest. *arXiv preprint arXiv:1605.00003*.

[22] Kim, K.J., 2003. Financial time series forecasting using support vector machines. *Neurocomputing*, *55*(1-2), pp.307-319.

[23] Larose, D.T. and Larose, C.D., 2014. *Discovering knowledge in data: an introduction to data mining*. John Wiley & Sons.

[24] Lee, M.C. and To, C., 2010. Comparison of support vector machine and back propagation neural network in evaluating the enterprise financial distress. *arXiv preprint arXiv:1007.5133*.

[25] Liaw, A. and Wiener, M., 2002. Classification and regression by random forest. *R news*, *2*(3), pp.18-22.

[26] Lin, C.H., Liu, J.C. and Ho, C.H., 2008, April. Anomaly detection using LibSVM training tools. In *2008 International Conference on Information Security and Assurance (isa 2008)*(pp. 166-171). IEEE.

[27] Lin, F., Yeh, C.C. and Lee, M.Y., 2011. The use of hybrid manifold learning and support vector machines in the prediction of business failure. *Knowledge-Based Systems*, *24*(1), pp.95-101.

[28] McNally, S., Roche, J. and Caton, S., 2018, March. Predicting the price of Bitcoin using Machine Learning. In *2018 26th Euromicro International Conference on Parallel, Distributed and Network-based Processing (PDP)* (pp. 339-343). IEEE.

[29] Najafabadi, M.M., Villanustre, F., Khoshgoftaar, T.M., Seliya, N., Wald, R. and Muharemagic, E., 2015. Deep learning applications and challenges in big data analytics. *Journal of Big Data*, *2*(1), p.1.

[30] Nakano, M., Takahashi, A. and Takahashi, S., 2018. Bitcoin technical trading with artificial neural network. *Physica A: Statistical Mechanics and its Applications*, *510*, pp.587-609.

[31] Nemmour, H. and Chibani, Y., 1993, August. Multi-class SVMs based on fuzzy integral mixture for handwritten digit recognition. In *Geometric Modeling and Imaging--New Trends (GMAI'06)* (pp. 145-149). IEEE.

[32] Panagiotidis, T., Stengos, T. and Vravosinos, O., 2018. On the determinants of bitcoin returns: A LASSO approach. *Finance Research Letters*, *27*, pp.235-240.

[33] Park, B. and Bae, J.K., 2015. Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data. *Expert Systems with Applications*, *42*(6), pp.2928-2934.

[34] Patel, J., Shah, S., Thakkar, P. and Kotecha, K., 2015. Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques. *Expert Systems with Applications*, *42*(1), pp.259-268.

[35] Pichl, L. and Kaizoji, T., 2017. Volatility analysis of bitcoin. *Quantitative Finance and Economics*, *1*, pp.474-485.

Copyright © 2021 GMP Press and Printing
ISSN: 2304-1013 (Online); 2304-1269 (CDROM); 2414-6722 (Print)

[36] Poyser, O., 2017. Exploring the determinants of Bitcoin's price: an application of Bayesian Structural Time Series. *arXiv preprint arXiv:1706.01437*.
[37] Quinlan, J.R., 1986. Induction of decision trees. *Machine learning*, *1*(1), pp.81-106.
[38] Sivalingam, K.C., Mahendran, S. and Natarajan, S., 2016. Forecasting gold prices based on extreme learning machine. *International Journal of Computers Communications & Control*, *11*(3), pp.372-380.
[39] Turban, E., Sharda, R. and Delen, D., 2010. Decision Support and Business Intelligence Systems (required). *Google Scholar*.
[40] Weston, J. and Watkins, C., 1998. *Multi-class support vector machines* (pp. 98-04). Technical Report CSD-TR-98-04, Department of Computer Science, Royal Holloway, University of London, May.
[41] Yoshida, T.M.M.K.K., 2001. Term weighting method based on information gain ratio for summarizing documents retrieved by IR systems. *Journal of Natural Language Processing*, *9*(4), pp.3-32.

Copyright © 2021 GMP Press and Printing
ISSN: 2304-1013 (Online); 2304-1269 (CDROM); 2414-6722 (Print)