Volume: 3 Issue: 1 Pages: 24-31

Classification of Crypto-Currency Data Using Data Mining Techniques

[*1] Atish Tangawade and [2]Aniket Muley

[*1] Research Scholar, School of Mathematical Sciences, Swami Ramanand Teerth Marathwada University, Nanded, Maharashtra, India.

[2]Associate Professor, School of Mathematical Sciences, Swami Ramanand Teerth Marathwada University, Nanded, Maharashtra, India.

Article Info.
E-ISSN: 2583-6528
Impact Factor (SJIF): 5.231
Peer Reviewed Journal
Available online:
www.alladvancejournal.com

Received: 17/Dec/2023
Accepted: 15/Jan/2024

***Corresponding Author**
Atish Tangawade
Research Scholar, School of Mathematical Sciences, Swami Ramanand Teerth Marathwada University, Nanded, Maharashtra, India.

Abstract
Crypto-currencies became popular with the emergence of Bitcoin and Litcoin have shown an unprecedented growth over the last few years. After more than a decade of existence, crypto-currencies may now be considered an important class of assets presenting some unique appealing characteristics but also sharing some features with real financial assets. Crypto-currencies price behaviour is still largely unexplored, presenting new opportunities for researchers to highlight similarities and differences with standard financial prices. Consequently, the cryptocurrency market can be a conducive arena for investors, as it offers many opportunities. However, it is difficult to understand. The current article focused on the Bitcoin and Litcoin Cryptocurrency informational collection with classification perspective. Our idea is to recognize the Artificial Intelligence (AI) device that, classify Cryptocurrency brings about more proficient way. To perform arrangement expectation, we have applied different data mining methods viz., Decision Tree (DT), Random Forest (RF), Extreme Boost model, Support Vector Machine (SVM) Model, Linear and Generalized Linear Models, brain organization or neural network (NN) model, beneficiary administrator trademark (ROC) bend. We found that, general linear model is more liked and it gives a lot of proficient outcomes. In this paper, various algorithms of data mining are demonstrated and outlined comprehensively for the Cryptocurrency data. Overall, dealing with the large scale dataset, the results obtained through the various machine learning techniques we found that, low price, high price, volume of the Crypto currency parameters plays significant role in classification and in future this study will be useful for prediction perspective.

Keywords: Classification, Cryptocurrency, Data Mining, Optimization.

1. Introduction

In the era of information as well as communication technologies, the exponential growth and development in financial technology have been observed in our daily life. Consequently, many accomplishments, precisely financial activities, have been merged digitally and they become more malleable and effective. Nowadays, to promote the financial activities in a virtual way, a new business phenomenon Crypto-currency, facilitate buying, selling and trading. A Crypt-currency is a form of virtual or digital asset based on a network in financial systems. A huge growth in number of online users has activated this virtual word concept which is secured by cryptography that allows them to exist outside the control of governments and central authorities. This delimits decentralized structure that does not collapse at a single point of failure, advancing the cheaper and faster money transfers. There are almost 5000 crypto-currencies and near about 6 million active users present in the market but bitcoin, which is established in 2009, is the most popular and valuable crypto-currency. Bitcoin is the most widely traded and covered Crypto-currency since it uses peer-to-peer technology to facilitate instant payments without cash.

Nakamoto (2008) [1] invented it and introduced it to the world via a white paper. Each record of bitcoin is encrypted and the balances of it kept on a public ledger so that everyone has transparent access. Bitcoin is a digital currency not backed by real assets or tangible securities. They are traded between consenting parties with no broker and tracked on digital ledgers. Bitcoin is created, distributed, traded, and stored with the use of a decentralized ledger system, known as a block chain, can be thought of as a collection of blocks having transactions with unique key. The new data is entered into a fresh block and once the block is filled, it is chained onto the previous block, which makes the data chained together in

chronological order and hence, it is called as block chain technology. The benefit of using this technology is, all of the computers running the block chain have the same list of blocks and transactions and can transparently see these new blocks as they're filled with new bitcoin transactions, no one can cheat the system since it is not editable. Block chain is a type of shared database that plays a crucial role in the system of Crypto-currency for maintaining a secure and decentralized record of transactions. It is trustworthy and secure to record data and generates trust without the need for a trusted third party. Various machine learning models have been studied by plenty of researchers to predict the accurate prices of crypto currencies, to discover the hidden patterns from the data, etc. due to price volatility and dynamism of Crypto-currency. We overviewed the literature which is given below.

Fatah *et al*. (2020) [5] predicts crypto-currency prices such as bitcoin, ethereum and ripple using data mining algorithms such as K-NN, Neural Network, SVM, Linear Regression, Random Forest and Decision Tree. Data mining modelling is done by dividing the dataset into each type of commodity and then analyzed using each algorithm. Hamayel and Owda (2021) [2] proposed three types of recurrent neural network algorithms used to predict the prices of three types of crypto-currencies, namely bitcoin, litecoin and ethereum. Tanwar *et al*. (2021) [3] proposes a deep learning based hybrid model to predict the price of litecoin and zcash with interdependency of the parent coin. Marne *et al*. (2021) predicts the price of crypto-currency using recurrent neural network, support vector machines, multilayer perceptron, etc. and compare the accuracy and efficiency within these models using root mean square error. Liu and Tsyvinski (2021) [7] establish that crypto-currency returns are driven and can be predicted by factors that are specific to crypto-currency markets. They constructed the network factors to capture the user adoption of crypto-currencies and the production factors to proxy for the costs of crypto-currency production using time series analysis. Fang *et al*. (2022) [6] provides a comprehensive survey of crypto-currency trading research, by covering 146 research papers on various aspects of crypto-currency trading, analyses datasets, research trends and distribution among research objects and technologies, concluding with some promising opportunities that remain open in crypto-currency trading.

Crypto-currency is anonymous form of transaction which is highly volatile, uncertain and unpredictable within price, high energy consumption for mining activities because anyone can mine them using a computer with an Internet connection and it can be easily used in criminal activities. On this background of review, are motivated to study the uncertainty within the transaction related to the crypto-currency and also, investing in crypto-currencies is highly risky and speculative.

In this paper, our interest is to identify the suitable technique for classification crypto-currency, to identify the significant parameters that help to classify it with different machine learning methods. As Cryptocurrency is one of way to invest your money in the respective form. Very few researchers focused on Cryptocurrency data. It is essential to investor that, specific parameters that might be helpful in appropriate selection for investment in Cryptocurrency based on the available market data. In the next subsequent sections, methodology, results of the study and conclusions is discussed in detail.

2. Materials and Methods

In this article, auxiliary information is accumulated from open-source site [8], to play out the investigation with the assistance of AI apparatuses. To perform classification, we have applied various machine learning techniques viz., Decision Tree (DT), Random Forest (RF), Extreme Boost model, Support Vector Machine (SVM) Model, Linear and Generalised Linear Models, neural network (NN) model, receiver operator characteristic (ROC) curve. To perform visualization and analysis of the dataset, R 3.4.0Version is used along-with rattle package.

Following are the steps used for identification of crypto-currency parameters and further, classify with machine learning tools.

Step 1: Collection of Cryptocurrency data from the Open source website [8].

Step 2: Extract the Bitcoin and Litcoin data from overall dataset.

Step 3: Clean the extracted dataset with summarization.

Step 4: To identify the important variables under study apply DT model, RF, Extreme Boost model, Linear and Generalised Linear Models.

Step 5: To classify the Cryptocurrency dataset using SVM model with different functions. Identify suitable NN model with different hidden neurons that optimizes the classification.

Step 6: To identify the suitable classifier, split the data into train: test: validation i.e., 70:15:15 percentage and overall count form.

Step 7: Use ROC curve to compare the classifier models that optimizes the Cryptocurrency classification.

3. Result and discussion

In this section, above proposed steps are implemented on the dataset and results are explored below:

3.1 Decision Tree

For the classification as well as for the regression problems, DT can be used. It is a supervised learning technique mostly preferred for solving classification problems. It gives all possible solutions to a problem/decision based on given conditions graphically. Here, bitcoin and lit coin are classified using DT because it usually imitates human thinking ability while making a decision. Table 1 represents the summary of the DT model showed that, as numbers of splits are increasing the respective error get decrease. It helps us to increase the accuracy of the classified data set. The DT models classification root node error is 0.49678 with sample of size 4187 with 0.10 secs.

Table 1: Summary of the DT model

S. No.	CP	NSPLIT	REL ERROR	XERROR	XSTD
1	0.905288	0	1.0000000	1.0168269	0.0155538
2	0.032212	1	0.0947115	0.0951923	0.0066031
3	0.012500	3	0.0302885	0.0307692	0.0038166
4	0.010000	5	0.0052885	0.0057692	0.0016630

Fig. 1 represents the DT of the crypto-currency dataset and it shows that, price and volume plays an important role in classification of bitcoin and lit coin.

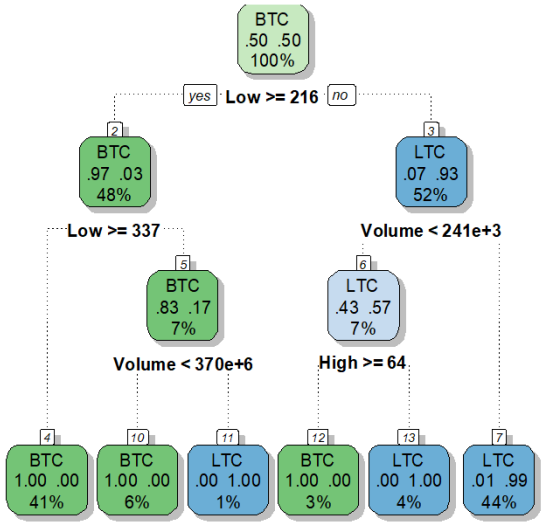

Fig 1: Decision Tree

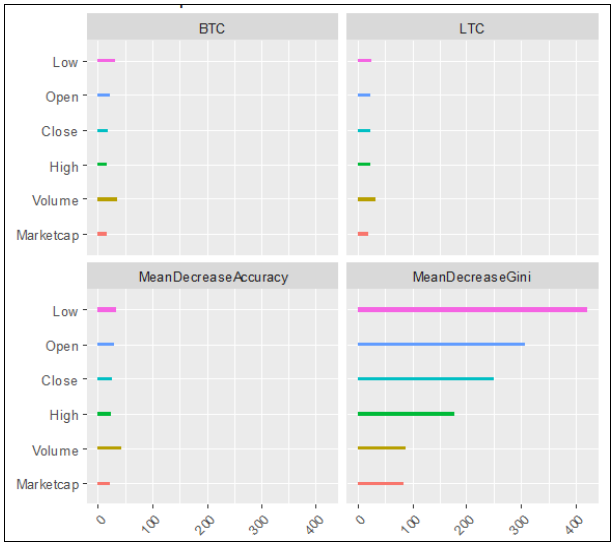

Fig 2: Variable importance

3.2 Random Forest Model

It is a gathering of un-pruned choice trees. RF is utilized many times when we have enormous preparation datasets and especially an exceptionally huge number of info factors. The calculation is productive as for countless factors since it over and over subsets of the accessible factors and is typically wound up tens or many DT and to see the pace of reduction of the model mistake as the quantity of trees increments. Fig. 2 zeroed in on the variable significance perception. Here, 4187 number of perceptions works in this the model with OOB gauge of blunder rate 0.05% and it is graphically envisioned in Fig. 3. Bitcoin gives us least variety as contrasted and litcoin. Table 2 addresses the disarray framework produced by RF. It investigates the grouping of bitcoin and litcoin. The probability of misclassification is least in the said model. Table 3 addresses the variable significance of bitcoin and litcoin for different boundaries considered in this review. Fig. 4 investigates region under bend and it gives hit rate to the misleading problem rate is 1. It essentially shows the effectiveness of the RF model.

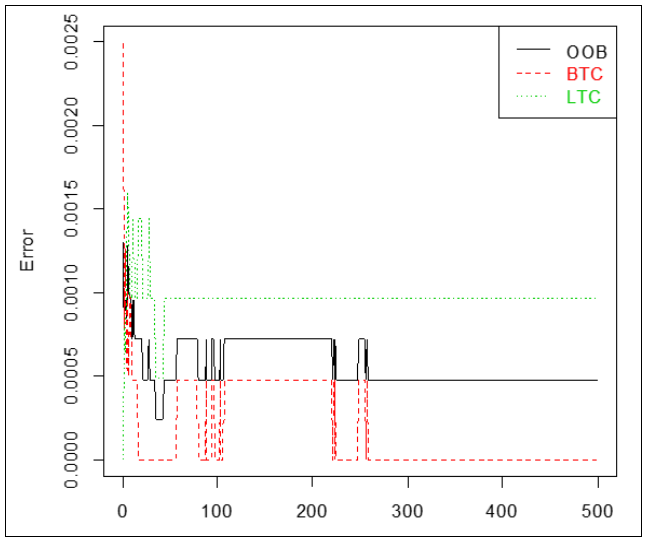

Fig 3: Error rate

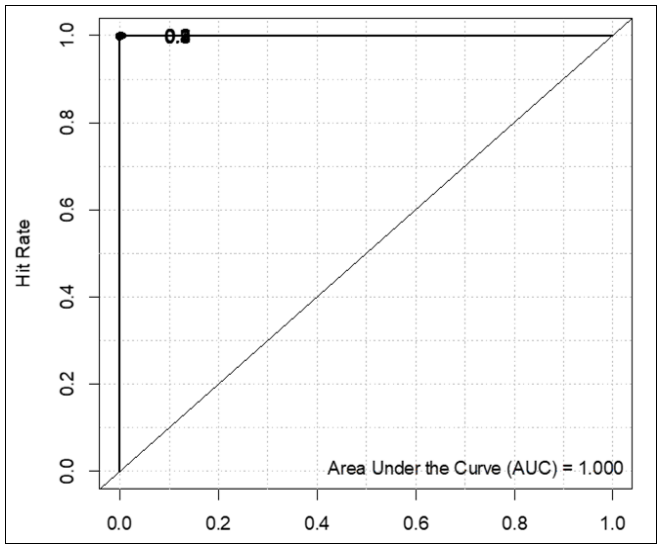

Fig 4: AU Curve

Table 2: Confusion matrix generated by RF

	BTC	LTC	Class Error
BTC	2107	0	0.0000000000
LTC	2	2078	0.0009615385

Table 3: Variable Importance

	BTC	LTC	Mean-Decrease-Accuracy	Mean-Decrease-Gini
Volume	34.14	33.06	41.94	87.40
Low	30.03	23.96	33.10	420.56
Open	22.73	23.37	28.87	304.97
Close	18.88	22.11	26.86	249.94
High	15.14	22.53	24.39	177.00
Marketcap	15.27	18.74	21.64	83.39

3.3 Boosting Model

This model lift to connect a load with every perception in the dataset and the loads are expanded assuming a model inaccurately orders the perception. Further, the subsequent series of DT structure is a troupe model. The Adaptive choice conveys the customary versatile supporting calculation as executed in the Ada-boost bundle. The Extreme choice conveys the outrageous inclination supporting calculation to assemble an angle helping model which gives an ideal way to deal with helping. The execution of the XG-boost bundle is utilized. At first, it works out with 50 emphases for train information and moreover it has 32 last cycles with log misfortune 0. Table 4 addresses the significance/recurrence of factors really utilized and Table 5 gives the disarray framework produced by support model. It just makes sense of the order of bitcoin and litcoin.

Table 4: Summary of the Extreme Boost model

S. No.	Feature	Gain	Cover	Frequency
1	Low	0.8260189799	0.5688322363	0.32098765
2	Volume	0.0972632140	0.2304582755	0.38888889
3	High	0.0751891561	0.1954245717	0.23456790
4	Close	0.0007714225	0.0038331709	0.03086420
5	Open	0.0005434326	0.0009263843	0.00617284
6:	Marketcap	0.0002137949	0.0005253613	0.01851852

Table 5: Final Confusion Matrix generated by Boost model

True value	BTC	LTC
BTC	2107	0
LTC	2	2080

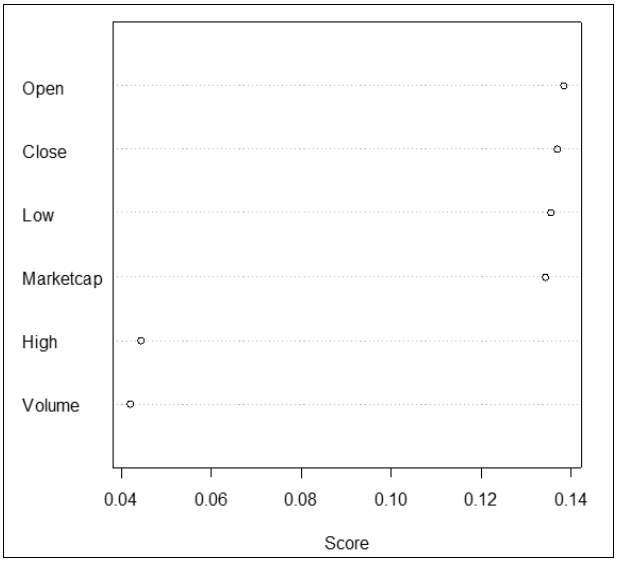

Fig 5: Variable importance plot

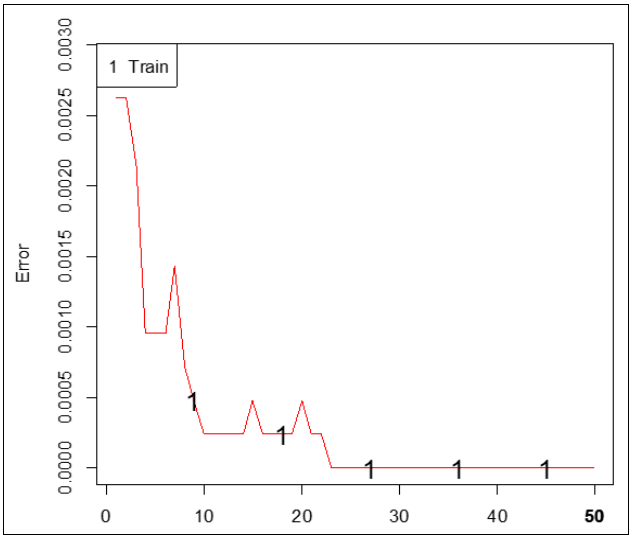

Fig 6: Training Error

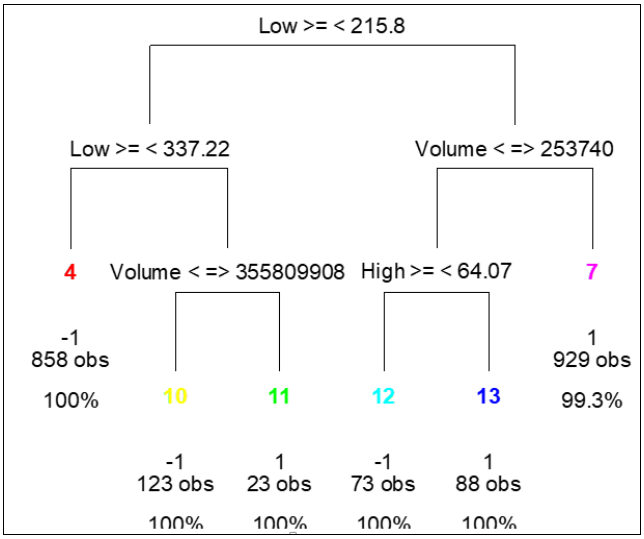

Fig 7: Extreme Boost model Tree

Here, in the wake of carrying out the boosting model, Fig. 5 investigates factors significance. It basic featured that, opening cost, shutting cost, low cost and market capital plays significant on bitcoin or litcoin. Fig. 6 addresses the preparation mistake happened in help model. Fig. 7 investigates the representation of outrageous lift model and it gives tree structure. Further it is seen that, low cost, excessive cost and volume assume the huge part in the grouping.

3.4 Support Vector Machine Model

A Support Vector Machine (SVM) looks for alleged help vectors which are information focuses on information that are found to lie at the edge of an area in space which is a limit starting with one class of focuses then onto the next. In the phrasing of SVM we discuss the space between areas containing pieces of information in various classes just like the edge between those classes. The help vectors are utilized to distinguish a hyperplane that isolates the classes.

Table 6: Summary results of various functions used in Support vector machine

SVM object of class	SV type	Parameter: cost	Function	Hyperparameter: sigma =	Number of Support Vectors	Objective Function Value	Training error	Time taken:
ksvm	C-svc	C = 1	Gaussian Radial Basis kernel function	576.02233359259	1063	-80.4311	0	2.55 secs
			Polynomial kernel function	degree = 1, scale = 1 offset = 1	1012	-756.54	0.033198	0.60 secs
			Linear (vanilla) kernel function		1012	-756.5399	0.033198	0.36 secs

			Hyperbolic Tangent kernel function	scale = 1 offset = 1	2126	-2259.546	0.252448	3.63 secs
			Laplace kernel function	Sigma = 576.02233359259	3004	-760.5265	0	10.21 secs
			Bessel kernel function	sigma = 1 order = 1 degree = 1	1580	-1195.671	0.086697	8.24 secs
			Anova RBF kernel function	sigma = 1 degree = 1	810	-590.6928	0.030093	7.63 secs
			Spline kernel function.		951	-718.2728	0.119417	12.18 secs

Table 6 summed up the outcomes applied on the dataset. SVM apparatus used to examine and performed correlation with the different capacities. It essentially shows the Gaussian Radial Basis portion work gives the most un-preparing mistake, least figuring time, number of help vectors and the base goal work cost esteem.

3.5 Linear and Generalised Linear Models

A linear regression model is the conventional technique for fitting a factual model to information. It is suitable when the objective variable is numeric and nonstop. The group of summed up straight models stretches out conventional direct relapse to focuses with non-typical (non-gaussian) circulations. Direct relapse models are iteratively fit to the information subsequent to changing the objective variable to a nonstop numeric. Table 8 shows the synopsis of the summed up straight model and the particular coefficients of the factors.

Table 7: Deviance Residuals

Min	1Q	Median	3Q	Max
-0.00103732	-0.00000002	-0.00000002	0.00000002	0.00010542

Table 8: Coefficient's summary

Coefficients	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.947e+01	8.512e+02	0.023	0.982
High	2.246e-03	3.255e+01	0.000	1.000
Low	2.619e-02	7.374e+01	0.000	1.000
Open	-1.285e-02	3.919e+01	0.000	1.000
Close	- 1.237e+00	7.625e+01	-0.016	0.987
Volume	2.234e-09	3.978e-07	0.006	0.996
Marketcap	6.497e-08	2.132e-06	0.030	0.976

Table 9: Analysis of Deviance Table (Model: binomial, link: logit)

Source	Df	Deviance Resid.	Df	Resid. Dev	Pr(>Chi)
NULL		4186	5804		
High	1	4596	4185	1208	<2e-16 ***
Low	1	173	4184	1035	<2e-16 ***
Open	1	0	4183	92488	1
Close	1	91458	4182	1030	<2e-16 ***
Volume	1	0	4181	46208	1
Marketcap	1	46208	4180	0	<2e-16 ***

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 9 investigates ANOVA with binomial model having the connection with logit work. It is seen that, opening cost and volume establishes to be insignificant.

3.6 Neural Network Model

Assemble a model that depends on the possibility of numerous layers of neurons associated with one another, taking care of the numeric information through the organization, joining the numbers, to create a last response. Outline of the Neural Net model (fabricated utilizing nnet):

Table 10: Summary of the Neural Net model

Network	Weight	Inputs	Output	Sum of Squares Residuals (SSR)
6-6-1	55	High Low Open Close Volume Market capital	Symbol	2125.00
6-12-1	103			2535.00
6-18-1	151			2255.00
6-24-1	199			2073.00
6-30-1	247			2071.00
6-36-1	295			2080.00
6-42-1	343			2080.00
6-48-1	391			2080.00

Table 10 summarises the various network models assuming different weights. After, comparison we observed that, ANN model 6-30-1 gives least SSR. Hence, this model gives more efficiency which contains least SSR.

3.7 Error Matrix

An error matrix shows the true outcomes against the predicted outcomes. Two tables will be presented here. The first will be the count of observations and the second will be the proportions. For a binary classification model the cells of the error matrix are referred to, from the top left going clockwise, as the True Negatives, False Positives, True Positives, and False Negatives. An error matrix is also known as a confusion matrix.

Table 11: Error matrix

Method		Predicted			Overall error	Averaged class error
	Actual	BTC	LTC	Error		
Decision Tree model [validate]	BTC	451	1	0.2	0.2%	0.2%
	LTC	1	444	0.2		
Extreme Boost model [validate]	BTC	452	0	0.0	0.1%	0.1%
	LTC	1	444	0.2		
Random Forest model [validate]	BTC	452	0	0.0	0.1%	0.1%
	LTC	1	444	0.2		
SVM model [validate]	BTC	361	91	20.1	10.2%	10.05%
	LTC	0	445	0.0		
Linear model validate]	BTC	451	1	0.2	0.1%	0.1%
	LTC	0	445	0.0		
Neural Net model [validate]	BTC	452	0	0	49.6%	50%
Decision Tree model [train]	BTC	2096	11	0.5	0.2%,	0.25%
	LTC	0	2080	0.0		
Extreme Boost model [train]	BTC	2107	0	0	0%	0%
	LTC	0	2080	0		
Random Forest model [train]	BTC	2107	0	0	0%	0%
	LTC	0	2080	0		
SVM model [train]	BTC	1998	109	5.2	12%	12%
	LTC	391	1689	18.8		
Linear model [train]	BTC	2107	0	0	0%	0%
	LTC	2080	0	0		
Neural Net model [train]	BTC	2107	0	0	49.7%	50%
	LTC	2080	0	100		
Decision Tree model [test]	BTC	429	3	0.7	0.4%	0.45%
	LTC	1	465	0.2		
Extreme Boost model [test]	BTC	432	0	0.0	0.1%	0.1%
	LTC	1	465	0.2		
Random Forest model [test]	BTC	432	0	0.0	0.1%	0.1%
	LTC	1	465	0.2		
SVM model [test]	BTC	408	24	5.6	17.2%	16.75%
	LTC	130	336	27.9		
Linear model [test]	BTC	432	0	0	0%	0%
	LTC	0	466	0		
Neural Net model [test]	BTC	432	0	0.0	51.7%	49.8%
	LTC	464	2	99.6		
Decision Tree model (counts):	BTC	2976	15	0.5	0.3%,	0.3%
	LTC	2	2989	0.1		
Extreme Boost model (counts)	BTC	2991	0	0.0	0%	0.05%
	LTC	2	2989	0.1		
Random Forest model (counts):	BTC	2991	0	0.0	0%	0.05%
	LTC	2	2989	0.1		
SVM model (counts)	BTC	2807	184	6.2	16%	16%
	LTC	771	2220	25.8		
Linear model (counts)	BTC	2990	1	0	0%	0%
	LTC	0	2991	0		
Neural Net model (counts)	BTC	2991	0	0.00	50%	49.95%
	LTC	2989	2	99.9		

Table 11 explains the comparative results of the various methods applied on our study data set. Our data is spiltted into train, test and validity as well as full dataset form. The classification results are stored in the form of error matrix. The results obtained reveal the information that, random forest, extreme boost model and linear model applied on the data model gives more preferred results. In this study, linear model gives more efficient results.

3.8 ROC Curve

A receiver operator characteristic (ROC) curve compares the false positive rate to the true positive rate. We can access the trade off the number of observations that are incorrectly classified as positives against the number of observations that are correctly classified as positives. Table 8-11 represents the respective precision against recall for full, train, validity and test dataset for the various models applied the data.

Table 12-15 represents the respective sensitivity against specificity for full, train, validity and test dataset for the various models applied the data.

Table 12: AUROC curve

Model	Efficiency of dataset			
	validate	train	test	Full
rpart model	0.9979	0.9977	0.9958	0.9974
ada model	1	1	1	1
rf model	1	1	1	1
ksvm model	0.9814	0.8906	0.9473	0.848
glm model	0.9984	1	1	0.9998
nnet model	0.5	0.5	0.5021	0.5003

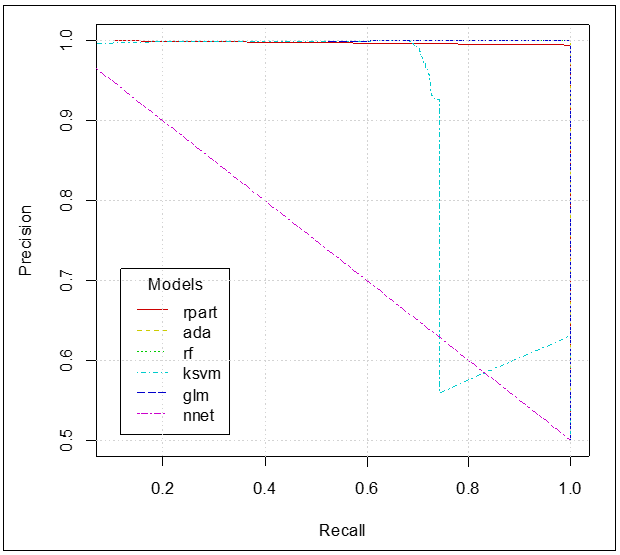

Fig 8: Full dataset

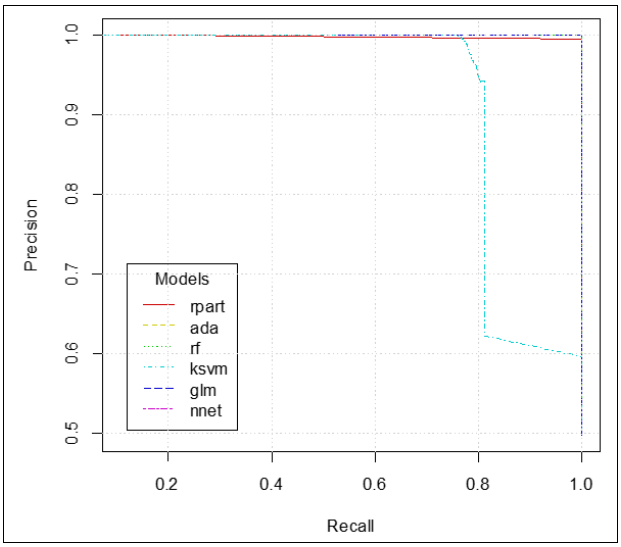

Fig 9: Train dataset

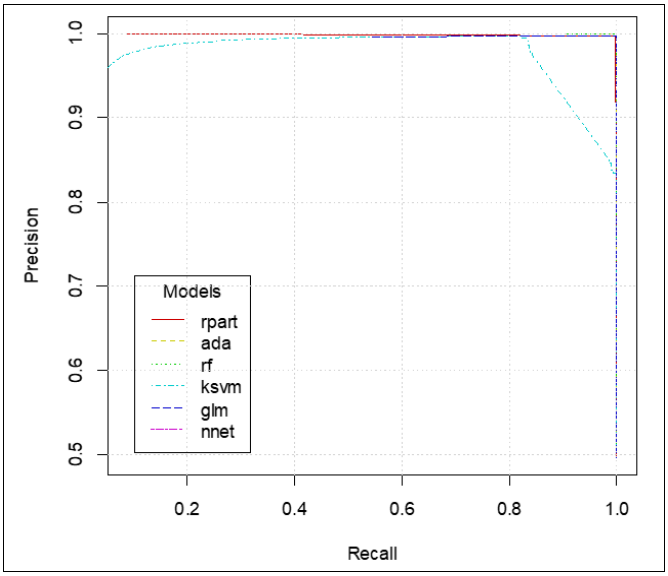

Fig 10: Validate dataset

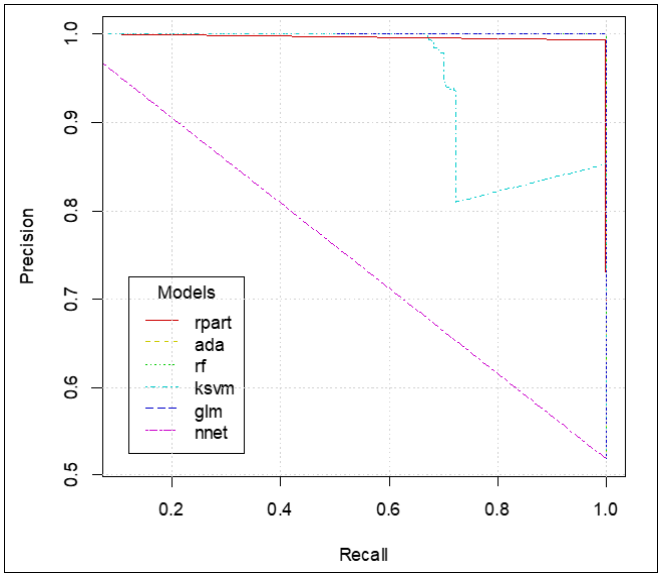

Fig 11: Test dataset

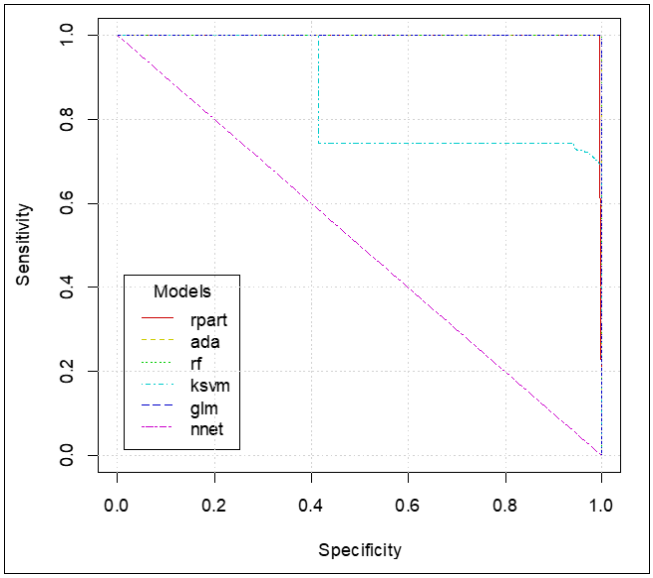

Fig 12: Full dataset

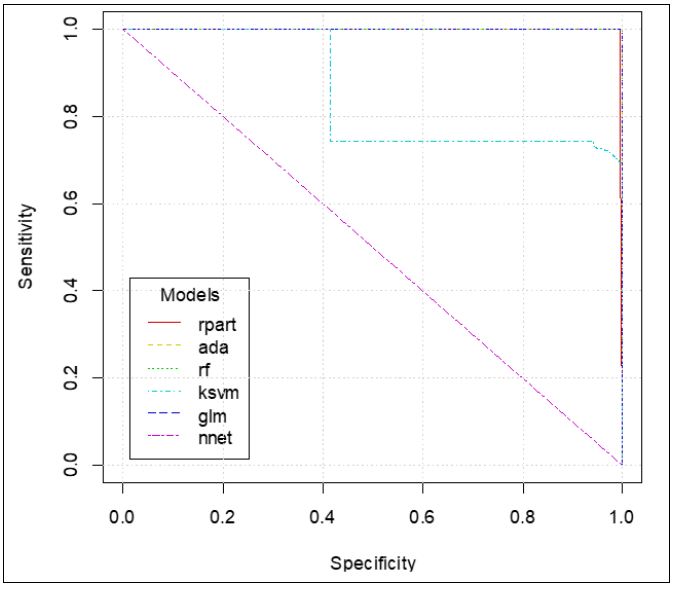

Fig 13: Train dataset

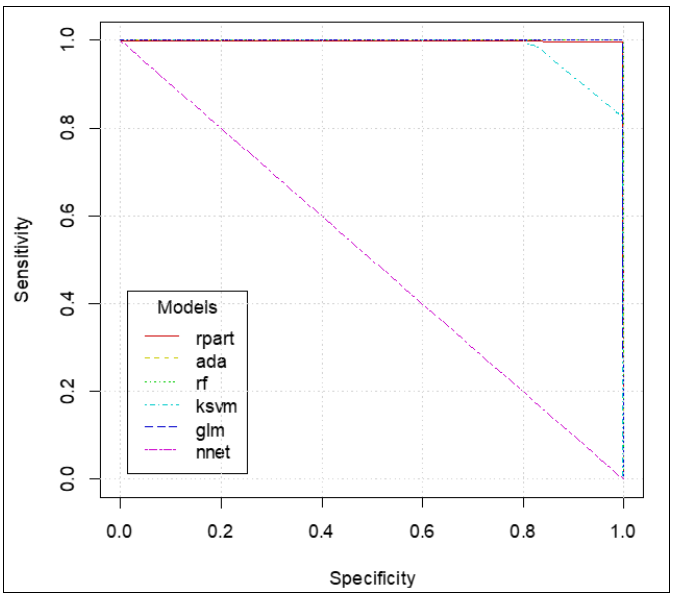

Fig 14: Validate dataset

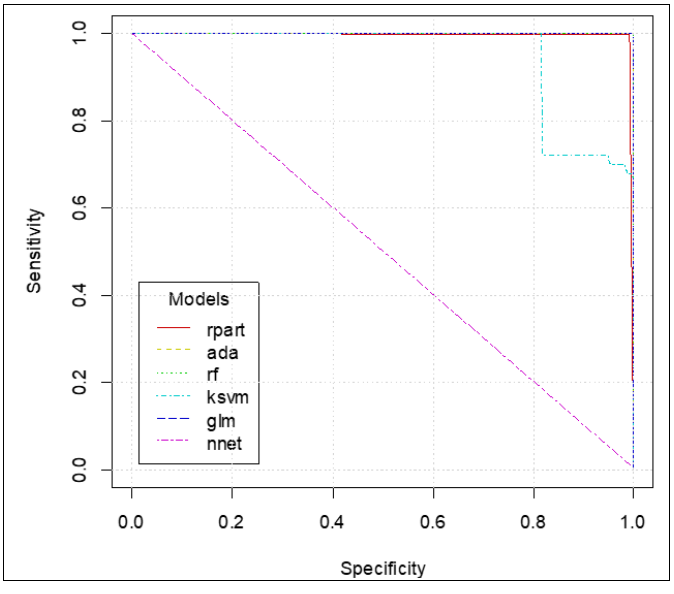

Fig 15: Test dataset

Conclusions

In this study, we have exceptionally centred on the Bitcoin and Litcoin Crypto currency informational index with different boundaries. It is seen that, it is irrelevant to open cost and volume. We have applied different Machine learning methods viz., DT, RF, Extreme Boost model, SVM model, Linear and Generalized Linear Models, NN model, OC curve. Overall, dealing with the large scale dataset, the results obtained through the various machine learning techniques we found that, low price, high price, volume of the Crypto currency plays significant role in classification. By and large, in view of the execution different streamlining devices we are come towards resolution that, general linear model is more akin and it gives a lot of productive outcome.

Acknowledgements: Author is thankful to the anonymous reviewers and the editor for their valuable suggestions to improve the quality of the paper.

Author contribution: Author has conceived the idea, secondary data collection, performed analysis of the data, and written the manuscript.

Funding: This study has received no funding.

Data availability statement: The secondary data will be shared if requested.

Declarations:

Conflict of interest: There is no conflict of interest.

Ethical approval: There is no ethical conflict.

Consent for publication: Author approves the consent to publish this paper.

References

1. Nakamoto S. Bitcoin: A peer-to-peer electronic cash system. Decentralized Business Review, 2008, 21260.
2. Hamayel MJ, Owda AY. A Novel Cryptocurrency Price Prediction Model Using GRU, LSTM and bi-LSTM Machine Learning Algorithms. AI. 2021; 2(4):477-496.
3. Tanwar S, Patel NP, Patel SN, Patel JR, Sharma G, Davidson IE. Deep Learning-Based Cryptocurrency Price Prediction Scheme with Inter-Dependent Relations. IEEE Access. 2021; 9:138633-138646.
4. Marne S, Correia D, Churi S, Gomes J. Predicting Price of Cryptocurrency-A Deep Learning Approach. NTASU. 2020; 9(3).
5. Fatah H, Anggraini RA, Supriadi D, Pertiwi MW, Warnilah AI, Ichsan N. Data mining for cryptocurrencies price prediction. In Journal of Physics: Conference Series. 2020; 1641(1):012059. IOP Publishing.
6. Fang F, Ventre C, Basios M, Kanthan L, Martinez-Rego D, Wu F, Li L. Cryptocurrency trading: a comprehensive survey. Financial Innovation. 2022; 8(1):1-59.
7. Liu Y, Tsyvinski A. Risks and returns of cryptocurrency. The Review of Financial Studies. 2021; 34(6):2689-2727.
8. https://www.kaggle.com/kajalkuchhadiya/crypto-currency-dataset/metadata