Performance of machine learning methods in predicting trend in price and trading volume of cryptocurrencies

Xuanjie Zhang

A Thesis in

The Department

of

Supply Chain and Business Technology Management

Presented in Partial Fulfillment of the Requirements

for the Degree of Master of Supply Chain

Management at

Concordia University

Montreal, Quebec, Canada

August 2023

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: Xuanjie Zhang

Entitled: Performance of machine learning methods in predicting trend in price and trading volume of cryptocurrencies

and submitted in partial fulfillment of the requirements for the degree of

Master of Supply Chain Management

complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the final Examining Committee:

Dr. Satyaveer S. Chauhan Chair

Chair's name

Dr. Danielle Morin Examiner

Examiner's name

Dr. Chaher Alzaman Examiner

Examiner's name

Dr. Navneet Vidyarthi Supervisor

Supervisor's name

Dr. Salim Lahmiri Supervisor

Supervisor's name

Approved by _____

Dr. Satyaveer S. Chauhan Chair of Department or Graduate Program Director

August 11th 2023

Dr. Anne-Marie Croteau Dean of Faculty

ABSTRACT

Performance of machine learning methods in predicting trend in price and trading volume of cryptocurrencies

Xuanjie Zhang

This study is motivated by the growing interest in cryptocurrency trading and the need for accurate forecasting tools to guide investment decisions. The main aim is to forecast price and trading volume changes of cryptocurrencies by determining their movement directions. Naïve Bayes, support vector machines, logistic regression, regression trees, and the K-nearest neighbors' algorithm are selected to solve the problem and compared. Performance measures such as accuracy, sensitivity, and specificity are used to assess the models. The study shows that some models are better at predicting volume trends than price trends in cryptocurrencies. Naïve Bayes is good at spotting positive trends, while Logistic Regression is accurate at identifying negative trends. Interestingly, the research reveals that shorter prediction times are more accurate for price forecasts, but intermediate times work better for specificity. These insights help us understand which models work well for different aspects of cryptocurrency forecasting.

Acknowledgements

I would like to express my deepest appreciation to my supervisor Dr. Salim Lahmiri. And I'm extremely grateful to my supervisor Dr. Navneet Vidyarth. I could not have undertaken this journey without their help. I am also thankful to the advisor Dr. Satyaveer Chauhan for the support. Sincere thanks should also go to Dr. Danielle Morin and Dr. Chaher Alzaman as reviewers for my thesis. Finally, I am extremely grateful to my parents for taking care of me during my injury and supporting me in the work. Thank you all.

Table of Contents

List of Figures

List of Tables

Chapter 1
Introduction

A cryptocurrency is a type of virtual currency system that works similarly to traditional currencies, it allows users to make virtual payments for goods and services but without the involvement of a centrally controlled, dependable authority (Farell 2015). The greatest achievement of cryptocurrency, and its technological dependency, is a peer-to-peer digital trading system that produces and distributes currency units. This system relies on cryptographic evidence rather than authority (Nakamoto, 2008).

One of the most notable examples of cryptocurrency is Bitcoin, which exemplifies the power and potential of this virtual currency system. Bitcoin is the first known cryptocurrency; it is published anonymously by Nakamoto in 2008. For the first two years, it attracted no attention and traded for less than a dollar. But to this day, it is the most popular and most traded cryptocurrency globally, with a peak trading price of $64,978 (November 2021), dominates around 38% of the overall market.

As the popularity of cryptocurrency trading continues to soar, the number of newly exploited coins has witnessed a significant rise, the exploitation of new coins has been rising year by year (see Figure 1). There are 21,844 cryptocurrencies in existence as of November 2022, yet not all of them are useful or active. There are about 9,314 active cryptocurrencies once numerous "dead" cryptocurrencies are excluded (coinmarketcap.com).

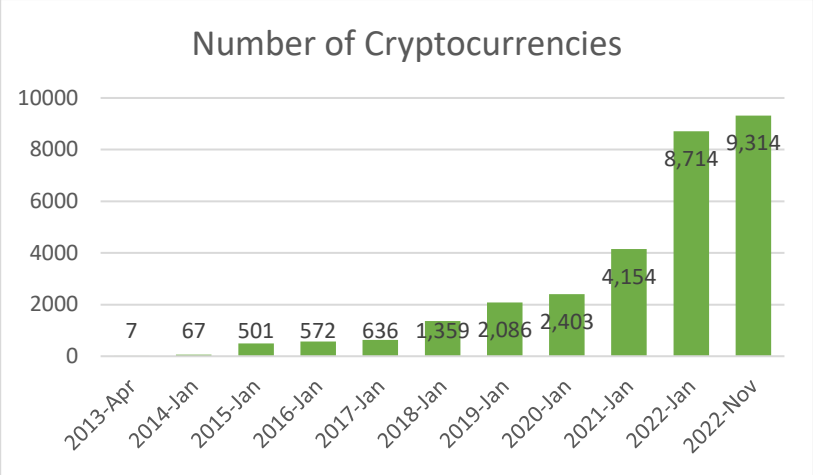


Figure 1 Number of Cryptocurrencies in existence as of November 2022 (CoinMarketCap, n.d.)

Blockchain is Bitcoin's greatest novel invention. Bitcoin is essentially "a chain of digital signatures" as they can only be transferred by digitally signing transactions on the public ledger. Therefore, cryptocurrencies are seen as devoid of any inherent value due to their innate nature; instead, their price is solely determined by supply and demand (Andreessen, 2014). Building upon Bitcoin's innovative blockchain technology, Cheah and Fry (2015) discovered that there is a significant massive bubble component in Bitcoin pricing, compared to the gold or foreign exchange markets, it is more volatile (Dwyer, 2015).

The market structure of cryptocurrencies is fundamentally different, even though they share many characteristics with conventional financial markets like foreign currency (Dyhrberg et al., 2018). Bitcoin, in addition to being a means of electronic exchange, also serves as an investment for speculation asset, with trading available every day of the week without oversight. According to Yermack (2015), most Bitcoin transactions are between investors who are speculative. Because just a small proportion of Bitcoin transactions are used to purchase products and services, it appears that Bitcoin behaves more like an unstable investment than a currency.

As the price of cryptocurrencies has risen year after year, enthusiasm for investing has continued to rise despite the volatility. In recent years, the realm of data analytics has transcended traditional industries and found its application in novel domains, including the dynamic landscape of supply chain management (Van Nguyen et al., 2018). The burgeoning

interest in cryptocurrency trading has mirrored the momentum of supply chain management's transformation through data analytics. In both areas, accurate predictions and informed decision-making play a pivotal role. Just as supply chain analytics leverages data-driven forecasts to optimize inventory, streamline logistics, and enhance demand planning, the cryptocurrency market necessitates precise forecasting tools to guide investment decisions and navigate the volatile landscape (Gunasekaran et al., 2016).

Recent research has shown the increasing relevance of cryptocurrency and blockchain technology and cryptocurrency in reshaping supply chain management across various industries. Blockchain technology possesses the capacity to revolutionize current supply chain arrangements. Its worth in supply chain management can be categorized into four key aspects: expanded transparency and traceability, the digitization and reduction of intermediaries within the supply chain, heightened data security, and the implementation of smart contracts (Wang et al., 2019). Koirala et al. (2019) proposed a blockchain-enabled model with traceability and ownership management, highlighting the transformative potential of smart contracts in supply chain operations. Choi (2020) delved into the impacts of agents' risk attitudes towards cryptocurrency on creating win-win scenarios in supply chains. In the construction and engineering sector, Hamledari and Fischer (2021) have demonstrated the application of blockchain-based crypto assets to seamlessly integrate physical and financial supply chains. Subramanian et al. (2021) introduced the concept of a "Crypto Pharmacy" utilizing a mobile application integrated with hybrid blockchain to address issues in pharmaceutical supply chains.

Taken together, in this environment, forecasting the price and volume of cryptocurrencies trading will become quite challenging and fascinating. This study addresses the challenge of predicting price and trading volume changes in cryptocurrencies, analogous to the supply chain's requirement for anticipating demand patterns and optimizing resource allocation.

One notable analogy lies in the prediction of trading volume, a critical factor in both cryptocurrency markets and supply chains. In the realm of cryptocurrencies, anticipating trading volume between buyers and sellers is essential for effective market participation and strategy formulation. Similarly, supply chain managers must forecast demand and supply

volumes to optimize inventory levels, minimize costs, and ensure seamless operations (Dubey et al., 2015).

In essence, this study not only delves into the predictive prowess of machine learning techniques for cryptocurrency trading but also uncovers a fascinating parallel to the supply chain management domain. As both realms embrace data analytics to navigate complex patterns, the interplay between predicting cryptocurrency trends and optimizing supply chain flows underscores the transformative power of data-driven decision-making across diverse industries.

The aim of our work is forecasting both price and trading volume changes of cryptocurrencies by determining the directions of movements. To the best of our knowledge, very limited recent publications have carefully examined the volume changes of cryptocurrencies. In addition, most of the price prediction studies concentrate on the most popular cryptocurrencies, but our work extends the scope to 20 cryptocurrencies, investigated the broader applicability of the models. In selecting the models of study, we have likewise taken a broad scope to contribute a comprehensive study of the role of machine learning on cryptocurrency prediction, they are naïve Bayes (NB) (Russell & Norvig, 1995), support vector machines (SVM) (Vapnik, 1995), logistic regression (Allison, 2012), regression trees (RT) (Breiman, 1984), K-nearest neighbors (k-NN) algorithm (Cover and Hart, 1967). We measure the performance of each model by using accuracy (correct classification rate), sensitivity and specificity. We apply statistical tests to check differences of performance measures across models.

The contributions of this thesis can be summarized as follows:

1. Utilizing machine learning algorithms to make informed cryptocurrency investment decisions, highlighting the importance of volume forecasting.
2. Investigating a broad range of 20 highly traded cryptocurrencies expands beyond popular cryptocurrencies. Examining the robustness of the results and to draw general conclusions.

3. Examining the effectiveness of number of days tracing back as predictors on performance of machine learning algorithms.

The remainder of the thesis is organized as follows: chapter 2 presents a comprehensive Literature Review, delving into the evolving landscape of predicting cryptocurrency trends using machine learning methods. Chapter 3 technically describes selected machine learning models for forecasting and performance measures of accuracy, specificity, and sensitivity. Chapter 4 introduces our data and provides forecasting results, as well as comparisons of different models' performance and their strengths or weakness. Chapter 5 concludes our main findings and discusses future research directions.

Chapter 2
Literature Review

There has been a lot of research done on forecasting cryptocurrency prices, where academics look for patterns or simulations in uncertainty to support decision making. Review by Khedr et al. (2021) points out that predicting cryptocurrencies using traditional statistical methods, machine Learning and deep Learning are increasing every year with a total of 87 papers between 2010 and 2020. Among them, machine learning and deep learning algorithms are widely used techniques for prediction. Lahmiri and Bekiros (2019) are among the first to use a deep learning approach to predict cryptocurrency prices. Their test results revealed that, while the overall computational cost of the long short term memory (LSTM) model is higher in nonlinear pattern recognition than brute force, the predictability of LSTM is much higher. Lahmiri and Bekiros (2021) further implemented and use deep feed-forward neural network (DFFNN) for high-frequency Bitcoin price analysis and forecasting. They investigated how common numerical training procedures affect the accuracy obtained by DFFNN and found the Levenberg-Marquardt algorithm-trained DFFNN is a powerful and simple tool for forecasting high-frequency price data for Bitcoin.

Wu et al. (2018) introduced a novel forecasting framework using LSTM networks for predicting daily Bitcoin prices. Their findings showed that the LSTM model with Auto Regressive (AR) integration outperformed the standard LSTM model, demonstrating superior forecasting accuracy. This was evident in the reduced prediction errors, including a decrease of 4574.12 in MSE (Mean Square Error), 9.08 in RMSE (Root Mean Square Error),

9.75 in MAPE (Mean Absolute Percentage Error), and 0.1 in MAE (Mean Absolute Error). Poongodi et al. (2020) compared linear regression (LR) and support vector machine (SVM) in predicting Ethereum price and concludes that SVP has a higher accuracy (96.06%). Similarly, Yiying and Yeze (2019) focused on analyzing and comparing the performance and efficiency of LSTM and ANN (Artificial neural network) on predicting price dynamics of Bitcoin, Etherum, and Ripple. They concluded that LSTM is more effective than ANN at using information that is hidden in historical memory, as LSTM utilizes shorter term dynamics (1, 3, 5 days) more. Shintate and Pichl (2019) conducted a study that contrasts LSTMs and MLPs (Multilayer perceptron) using 1-minute interval time-series data from Bitcoin and Litecoin exchanges. Additionally, they introduce their algorithm, the Random Sampling Method (RSM), inspired by advancements in deep learning from image processing. Notably, RSM achieves the highest accuracy (0.5353) in comparison to LSTMs (0.4688) and MLPs (0.4766) in their investigation.

Patel et al. (2020) proposed a hybrid prediction scheme of LSTM and GRU and tested its accuracy, which proved the hybrid scheme is better than LSTM. They focused on only two cryptocurrencies, namely Litecoin and Monero. The proposed hybrid model reduced the mean square error from 194.50 to 5.28 for Litecoin forecast, and from 230.93 to 10.10.7 for Monero price forecast. Jay et al. (2020) trained a stochastic neural network model which simulates market volatility by inducing layer-wise randomness into the observed feature activations of neural networks. They also compared the proposed model with LSTM and Multi-Layer Perceptron (MLP) for Bitcoin, Ethereum, and Litecoin and show its superiority. A stochastic neural network's average relative improvement over a normal neural network range from 1.56% to 1.76%. Monsalve et al. (2020) studied prediction of six well-known cryptocurrency’s exchange rate- whether their value will increase respect to USD in the next minute. In their four different network architectures, results showed Convolutional LSTM neural networks outperformed all the rest significantly. Adcock and Gradojevic (2019) employed feed-forward neural networks incorporating lagged returns and basic technical trading rules to predict returns for Bitcoin to USD. The study indicates the suitability of these architectures for Bitcoin return forecasting, noting that outcomes may fluctuate over time due to the influence of its rapid price fluctuations. Further, Chowdhury et al. (2020) expanded

the range of studied cryptocurrencies to nine, and considered four different models: neural net model, k-NN model, gradient boosted trees model and ensemble learning method, among which the ensemble learning method obtained 92.4% accuracy and considered to be the best. Jaquart et al. (2022) employed six machine learning algorithms to determine the binary relative daily performance of the top 100 cryptocurrencies by market capitalization, all models employed made statistically feasible predictions. The study presented evidence of statistical arbitrage opportunities in the bitcoin market.

In addition to predicting cryptocurrency prices, there are a few studies that concentrate on volume forecasting as well. To anticipate final trading volume, Lahmiri et al. (2020) offered an artificial neural networks ensemble forecasting model that combines radial basis function neural networks (RBFNN) and generalized regression neural networks (GRNN) together with feedforward artificial neural network (FFNN) to generate final trading volume prediction. The results showed that the ensemble prediction model can achieve improvement in volume prediction by significantly reducing errors compared to a single model. Lahmiri et al. (2022) explored the impact of kernel selection on the support vector regression's (SVR) capacity to predict cryptocurrency trade volume, as well as when the SVR's critical parameters are tuned by the Bayesian optimization (BO) method, referred as SVR-BO. The results of 180 trials showed that the SVR-BO with RBF kernel (RMSE= 0.2111 ± 0.1504) outperforms all other models (RMSE ≥0.2177±0.1546) when used to forecast trading volume for the following day, while the SVR-BO using polynomial kernel (RMSE= 0.1045±0.1001) outperforms all other models (RMSE ≥0.1101±0.1320) when used to forecast trading volume for the following week.

There are also studies that included other indicators in the predictions and examine the predictive power of these indicators. Nakano et al. (2018) applied an Artificial Neural Network (ANN) to forecast the price direction of Bitcoin at 15-minute intervals, utilizing both price data and technical indicators. Their intraday approach yields superior returns compared to buy-and-hold strategies and other basic technical trading methods, results of investment show under buy-and-hold strategy, final value equals to 2.28, while with ANN prediction, final value ranges from 12.14 to 64.46. Miura et al. (2019) used 3 hours returns for predicting aggregating realized volatility (RV) of continuously traded cryptocurrencies,

specifically Bitcoin. Machine learning techniques, including ANN (MLP, GRU, LSTM), SVM, and Ridge Regression, are employed to predict future RV values based on past samples. By measuring the MSE of each model, they found that Ridge Regression performs best, aligning with the auto-regressive dynamics of the Heterogeneous Auto-Regressive Realized Volatility (HARRV) model. Neural networks closely follow in performance, while SVM fares worst. Poongodi et al. (2021) made use of communication data in online bitcoin community bitcointalk.org to build a deep learning model and examined its connection with global bitcoin price trends, concluded that social media trends can be a strong indicator of cryptocurrency trends. In the study of Ortu et al. (2022) on the predictability of price movements, three indicators of technical, trading, and social are considered as inputs of the classification algorithm. Study showed that the unrestricted model that includes all three features produces a substantial improvement in the prediction and accuracy than the restricted model composed of technical feature only. Wang et al. (2022) selected 12 major cryptocurrencies and analyzed whether informed trading makes cryptocurrency returns more predictable. They suggested that informed trading helps anticipate some of cryptocurrencies returns, but it is unable to considerably increase prediction accuracy on a market-wide average. Dag et al. (2023) provided a data-driven Tree Augmented Naive (TAN) Bayes approach for determining the most crucial factors affecting Bitcoin price changes. Study showed making short-term investment decisions using the suggested methods is feasible without losing accuracy.

The summary table below summarizes the literature on using machine learning and deep learning to predict trends related to cryptocurrencies. These studies mainly focused on the prediction of final price levels, with limited attention given to forecasting binary price trends, which play a pivotal role in shaping trading decisions within the dynamic market landscape. Recognizing the critical influence of price fluctuations on trading outcomes, it becomes imperative to delve into the intricate realm of price trend prediction.

Furthermore, noted is the limited scope of earlier cryptocurrency prediction studies to specific currencies, restricting their broader market applicability. Additionally, studies analyzing alternative indicators often overlooked the importance of trading volume—a key metric reflecting market demand and liquidity.

Table 1 Summary of Literature Review of Cryptocurrencies Forecast Using Machine Learning Models

Study	Market	Inputs	Model	Performance
Lahmiri and Bekiros (2019)	Bitcoin, Digital Cash, Ripple	Price	LSTM, DLNN, GRNN	RMSE
Lahmiri and Bekiros (2021)	Bitcoin	Price	DFFNN	RMSE
Wu et al. (2019)	Bitcoin	Price	LSTM	RMSE, MSE, MAE, MAPE
Poongodi et al. (2020)	Ethereum	Price	Linear regression, SVM	Accuracy
Yiying and Yeze (2019)	Bitcoin, Ethereum, Ripple	Price	LSTM, ANN	MSE
Shintate and Pichl (2019)	Bitcoin, Litecoin	Price	LSTM, MLP, REM	Accuracy, Sensitivity, Precision, F1 Score
Patel et al. (2020)	Litecoin, Monero	Price	LSTM and GRU hybrid	MSE, RMSE, MAE, MAPE
Jay et al. (2020)	Bitcoin, Ethereum, Litecoin	Price	Stochastic neural network, LSTM and MLP	MAPE
Adcock and Gradojevic (2019)	Bitcoin	Price	Feedforward ANN	Accuracy
Chowdhury et al. (2020)	Bitcoin Cash, Bitcoin, Dash, DOGE, Ethereum, IOTA, Litecoin, NEM, NEO	Price	Neural networks, k-NN, GBT, ensemble learning	Accuracy, RMSE
Monsalve et al. (2020)	Bitcoin, Dash, Ether, Litecoin, Monero, Ripple	Price	CNN, CLSTM, MLP, RBFNN	Accuracy, Sensitivity
Lahmiri et al. (2020)	Bitcoin	Trading volume	ANN Ensemble model of RBFNN, GRNNM FFNN	RMSE
Lahmiri et al. (2022)	30 cryptocurrencies	Trading volume	SVR-BO with different kernels	RMSE, MAE
Nakano et al. (2018)	Bitcoin	Price and technical indicators	ANN	Risk-return
Miura et al. (2019)	Bitcoin	Realized volatility (RV)	ANN (MLP, GRU, LSTM), SVM, Ridge Regression, HARRV	MSE
Poongodi et al. (2021)	Bitcoin	Social media trend	LDA	RMSE
Ortu et al. (2022)	Bitcoin, Ethereum	Technical, trading, social indicators	MLP, CNN, LSTM, ALSTM	Accuracy
Wang et al. (2022)	12 cryptocurrencies	Informed trading	RF, LR, SVM, LASTM, ANN	Accuracy
Dag et al. (2023)	Bitcoin	188 new variables	TAN	Accuracy, Sensitivity, Specificity, AUC
Jaquart et al. (2022)	Top 100 cryptos	Market capitalization	LSTM, GRU, TCN, GBC, RF, LR	Accuracy

This work	20 cryptocurrencies	Price, trading volume	NB, LR, Regression Trees, k-NN, SVM	Accuracy, Sensitivity, Specificity

Addressing these gaps in the existing body of literature, the present study undertakes a comprehensive approach by not only forecasting price trends but also delving into the prediction of trading volume trends. There are previous studies that target 20, 30 and even 100 cryptocurrencies, we also expanded the study's target market to encompass 20 highly traded cryptocurrencies safeguards against bias, ensuring the impartial evaluation of machine learning models across diverse cryptocurrency domains.

Chapter 3
Methods

Selected machine learning models are described in this chapter, including NB, logistic regression, RT, SVM and k-NN. We also use three performance measures: accuracy, sensitivity, and specificity, briefly described below.

3.1. Machine Learning Classifiers

3.1.1. Naïve Bayes (NB)

Naive Bayes is a probabilistic algorithm that employs the conditional independence assumption to calculate the likelihood of a particular class membership based on observed features (Russell & Norvig, 1995). The algorithm models the classes in the training data using probability density functions and assigns objects to the class with the highest likelihood. Given a set of features $(f = f_1, f_2, \dots, f_n)$. It assumes that features are independent of each other, which is why it's called "naïve". Despite its seemingly "naive" assumption of feature independence, Naive Bayes has demonstrated remarkable success in various real-world applications, ranging from text classification and spam filtering to medical diagnosis and image recognition (Langarizadeh & Moghbeli, 2016; Duan et al., 2014).

The NB classifier estimates the most probable target class (c) as the one with the highest posterior probability, which can be computed using Bayes' theorem.

Mathematically, the posterior probability can be expressed as:

$$c = \arg\max\big(Prob(c|f_1, f_2, \cdots, f_n)\big) \quad (1)$$

$$c = \arg\max\left(\frac{Prob(f_1,f_2,\cdots f|c)\cdot Prob(c))}{Prob(f_1,f_2,\cdots f)}\right) \quad (2)$$

Assuming that the features are uniformly distributed and using the chain rule, the most probable target class (c) can be computed as:

$$c = \arg\ \max\ \left(Prob\ (c)\prod_{i=1}^{n} Prob(f_i|c)\right) \quad (3)$$

Here, Prob(c) is estimated by the frequency of c in the training data, and Prob(f_i|c) is estimated by a Gaussian distribution function. The NB classifier is simple to use and requires only one iteration during the learning process to generate probabilities, making it efficient for large datasets.

One key strength of NB is its ability to handle high-dimensional data gracefully. As the number of features increases, Naive Bayes retains its computational efficiency because it independently estimates the distribution of each feature given the class. This characteristic makes it particularly well-suited for tasks involving text data, where the vocabulary can grow significantly. Naive Bayes classifiers also lend themselves well to incremental learning scenarios. As new data becomes available, the algorithm can be easily updated with minimal computational overhead, making it adaptable to changing conditions and evolving datasets.

In practice, there are different variants of NB classifiers, such as Gaussian Naive Bayes, Multinomial Naive Bayes, and Bernoulli Naive Bayes, each tailored to specific types of data and assumptions about feature distributions. Researchers and practitioners have also explored ways to relax the independence assumption through techniques like feature engineering or using more advanced variants like Tree-Augmented Naive Bayes.

Chen et al. (2019) conducted research predicting price changes in Ethereum. They gathered data on Ethereum price fluctuations at around 1-hour intervals spanning from August 30, 2015, to December 2, 2017. They applied six distinct techniques to forecast Ethereum price changes, the NB classifier is one of them and performed 51.78% in accuracy.

3.1.2. Logistic Regression

Logistic regression is a statistical approach that is typically used to model a dichotomous dependent variable, which takes on only two distinct values. The logistic regression model is a sort of generalized linear model that models the connection between the dependent variable and one or more independent variables using the logistic function, often known as the sigmoid function (Allison, 2012). The logistic function transforms the linear combination of independent variables into a probability that the dependent variable is equal to 1. If the probability is higher than or equal to 0.5, the observation is categorized as positive (y=1). Otherwise, it is identified as being of the negative kind (y=0). The logistic regression model can be used to classify observations into different categories based on the probabilities produced by the model.

The logistic regression function can be represented by the following equation:

$$P_i = \frac{1}{1+e^{-(\alpha+\beta_1 x_{i1}+\beta_2 x_{i2}+\cdots+\beta_m x_m)}} \quad (4)$$

Here, P_i is the probability that the dependent variable y_i takes on a value of 1, given each explanatory variable $x_{ij}(j = 1,1,\cdots,m)$, α is the intercept of the model, $\beta_j$is the coefficient of the model.

Within the framework of logistic regression, a logit transformation is employed on the odds, which signifies the ratio between the likelihood of success and the likelihood of failure. This transformation is commonly referred to as the log odds, or alternatively, the natural logarithm of odds. The model's beta parameter, also denoted as the coefficient, is conventionally estimated utilizing the technique of maximum likelihood estimation (MLE). This approach engages in a systematic exploration of diverse beta values across multiple iterations, thereby attaining an optimal alignment with the log odds' optimal fit.

The culmination of these iterations engenders the log likelihood function, which logistic regression endeavors to maximize for the attainment of the most suitable parameter

estimation. Once the pinnacle coefficient (or coefficients in scenarios involving multiple independent variables) is ascertained, the conditional probabilities for each individual observation can be meticulously computed, logged, and subsequently aggregated to yield a projected probability. Following the calculation of the model, ideally it is a good practice to assess how well the model predicts the dependent variable, which is known as the goodness of fit. The Hosmer-Lemeshow test is a common method of assessing the fit of a model.

Logistic regression serves as a versatile tool for predictive and classification tasks with various applications. Notably, it aids in fraud detection by identifying behaviors linked to fraudulent activities, benefiting financial institutions and SaaS companies in data analysis and user account authentication (Şahin & Duman, 2011). In the medical domain, it enables disease prediction, facilitating proactive healthcare interventions (Bender & Grouven, 1998). Furthermore, logistic regression plays a crucial role in churn prediction, guiding actions to retain high-performing employees or clients, thereby enhancing organizational well-being and revenue retention (Nie et al., 2011).

Bouri et al. (2019) analyzed a daily dataset encompassing seven cryptocurrencies (Bitcoin, Ripple, Ethereum, Litecoin, Nem, Dash, and Stellar) spanning from August 7, 2015, to December 31, 2017. Their study employed logistic regression to investigate the probability of price fluctuations in one cryptocurrency because of price changes in other cryptocurrencies. Their findings revealed an interconnectedness, demonstrating that alterations in the price of one cryptocurrency had effects on the prices of others.

3.1.3. Regression Trees

A non-parametric approach to estimating a regression function is the use of regression trees (RTs), often referred to as Classification and Regression Trees (CART) (Breiman, 1984). They divide the learning dataset recursively into two subsets using binary splits until terminal nodes are reached, where homogeneity is attained, and then they determine a set of if-then rules. The main benefit of RTs is that they can handle highly skewed numerical data and categorical inputs by employing ordinal or non-ordinal tree building and do not require assumptions about the distribution of predictors.

The graphical representation of the classification tree produced by RT consists of nodes and branches, where each node denotes a choice about one of the attributes and generates two branches.

The Gini index is employed to reduce impurities in tree construction. The formula for the Gini index G(t) of an impurity in a node t is:

$$G(t) = \sum_{j \neq i} P(j|t)P(i|t) \tag{5}$$

Where i and j are classes of the output, and $P(t)$ refers to the relative frequency of the first class.

For instance, the goodness of the split of a data set D into subset D_1 and D_2 is defined by:

$$G_{split}(D) = \frac{n_1}{n(G(D_1))} + \frac{n_2}{n(G(D_2))} \tag{6}$$

Where n, n_1 and n_2 are the size of D, D_1 and D_2.

Li et al., (2019) examined Twitter signals predictor in forecasting price changes in ZClassic. To build the predictive model, they employed an extreme gradient boosting (GB) regression tree model and assessed its performance against historical price data.

3.1.4. k-NN

Cover and Hart (1967) developed the k-NN method as a nonparametric supervised classifier based on the idea of similarity. It organizes a collection of data points into groups and classifies new data based on a measure of similarity between the new data's attributes and those in the training set. Its key benefit is that it is fully data-driven and does not take the shape of a fitted model.

The k-NN technique locates the k nearest neighbors of I in the sample set based on a distance metric such as Euclidean distance, given a number k and a feature vector to classify I. The new object's class is then determined using a voting procedure such as majority voting or weighted-sum voting based on the classes of its k-nearest neighbors (He & Wang, 2007).

The typical k-NN algorithm is as follows:

1. Determine the Euclidean distances between a new object *o* and all of the items in the learning set.

2. Based on the computed distances, select the k objects from the learning set that are closest to o.

3. Assign o to the group that has the greatest number of the k items.

The formal k-NN classifier algorithm can be expressed as (Ergen et al., 2014):

$$\arg\min\big(d_e(t,o,k)\big) \rightarrow \ identify\ P \tag{7}$$

where k is the number of nearest neighbors to be taken into account, t is the training data, o represents the object to be categorized, P is the assigned class of the new object, and de is the Euclidean distance determined by:

$$d_e(t,o,k) = \sqrt{\sum_{i=1}^{L}(t_{i,k} - o_{i,k})^2} \tag{8}$$

where L represents length of each of data vector.

The k-NN algorithm is computationally efficient and can handle both numerical and categorical data. It can also be used for regression tasks by taking the average of the k-nearest neighbors' output values. However, the performance of k-NN heavily depends on the quality of the distance metric used and the choice of k. The curse of dimensionality can also be a challenge, as the algorithm becomes less effective in high-dimensional spaces. In our study, Euclidean distance metric was used as the distance metric and 5 nearest neighbors were considered ($k = 5$).

Chowdhury et al. (2020) compared four models in predicting the closing prices for CCI30 cryptocurrency components, aiming to contribute to risk mitigation within the cryptocurrency market. In their study, k-NN was used as a comparison model with the ensemble learning method.

3.1.5. SVM

In a high-dimensional space, SVMs build a hyperplane or a series of hyperplanes that can be utilized for classification (Vapnik, 1995). The structural risk minimization principle is used to determine the hyperplane, it aims to minimize the expected risk of the classifier on new and unseen data. By increasing the distance between classes and the hyper-plane, this is achieved. Most importantly, the SVM can avoid local minima and has a better ability to generalize results than other approaches. The linear SVM is a specific type of SVM that assumes the data can be separated by a hyperplane. The linear SVM constructs a hyperplane represented by:

$$y = f(x) = w^T x - b \tag{9}$$

where b is the bias term and w is the weight vector. The goal of the linear SVM is to correctly categorize the training data while locating the weight vector and bias term that maximize the margin between the classes. Once the weight vector and bias term are determined, the linear SVM can be used to classify new data by evaluating the sign of the equation y=f(x).

A kernel function K is used by the SVM classifier to separate nonlinear data. It is said in the following way:

$$f(x_i) = sign(\sum_{i=1}^{n} y_i \alpha_i K\langle x, x_i\rangle + b) \quad (10)$$

where b is a constant, K is a kernel function, and α is the Lagrange multiplier.

In our study, RBF (radial basis function) kernel is used for SVM classifier, given by:

$$K(x, x\_i\ ) = \exp\ (-\delta \|x - x_i\|^2) \quad (11)$$

Let δ represent the scale parameter, defined as $1/\sigma^2$, where σ denotes the width of the radial basis function.

Support Vector Machines (SVM) have garnered notable prominence within the domain of supervised machine learning due to their pronounced efficacy in classification and regression tasks. These algorithms exhibit distinctive advantages, including adeptness in high-dimensional spaces and a robust capacity for generalization, underpinned by their propensity to optimize margins. SVM's resilience to outliers and the facilitation of non-linear modeling via kernel functions are pivotal attributes. However, it is imperative to acknowledge certain disadvantages intrinsic to SVM, encompassing its computational intensity, memory consumption, intricate hyperparameter tuning, and potentially convoluted interpretability, thereby warranting judicious consideration in application selection (L. Wang, 2005).

The purview of SVM's applicability spans diverse arenas, with classification serving as a prominent application context, especially in scenarios involving intricate categorization endeavors such as email filtering, text sentiment analysis, and medical diagnostic endeavors. In image recognition, SVM finds substantial utility in object classification and hand-written digit recognition (Kim et al., 2005). Furthermore, SVM plays a pivotal role in fields like bioinformatics, finance, medicine, remote sensing, natural language processing, and speech recognition, thus accentuating its pertinence across scientific, technological, and sociodemographic contexts (Feng et al., 2005). As SVM engenders an equilibrium between robust classification proficiency and the facilitation of complex data relations, cognizance of its computational intricacies and parameter tuning imperatives augments its judicious deployment within the manifold domains it encapsulates.

To ascertain whether these characteristics had an impact on the values of bitcoin, ripple, or Ethereum, Kim et al. (2016) examined the social activities of cryptocurrency communities. To predict the price changes and the number of transactions, user comments and responses

from these forums were gathered and modelled using ML models such as SVM. The study's findings showed that the strategy accurately anticipated the price volatility of low-cost cryptocurrencies. According to Mallqui and Fernandes (2019), the accuracy rate of bitcoin price prediction can be increased by using a chosen collection of features in conjunction with the best data-mining model. Evaluations were conducted on ML models such ANN, SVM, and ensemble methods (based on RNN and k-means clustering). The suggested model was employed to forecast the highest, minimum, and closing bitcoin prices. In all time intervals, the SVM model with relief approach for attribute selection achieved the greatest and most reliable accuracy rate. Comparing the suggested model to models from earlier research works, it exhibited a 10% increase in prediction accuracy.

3.2. Performances Measures

Three performance measures are used to evaluate the predicting power and effectiveness of the above machine learning classifiers in forecasting cryptocurrencies' price and trading volume changing directions. They are described as follows:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (12)$$

$$Sensitivity = \frac{TP}{TP+FN} \quad (13)$$

$$Specificity = \frac{TN}{TN+FP} \quad (14)$$

where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives in confusion matrix respectively. In our case, positives refer to an upward trend in price or trading volume (output = 1) whilst negatives refer to a downward trend (output = 0).

Accuracy serves as a fundamental performance metric, reflecting the model's ability to make correct predictions across all instances. The overall gap between estimated (or observed) values and the genuine value is what is meant by accuracy (Walther & Moore, 2005). It quantifies the proportion of predictions that match the actual outcomes, offering a straightforward evaluation of overall correctness. While high accuracy is desirable, it may not provide a complete picture, especially in situations where class imbalances or specific goals exist. It is crucial to consider accuracy alongside other metrics to gain a comprehensive understanding of the model's effectiveness.

Barnwal and colleagues (2019) solely employed accuracy as the performance metric to evaluate the Random Forests (RFs) model's effectiveness in predicting cryptocurrency prices, including Bitcoin. Hitam et al., (2019) similarly utilized accuracy as the performance criterion to assess the optimized SVM–PSO model's predictive performance for open and close prices across various cryptocurrencies, including Bitcoin, Litecoin, Ethereum, Ripple, Nem, and Stellar.

Sensitivity, also known as the true positive rate or recall, gauges the model's proficiency in correctly identifying positive instances. It quantifies the percentage of actual positive instances that the model accurately predicts as positive (Parikh et al., 2008). Sensitivity is particularly valuable in applications where missing positive instances could have serious implications. A high sensitivity indicates that the model is adept at capturing the occurrences of the positive trend, showcasing its sensitivity to detecting important events. Mcnally et al. (2018) used sensitivity together with RMSE, accuracy and precision as performance measures to test the performance of LSTM AND RNN in predicting daily price of Bitcoin.

Complementing sensitivity, specificity focuses on the model's performance with negative instances. It measures the proportion of actual negative instances that the model correctly classifies as negative (Trevethan, 2017). Specificity reflects the model's ability to discern and classify instances that do not exhibit the trend or condition of interest. High specificity indicates the model's proficiency in accurately detecting cases where the negative trend is absent, underscoring its capability to avoid false alarms. Madan et al., (2015) assessed the performance of their predictive models using accuracy, sensitivity, and specificity as evaluation metrics. They applied binomial logistic regression, SVM, RF, and binomial GLM to predict Bitcoin price changes at different intervals—daily, every 10 minutes, and every 10 seconds. Similarly, Ji et al., (2019) employed various performance metrics, including accuracy, sensitivity, specificity, and the F1 score, to evaluate their models. These models, which encompassed NN, LSTM, CNN, and a deep residual network, were utilized to predict daily Bitcoin prices.

By considering accuracy, sensitivity, and specificity collectively, we gain a comprehensive insight into the model's strengths and weaknesses. While accuracy provides a general overview of correctness, sensitivity and specificity delve deeper into the model's ability to detect positive and negative instances, respectively. This multifaceted evaluation ensures a more thorough assessment of the model's performance, aiding in making well-informed decisions and optimizations to suit specific application requirements.

Chapter 4
Dataset and Results

4.1 Dataset

The dataset used in this study consists of daily trading prices and volumes for 20 different cryptocurrencies including TRX, BCH, DOGE, ETC, ETH, LINK, USDT, XLM, XMR, XRP, XTZ, BAT, BTG, CVC, DASH, DCR, DGB, ENJ, ERG, and GLM. The data was collected from Yahoo Finance website and the sample period ranges from 09 November 2017 to 24 October 2022. As a result, the sample has 1808 observations. It is imperative to underscore that the selection of these cryptocurrencies for inclusion within the research emanated from a meticulous consideration of liquidity metrics, avoiding reliance upon market capitalization. Indeed, the chosen set constitutes the upper tier of crypto assets in terms of trading activity, proved their dominant position in the trading environment. The careful selection of cryptocurrencies based on liquidity criteria rather than pure market capitalization highlights the rigidity and empirical nature of the analytical approach.

The past cryptocurrencies price data were used as input, the next-day price trends were generated as output (1 if price goes up, 0 if goes down). Same for volume forecasting, past trading volume data were used as input to generate the next day trading volume trend (1 if volume increase, o if decrease). The dataset was then split into a training set and a testing set,

with the training set consisting of the first 80% of the data and the testing set consisting of the remaining 20%.

To investigate the impact of backdating length on prediction effectiveness, the known prices, and volumes for the past 7, 14, 21, and 30 days were utilized as inputs to generate predictions.

We employed the Python software to perform all classification tasks. The prediction process was carried out for each of the 20 cryptocurrencies in the dataset. The use of multiple cryptocurrencies ensures that the performance of the machine learning models is not biased towards any cryptocurrency. Furthermore, the use of multiple prediction runs helps to account for any variability in the results and provides a more robust estimate of the performance of each model. Finally, the results of each cryptocurrency were averaged to obtain the final performance metrics reported in this study.

4.2 Results and Analysis

4.2.1 Average Forecasting Results

Tables 2 and 3 show the predictive ability of different types of machine learning classifier when they are given different backdating lengths (7, 14, 21 and 30 days in the past respectively). For most prediction methods, the accuracy hovers around 0.5. In addition, the sensitivity and specificity obtained by the NB method exhibit a large difference for both price and volume trend prediction, sensitivity up to greater than 0.8 but specificity down to less than 0.2. When predicting volume trend, LR and SVM both obtained high specificity but low sensitivity.

Among the classifiers tested for price trend prediction, the best performances of each classifier in terms of accuracy are as follows: NB achieved 0.5010±0.02 given 14 days price data. LR achieved the highest accuracy of 0.5161±0.02 using 7 days data. K-NN achieved its

highest of 0.5026±0.04 given 7 days data. RT achieved its highest 0.5010±0.04 accuracy also using 7 days data. And SVM achieved 0.5103±0.03 accuracy given 30 days price data.

The best performance by sensitivity achieved by NB is 0.7566±0.09 when using 7 days data, 0.5276±0.34 by LR, 0.4934±0.05 by k-NN, also using 7 days data. RT achieved its highest sensitivity 0.4956±0.04 by using 30 days data. SVM achieved its highest 0.5471±0.30 in sensitivity given 7days price data.

The best performances by specificity are as follows: NB achieved 0.4174±0.25 given 21 days data. LR achieved the highest specificity of 0.5945±0.24 using 21 days data. K-NN achieved its highest of 0.5109±0.06 given 7 days data. RT achieved its highest 0.5148±0.06 specificity also using 7 days data. And SVM achieved 0.5252±0.29 specificity given 21 days price data.

For trading volume trend prediction, the following are the best performances in terms of accuracy: Naive Bayes (NB) achieved 0.5139±0.03 using 14 days price data. LR achieved the highest accuracy of 0.5761±0.03 with 30 days of data. K-Nearest Neighbors (K-NN) reached its peak accuracy of 0.5458±0.03 with a 7-day data window. RT achieved its highest accuracy of 0.5370±0.03 when using 21 days of data. SVM achieved an accuracy of 0.5474±0.03 when considering 7 days of price data for prediction.

For trading volume trend prediction, the best performance by sensitivity achieved by NB is 0.8400±0.17 with 7 days data. Logistic Regression (LR) attained a sensitivity of 0.4267±0.13, using 30 days of data. K-NN achieved a sensitivity of 0.5314±0.05 with a 7-day data window. RT obtained its highest sensitivity of 0.5149±0.04 when using 30 days of data. SVM achieved a sensitivity of 0.3390±0.31 when considering 7 days of price data for prediction.

The best performances by specificity are as follows: Naive Bayes (NB) achieved 0.2789±0.27 using 30 days price data. LR achieved the highest specificity of 0.8190±0.14 with 7 days of data. K-NN reached its highest specificity of 0.5581±0.04 also with 7 days of data. RT obtained its highest specificity of 0.5619±0.03 when using 30 days of data. SVM obtained the highest specificity of 0.8153±0.22 given 21 days of volume data.

Table 2 Summary of average forecasting results for prices

	NB			LR			k-NN		
	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity
7 Days	0.4972±0.02	0.7566±0.09	0.2426±0.07	0.5161±0.02	0.5276±0.34	0.5055±0.33	0.5026±0.04	0.4934±0.05	0.5109±0.06
14 Days	0.5010±0.02	0.7553±0.10	0.2534±0.08	0.5116±0.03	0.5049±0.30	0.5143±0.30	0.4936±0.03	0.4822±0.06	0.5031±0.05
21 Days	0.4888±0.03	0.5575±0.27	0.4174±0.25	0.5041±0.03	0.4069±0.22	0.5945±0.24	0.4862±0.02	0.4645±0.05	0.5076±0.04
30 Days	0.4983±0.02	0.6959±0.19	0.3161±0.18	0.4990±0.03	0.5085±0.25	0.4906±0.26	0.4913±0.02	0.4703±0.04	0.5100±0.05

	RT			SVM		
	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity
7 Days	0.5010±0.04	0.4862±0.05	0.5148±0.06	0.5090±0.04	0.5471±0.30	0.4750±0.31
14 Days	0.4942±0.02	0.4951±0.04	0.4929±0.04	0.5102±0.03	0.5356±0.32	0.4775±0.31
21 Days	0.4941±0.03	0.4834±0.05	0.5046±0.03	0.5041±0.03	0.4713±0.30	0.5252±0.29
30 Days	0.5008±0.03	0.4956±0.04	0.5043±0.05	0.5103±0.03	0.5315±0.27	0.4828±0.28

Table 3 Summary of average forecasting results for volumes

	NB			LR			k-NN		
	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity
7 Days	0.5025±0.02	0.8211±0.24	0.2142±0.22	0.5687±0.03	0.2863±0.21	0.8190±0.14	0.5458±0.03	0.5314±0.05	0.5581±0.04
14 Days	0.5139±0.03	0.8400±0.17	0.2095±0.18	0.5696±0.04	0.3556±0.17	0.7665±0.09	0.5237±0.03	0.5013±0.06	0.5455±0.05
21 Days	0.5036±0.03	0.8291±0.16	0.2044±0.16	0.5751±0.04	0.3943±0.17	0.7396±0.11	0.5184±0.02	0.4896±0.04	0.5440±0.05
30 Days	0.4978±0.03	0.7496±0.28	0.2789±0.27	0.5761±0.03	0.4267±0.13	0.7040±0.09	0.4899±0.03	0.4633±0.06	0.5127±0.04

	RT			SVM		
	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity
7 Days	0.5184±0.03	0.5111±0.03	0.5254±0.05	0.5474±0.03	0.3390±0.31	0.7401±0.25
14 Days	0.5311±0.03	0.5015±0.03	0.5585±0.04	0.5383±0.03	0.2694±0.29	0.7876±0.25
21 Days	0.5370±0.03	0.5092±0.04	0.5619±0.03	0.5332±0.03	0.2182±0.23	0.8153±0.22
30 Days	0.5369±0.03	0.5149±0.04	0.5553±0.04	0.5368±0.02	0.2104±0.24	0.8127±0.23

4.2.2 Performance by Time Horizons

To investigate whether the number of days tracing back influences the overall forecasting performance, we averaged the performance of the five methods in four backdating lengths, as shown in Figures 2 and 3.

When forecasting price trends as shown in Figure 2, it becomes evident that the model performs most effectively at the shortest backtracking time of 7 days. At this interval, the accuracy score reaches its highest value of 0.5052, while sensitivity achieves its peak value of 0.5622. However, as the backtracking time increases to 14 days, both accuracy and sensitivity scores drop, reaching their lowest levels of 0.5021 and 0.5546, respectively, at 21 days. Notably, as the backtracking time extends further to 30 days, there is a partial recovery in both accuracy and sensitivity, although they do not reach the levels observed at 14 days. The accuracy score at 30 days is 0.4999, and sensitivity reaches 0.5404. In contrast, the behavior of specificity values exhibits a symmetrical pattern. At the shortest backtracking time of 7 days, the model's specificity is the lowest, with a value of 0.4498. As the backtracking time increases to 14 days, specificity values improve, reaching their highest value of 0.5099 at 21 days. However, with a further increase in backtracking time to 30 days, specificity values decline to 0.4607, although they do not reach the level observed at 14 days.

Figure 3 displays the accuracy, sensitivity, and specificity of the volume forecasting model over different time intervals (7 days, 14 days, 21 days, and 30 days). According to the accuracy scores, it is evident that there is a slight decrease over time. The accuracy values range from 0.5366 for the 7-day interval to 0.5275 for the 30-day interval. This indicates that the model's overall predictive accuracy slightly declines as the time horizon increases. Similarly, the sensitivity scores, which measure the model's ability to correctly identify positive cases, exhibit a decreasing trend as the time interval lengthens, ranging from 0.4700 to 0.5039. These results suggest that the model's ability to correctly identify positive cases experiences a slight reduction over longer prediction periods. In terms of specificity, the scores remain relatively stable across the different time intervals, with values ranging from 0.5714 to 0.5735.

This suggests that the model maintains a consistent level of accuracy in identifying negative cases, regardless of the length of the prediction period.

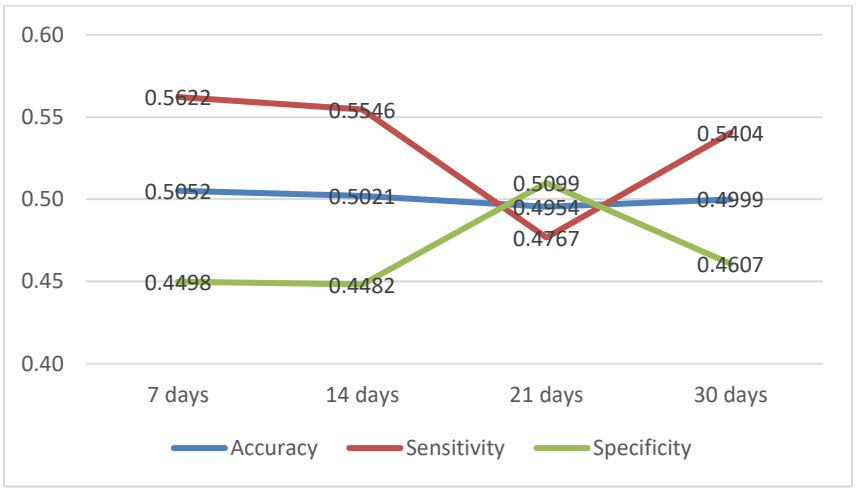


Figure 2 Averaged performance of the five methods in price forecasting given 7, 14, 21 and 30 days data.

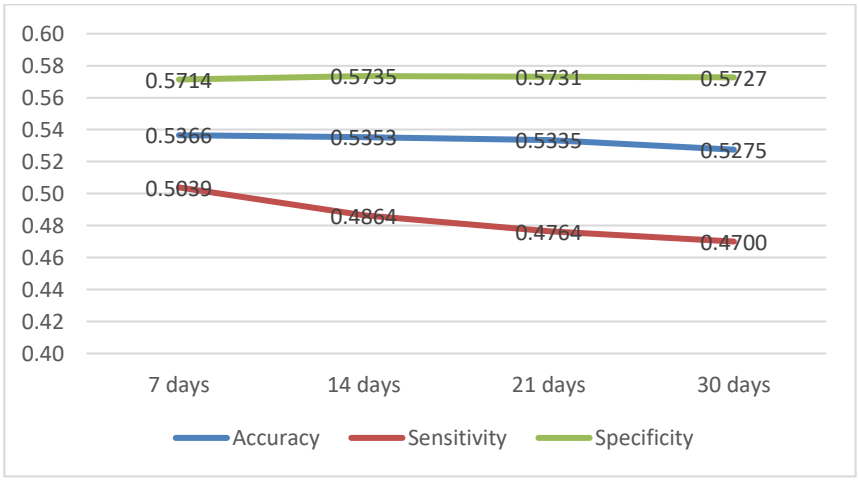


Figure 3 Averaged performance of the five methods in volume forecasting given 7, 14, 21 and 30 days data.

4.2.3 Performance by Methods

We seek to examine the predictive power and applicability of the five methods excluding the number of days of backdating. Tables 3 and 4 show the average performance values obtained for each machine learning classifier after averaging over four time periods. One can observe that the accuracy scores for volume forecasting (Table 4) tend to be higher compared to those for price forecasting (Table 3). This suggests that the models perform relatively better in changes in trading volume compared to price. The high accuracy of volume forecasts can be attributed to a variety of factors. Volume is a fundamental indicator of market activity and liquidity, and the factors it is subject to may exhibit more discernible patterns that can be captured by the model, leading to higher accuracy. Price forecasting, on the other hand, is often more challenging as it is influenced by numerous complex and interrelated factors that often introduce greater uncertainty and noise, making it more difficult for models to accurately predict price trends.

Table 4 Average performances of price forecasting by different classification methods

	Performance Measures		
	Accuracy	Sensitivity	Specificity
NB	0.4963	0.6913	0.3074
LR	0.5077	0.4870	0.5262
k-NN	0.4934	0.4776	0.5079
RT	0.4975	0.4901	0.5042
SVM	0.5084	0.5214	0.4901

Table 5 Average performances of volume forecasting by different classification methods

	Performance Measures		
	Accuracy	Sensitivity	Specificity
NB	0.5045	0.8099	0.2267
LR	0.5724	0.3658	0.7573
k-NN	0.5195	0.4964	0.5401
RT	0.5309	0.5092	0.5503
SVM	0.5389	0.2593	0.7889

In addition to the focus on accuracy, we can draw some patterns and observations about each model. For Naïve Bayes predicting changes in price, the recall score is 0.6913, indicating that it correctly identifies positive price trends (increases) 69.13% of the time, and the specificity is 0.3074, indicating that it correctly identifies negative price trends (decreases) only 30.74% of the time. This difference in predictive power is similarly enlarged when forecasting changes in trading volumes. The sensitivity score is 0.8099, indicating a superior ability to identify positive volume trends. However, the specificity is low at 0.2267, indicating a poor ability to identify negative volume trends.

Logistic regression, RT and k-NN all show better ability in identifying negative trends rather than identifying positive trends. Again, this difference is magnified when forecasting changes in trading volumes. Especially, LR achieves an accuracy of 0.5724, the highest among the methods for volume forecasting. The recall is 0.3658, indicating a moderate ability to identify positive volume trends. The specificity is 0.7573, indicating a good ability to identify negative volume trends.

SVM's performance patterns differ between price and volume forecasting. In price forecasting, SVM achieves the highest accuracy among the models, indicating its overall reliability. It shows relatively balanced recall and specificity scores, suggesting that it performs reasonably well in identifying both positive and negative price trends. However, in volume forecasting, SVM's recall score is notably lower compared to other models, indicating a reduced ability to detect positive volume trends. Nevertheless, its high specificity score suggests that SVM can still identify negative volume trends relatively well.

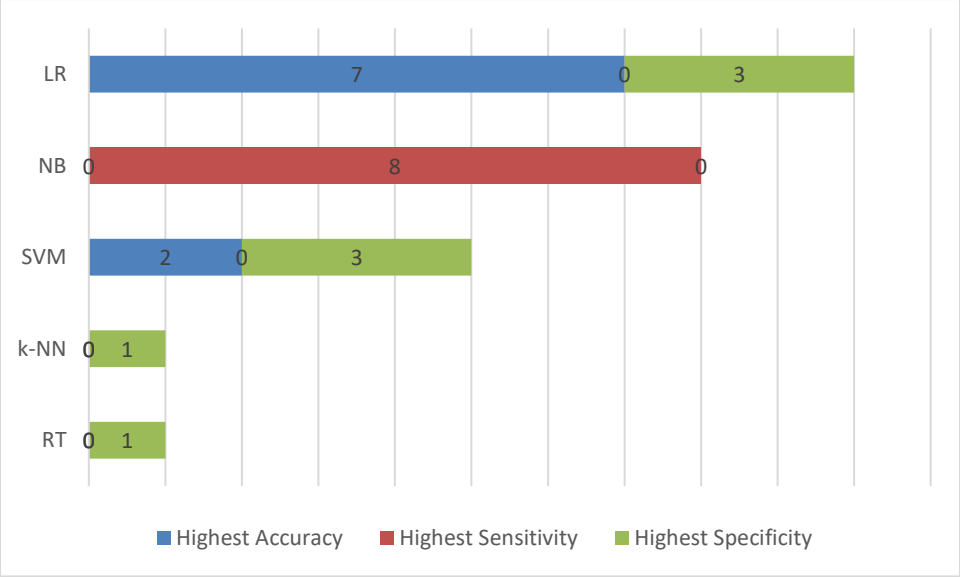


Figure 4 Overall counts of the five machine learning classifiers that received the highest values in accuracy, sensitivity, or specificity

We divided all forecasts into four groups by backdating length, namely 7 days, 14 days, 21 days, and 30 days, and since there are four groups each for price forecasts and volume forecasts, a total of eight rankings are known. Figure 4. provides an overview of the highest counts achieved by five machine learning classifiers (Logistic Regression, Naive Bayes, Support Vector Machine, k-Nearest Neighbors, and Regression Tree) in accuracy, sensitivity, specificity in eight instances. One can see that among these classifiers, Logistic Regression demonstrated the highest accuracy score, achieving the highest count of 7 out of 8 instances. Naive Bayes, on the other hand, exhibited the highest sensitivity count in all 8 instances, demonstrated its competence identifying positive directions. In terms of specificity, both Logistic Regression and Support Vector Machine achieved the highest counts, with 3 instances each out of a total of 10. It is important to note that k-Nearest Neighbors and Regression Tree models achieved the lowest counts across all metrics, with only 1 instance of achieving the highest specificity count.

4.2.4 Performance by Cryptocurrencies

While our initial approach involved the calculation of averaged predictions across all twenty cryptocurrencies, thereby providing an overarching assessment of the performance of each machine learning method, a more intricate analysis is warranted to unravel the diverse facets underlying these outcomes. Within this context, an exploration of the individual cryptocurrency's prediction results assumes a paramount role, enabling us to meticulously discern the categories of cryptocurrencies that stand out as exemplars of prediction success across varying contextual dimensions.

In pursuit of this endeavor, we embarked on a meticulous examination of the results, meticulously identifying the optimal single cryptocurrency for each model and time horizon, based on the dual performance measures of sensitivity and specificity. This selection allows us to take a more nuanced look at unique cryptocurrencies that show higher applicability within each machine learning method and time horizon.

Table 6 Best Single Crypto in Price Trend Predictions

	NB		LR		k-NN		RT		SVM	
	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity
7 Days	BTG	XLM	ERG	DOGE	BTG	ERG	ERG	USDT	LINK	DOGE
14 Days	DASH	XLM	GLM	DOGE	USDT	ERG	USDT	DGB	GLM	DOGE
21 Days	DASH	BTG	GLM	DOGE	USDT	ERG	USDT	DOGE	XMR	GLM
30 Days	BCH	DGB	TRX	USDT	LINK	DOGE	XLM	ERG	TRX	DOGE

Table 7 Best Single Crypto in Trading Volume Trend Predictions

	NB		LR		k-NN		RT		SVM	
	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity
7 Days	DASH	DCR	ERG	CVC	ERG	GLM	ERG	CVC	GLM	DOGE
14 Days	DASH	DCR	DCR	BTG	DCR	ETC	GLM	DOGE	BAT	DOGE
21 Days	DOGE	DCR	USDT	CVC	XTZ	BTG	XRP	LINK	GLM	DOGE
30 Days	DOGE	DCR	GLM	DOGE	ENJ	XTZ	XLM	XTZ	GLM	DOGE

The experiment indicates that the effectiveness of machine learning methods for cryptocurrency prediction tasks is nuanced and context-dependent, varying between price and trading volume prediction. Certain cryptocurrencies consistently demonstrate strong predictive potential across multiple models and time periods. For example, for prize trend prediction, in terms of specificity, Dogecoin (DOGE) achieved the best-performing cryptocurrency of SVM for 7, 21, 30 days' time lags in terms of specificity, also in trading volume prediction, DOGE again achieved the best-performing cryptocurrency of SVM for all four time lags.

While certain cryptocurrencies exhibit consistent predictive potential across both price and trading volume prediction tasks, there are notable differences in the optimal choices for each task. Dash (DASH) performs stable in terms of sensitivity when utilizing NB model. In price trend prediction, DASH achieved best-performing cryptocurrency in 14 days and 21 days instances, it also achieved best-performing cryptocurrency in 7 days and 14 days instances for trading volume trend prediction. However, the sensitivity obtained from the predictions using other machine learning methods did not make the DASH perform more prominently, this superior performance was only present in the predictions of NB.

These findings underscore the importance of adopting a tailored approach when selecting machine learning methods for different prediction tasks and cryptocurrencies. The results further emphasize the need for continuous refinement and adaptation of prediction models to capture the evolving dynamics of the cryptocurrency market.

To summarize our results, in terms of accuracy, the best performer for price prediction is SVM, the worst is k-NN, the best performer for volume prediction is logistic regression and the worst is NB. The time horizon effect shows that shorter backtracking times yield higher accuracy in predicting both cryptocurrency prices and volumes.

From a managerial perspective, SVM should be adopted by traders to predict price trend and LR should be adopted to predict volume trend as they showed relatively stable higher accuracy in each experiment. To be more precise, NB can be used as an aid in predicting upward trend for both price and trading volume as it performs well in sensitivity. For trading volume prediction, both LR and SVM can be used to predict downward trends.

Need to know that, while some cryptocurrencies maintain consistent applicability across different prediction tasks, others show variability in their predictive potential. It is important to tailor the predictive model to the task and the specific attributes of the cryptocurrency under consideration.

The limitation of moderate accuracy in our study arises primarily from the inherent difficulty of predicting directions in a classification problem compared to predicting price or volume levels. Unlike predicting price or volume levels, where the model can generate continuous values, predicting directions involves determining whether the market will move up or down. This binary nature of the prediction task makes it more susceptible to noise and uncertainties. Another notable limitation of our study is that the parameters of the Regression Trees (RT) and Support Vector Machine (SVM) models were not fully optimized, different parameter configurations can significantly impact the model's accuracy and generalization capabilities. To address this limitation, future research should prioritize an in-depth exploration of parameter optimization techniques to enhance the accuracy and reliability of predictive models.

Chapter 5
Conclusion

Motivated by growing interest in cryptocurrency trading and the need for accurate forecasting tools, the study expands the scope beyond popular cryptocurrencies and investigates 20 different cryptocurrencies, offering a comprehensive analysis of machine learning models' applicability in cryptocurrency prediction. The selected models for study include naïve Bayes, support vector machines, logistic regression, regression trees, and the K-nearest neighbors' algorithm. Performance measures such as accuracy, sensitivity, and specificity are used to assess the models.

In terms of major findings, when examining the predictive characteristics of different models, the study reveals that the predictive power and applicability of different models are more pronounced in predicting volume trends than in price forecasting. In addition, Naïve Bayes shows exclusive ability in identifying positive trends. Logistic regression achieves the highest accuracy and specificity the most times, demonstrating greater applicability. Both regression trees and k-NN performed stably and did not reflect exceptional predictive power. SVM achieves the highest accuracy among the models in price trends forecasting and shows good ability in predicting negative changes in trading volume.

When investigating whether the number of days tracing back influences the overall forecasting performance, we find that shorter backtracking times yield higher accuracy and sensitivity in predicting cryptocurrency prices. However, specificity shows a different pattern, with improved performance at intermediate backtracking times.

The research also highlights the importance of considering volume changes in cryptocurrency forecasting. The accuracy and sensitivity of the volume forecasting model exhibited a slight decrease as the time interval lengthened. The specificity scores remained relatively stable across different time intervals, indicating consistent accuracy in identifying negative cases regardless of the prediction period.

Due to their unique attributes, market behaviors and external influences, cryptocurrencies exhibit varying degrees of predictive potential across different forecasting tasks (price and trading volume). Certain cryptocurrencies like DOGE may consistently demonstrate predictive capabilities across various tasks. Conversely, cryptocurrencies like DASH may specialize or excel in specific aspects of prediction. This emphasizes the importance of carefully selecting appropriate machine learning methods and considering cryptocurrency characteristics when designing prediction models.

In conclusion, this thesis contributes to the field of cryptocurrency prediction by investigating the price and trading volume changes of a broad range of cryptocurrencies. The study emphasizes the importance of machine learning techniques and the incorporation of relevant indicators in forecasting models. The findings shed light on the optimal backtracking time for accurate predictions and underscore the significance of volume analysis in understanding cryptocurrency market trends. The research findings have implications for investors and researchers interested in leveraging machine learning algorithms to forecast cryptocurrency movements and make informed investment decisions. In ever-changing supply chain operations, accuracy, efficiency, and adaptability are critical, and insights are provided by our study. By harnessing the predictive potential of machine learning to anticipate price fluctuations and volume dynamics in the cryptocurrency space, supply chain practitioners can leverage this powerful technology to forecast demand patterns, optimize inventory levels, and streamline distribution networks. Thus, the response to our experiment extends beyond the financial sector, bringing about a paradigm shift in the supply chain, propelling it into an era defined by data-driven foresight and operational excellence.

References

Adcock, R., & Gradojevic, N. (2019). Non-fundamental, non-parametric Bitcoin forecasting. *Physica D: Nonlinear Phenomena*, *531*, 121727.

Allison, P. D. (2012). Logistic Regression Using SAS: Theory and Application, Second Edition. *SAS Institute.*

Andreessen, M. (2014, January 22). *Why Bitcoin Matters*. DealBook.

Barnwal, A., Bharti, H. P., Ali, A., & Singh, V. (2019). Stacking with Neural Network for Cryptocurrency investment. *2019 New York Scientific Data Summit (NYSDS)*, 1.

Bender, R., & Grouven, U. (1998). Ordinal logistic regression in medical research. *PubMed*, *31*(5), 546–551.

Bouri, E., Lau, C. K. M., Lucey, B. M., & Roubaud, D. (2019). Trading volume and the predictability of return and volatility in the cryptocurrency market. *Finance Research Letters*, *29*, 340–346.

Breiman, L. (1984). Classification and Regression Trees. *Biometrics*, *40*(3), 874.

Cheah, E., & Fry, J. C. (2015). Speculative bubbles in Bitcoin markets? An empirical investigation into the fundamental value of Bitcoin. *Economics Letters*, *130*, 32–36.

Chen, M., Narwal, N., & Schultz, M. (2019). Predicting Price Changes in Ethereum. *Available at http://cs229.stanford.edu/proj2017/final-reports/5244039.pdf*

Choi, T. (2020). Creating all-win by blockchain technology in supply chains: Impacts of agents' risk attitudes towards cryptocurrency. *Journal of the Operational Research Society*, *72*(11), 2580–2595.

Chowdhury, R. A., Rahman, M. M., Rahman, M. M., & Mahdy, M. (2020). An approach to predict and forecast the price of constituents and index of cryptocurrency using machine learning. *Physica D: Nonlinear Phenomena*, *551*, 124569.

CoinMarketCap. (n.d.). *Cryptocurrency Prices, Charts And Market Capitalizations | CoinMarketCap*. https://coinmarketcap.com/

Cover, T. M., & Hart, P. E. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, *13*(1), 21–27.

Dag, A., Dag, A., Asilkalkan, A., Simsek, S., & Delen, D. (2023). A Tree Augmented Naïve Bayes-based methodology for classifying cryptocurrency trends. *Journal of Business Research*, *156*, 113522.

Duan, L., Peng, D., & Li, A. (2014). A new naive Bayes text classification algorithm. *Telkomnika: Indonesian Journal of Electrical Engineering*, *12*(2).

Dubey, R., Gunasekaran, A., Childe, S. J., Wamba, S. F., & Papadopoulos, T. (2015). The impact of big data on world-class sustainable manufacturing. *The International Journal of Advanced Manufacturing Technology*, *84*(1–4), 631–645.

Dwyer, G. P. (2015). The economics of Bitcoin and similar private digital currencies. *Journal of Financial Stability*, *17*, 81–91.

Dyhrberg, A. H., Foley, S., & Svec, J. (2018). How investible is Bitcoin? Analyzing the liquidity and transaction costs of Bitcoin markets. *Economics Letters*, *171*, 140–143.

Ergen, B., Baykara, M., & Polat, C. (2014). An investigation on magnetic imaging findings of the inner ear: A relationship between the internal auditory canal, its nerves and benign paroxysmal positional vertigo. *Biomedical Signal Processing and Control*, *9*, 14–18.

Farell, R. (2015). An Analysis of the Cryptocurrency Industry. *Wharton Research Scholars*, 130.

Feng, C., Jin, G., & Wang, L. (2005). Cancer diagnosis and protein secondary structure prediction using support vector machines. In *Studies in fuzziness and soft computing* (pp. 343–363).

Gunasekaran, A., Tiwari, M. K., Dubey, R., & Wamba, S. F. (2016). Big data and predictive analytics applications in supply chain management. *Computers & Industrial Engineering*, *101*, 525–527.

Hamledari, H., & Fischer, M. (2021). The application of blockchain-based crypto assets for integrating the physical and financial supply chains in the construction & engineering industry. *Automation in Construction*, *127*, 103711.

He, Q., & Wang, J. (2007). Fault Detection Using the k-Nearest Neighbor Rule for Semiconductor Manufacturing Processes. *IEEE Transactions on Semiconductor Manufacturing*, *20*(4), 345–354.

Hitam, N. A., Ismail, A. R., & Saeed, F. (2019). An Optimized Support Vector Machine (SVM) based on Particle Swarm Optimization (PSO) for Cryptocurrency Forecasting. *Procedia Computer Science*, *163*, 427–433.

Jaquart, P., Köpke, & Weinhardt, C. (2022). Machine learning for cryptocurrency market prediction and trading. *The Journal of Finance and Data Science*, *8*, 331–352.

Jay, P., Kalariya, V., Parmar, P., Tanwar, S., Kumar, N., & Alazab, M. (2020). Stochastic Neural Networks for Cryptocurrency Price Prediction. *IEEE Access*, *8*, 82804–82818.

Ji, S., Kim, J., & Im, H. (2019). A comparative study of Bitcoin price prediction using deep learning. *Mathematics*, *7*(10), 898.

Khedr, A. M., Arif, I., P, P. R., El-Bannany, M., Alhashmi, S. M., & Sreedharan, M. (2021). Cryptocurrency price prediction using traditional statistical and machine-learning techniques: A survey. *International Journal of Intelligent Systems in Accounting, Finance & Management*, *28*(1), 3–34.

Kim, K. I., Jung, K., & Kim, J. H. (2005). Fast Color Texture-Based object Detection in images: application to license plate localization. In *Studies in fuzziness and soft computing* (pp. 297–320).

Kim, Y. B., Kim, J. G., Kim, W., Im, J. H., Kim, T. H., Kang, S., & Kim, C. H. (2016). Predicting fluctuations in cryptocurrency transactions based on user comments and replies. *PLOS ONE*, *11*(8), e0161197.

Koirala, R. C., Dahal, K., & Matalonga, S. (2019). Supply Chain using Smart Contract: A Blockchain enabled model with Traceability and Ownership Management. *2019 9th International Conference on Cloud Computing, Data Science & Engineering*.

Lahmiri, S., & Bekiros, S. (2019). Cryptocurrency forecasting with deep learning chaotic neural networks. *Chaos Solitons & Fractals*, *118*, 35–40.

Lahmiri, S., & Bekiros, S. (2021). Deep Learning Forecasting in Cryptocurrency High-Frequency Trading. *Cognitive Computation*, *13*(2), 485–487.

Lahmiri, S., Bekiros, S., & Bezzina, F. (2022). Complexity analysis and forecasting of variations in cryptocurrency trading volume with support vector regression tuned by Bayesian optimization under different kernels: An empirical comparison from a large dataset. *Expert Systems with Applications*, *209*, 118349.

Lahmiri, S., Saadé, R. G., Morin, D., & Nebebe, F. (2020). An Artificial Neural Networks Based Ensemble System to Forecast Bitcoin Daily Trading Volume. In *International Conference on Cloud Computing*.

Langarizadeh, M., & Moghbeli, F. (2016). Applying Naive Bayesian Networks to Disease Prediction: a Systematic Review. *Acta Informatica Medica : AIM : Journal of the Society for Medical Informatics of Bosnia & Herzegovina : ČAsopis DrušTva Za Medicinsku Informatiku BiH*, *24*(5), 364.

Li, T. R., Chamrajnagar, A. S., Fong, X. R., Rizik, N. R., & Fu, F. (2019). Sentiment-Based prediction of alternative cryptocurrency price fluctuations using gradient boosting tree model. *Frontiers in Physics*, *7*.

Madan, I., Saluja, S., & Zhao, A. (2015). Automated bitcoin trading via machine learning algorithms. *Stanford University*.

Mallqui, D. C. A., & Fernandes, R. a. S. (2019). Predicting the direction, maximum, minimum and closing prices of daily Bitcoin exchange rate using machine learning techniques. *Applied Soft Computing*, *75*, 596–606.

Mcnally, S., Roche, J. T., & Caton, S. (2018). Predicting the Price of Bitcoin Using Machine Learning. *2018 26th Euromicro International Conference on Parallel, Distributed and Network-based Processing (PDP)*.

Miura, R., Pichl, L., & Kaizoji, T. (2019). Artificial neural networks for realized volatility prediction in cryptocurrency time series. In *Lecture Notes in Computer Science* (pp. 165–172).

Monsalve, S. A., Suárez-Cetrulo, A. L., Cervantes, A., & Quintana, D. (2020). Convolution on neural networks for high-frequency trend prediction of cryptocurrency exchange rates using technical indicators. *Expert Systems with Applications*, *149*, 113250.

Nakamoto, S. (2006). Bitcoin: A Peer-to-Peer Electronic Cash System. *www.bitcoin.org*.

Nie, G., Rowe, W., Zhang, L., Tian, Y., & Shi, Y. (2011). Credit card churn forecasting by logistic regression and decision tree. *Expert Systems with Applications*, *38*(12), 15273–15285.

Ortu, M., Uras, N., Conversano, C., Bartolucci, S., & Destefanis, G. (2022). On technical trading and social media indicators for cryptocurrency price classification through deep learning. *Expert Systems with Applications*, *198*, 116804.

Parikh, R., Mathai, A., Parikh, S., Sekhar, G. C., & Thomas, R. (2008). Understanding and using sensitivity, specificity and predictive values. *Indian Journal of Ophthalmology*, *56*(1), 45.

Patel, M. M., Tanwar, S., Gupta, R., & Kumar, N. (2020). A Deep Learning-based Cryptocurrency Price Prediction Scheme for Financial Institutions. *Journal of Information Security and Applications*, *55*, 102583.

Poongodi, M., Nguyen, T. D., Hamdi, M., & Cengiz, K. (2021). Global cryptocurrency trend prediction using social media. *Information Processing and Management*, *58*(6), 102708.

Poongodi, M., Sharma, A., Vijayakumar, V., Bhardwaj, V., Sharma, A., Iqbal, R., & Kumar, R. (2020). Prediction of the price of Ethereum blockchain cryptocurrency in an industrial finance system. *Computers & Electrical Engineering*, *81*, 106527.

Russell, S. S., & Norvig, P. (1995). Artificial Intelligence: A Modern Approach. *Prentice Hall, Englewood Cliffs, NJ*.

Şahin, Y., & Duman, E. (2011). Detecting credit card fraud by ANN and logistic regression. *IEEE*, 315–319.

Shintate, T., & Pichl, L. (2019). Trend Prediction Classification for High Frequency Bitcoin Time Series with Deep Learning. *Journal of Risk and Financial Management*, *12*(1), 17.

Subramanian, G. B. V., SreekantanThampy, A., Ugwuoke, N. V., & Ramnani, B. (2021). Crypto Pharmacy – Digital Medicine: A mobile application integrated with hybrid blockchain to tackle the issues in pharma supply chain. *IEEE Open Journal of the Computer Society*, *2*, 26–37.

Trevethan, R. (2017). Sensitivity, Specificity, and Predictive Values: foundations, pliabilities, and pitfalls in research and practice. *Frontiers in Public Health*, *5*.

Van Nguyen, T., Zhou, L., Spiegler, V., Ieromonachou, P., & Lin, Y. (2018). Big data analytics in supply chain management: A state-of-the-art literature review. *Computers & Operations Research*, *98*, 254–264.

Vapnik, V. (1995). The Nature of Statistical Learning Theory. *Springer eBooks*.

Walther, B., & Moore, J. L. (2005). The concepts of bias, precision and accuracy, and their use in testing the performance of species richness estimators, with a literature review of estimator performance. *Ecography*, *28*(6), 815–829.

Wang, L. (2005). Support Vector Machines: Theory and applications. In *Springer eBooks*.

Wang, Y., Han, J. W., & Beynon-Davies, P. (2019). Understanding blockchain technology for future supply chains: a systematic literature review and research agenda. *Supply Chain Management*, *24*(1), 62–84.

Wang, Y., Wang, C., Sensoy, A., Yao, S., & Cheng, F. (2022). Can investors' informed trading predict cryptocurrency returns? Evidence from machine learning. *Research in International Business and Finance*, *62*, 101683.

Wu, C., Lu, C., Ma, Y., & Lu, R. (2018). A New Forecasting Framework for Bitcoin Price with LSTM. *IEEE International Conference on Data Mining Workshops (ICDM Workshops)*.

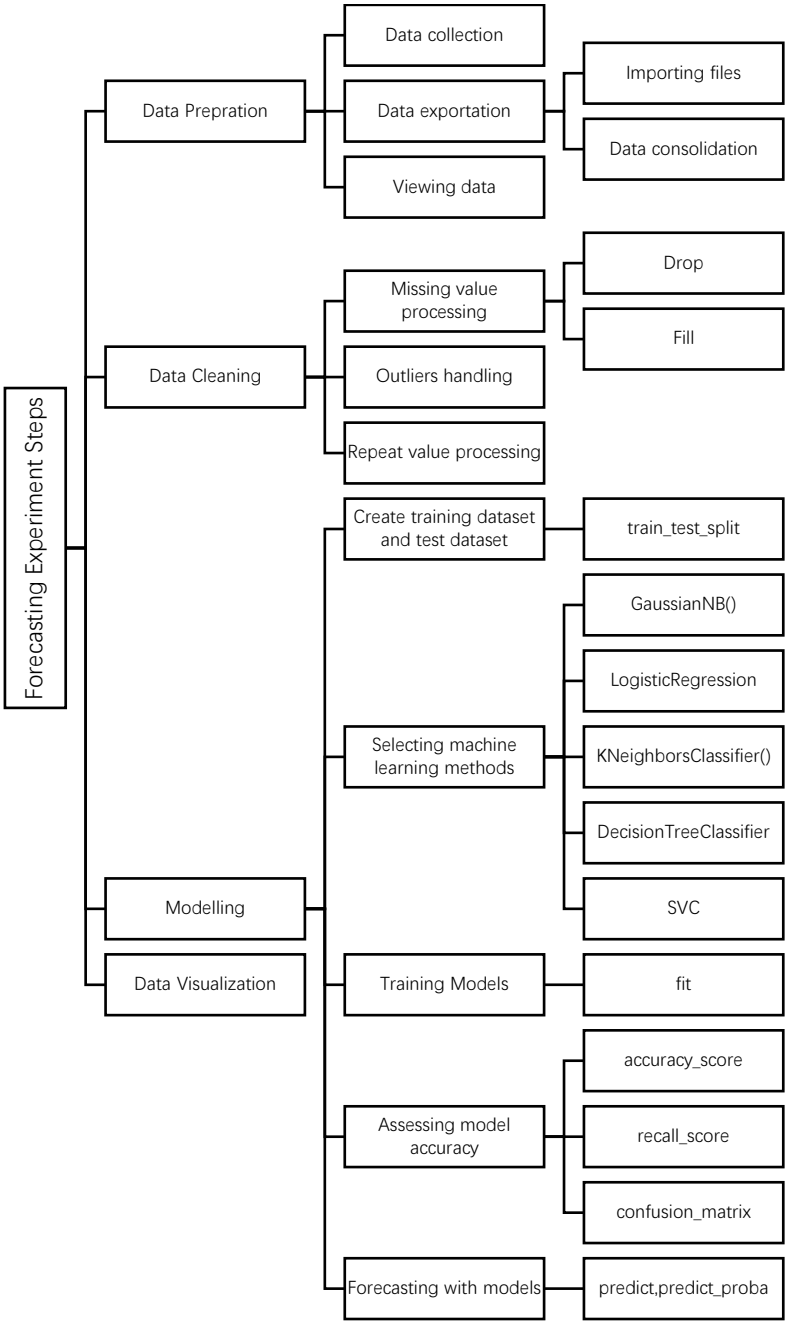
Yermack, D. (2015). Is Bitcoin a Real Currency? An Economic Appraisal. In *Elsevier eBooks* (pp. 31–43). Elsevier BV.

YiYing, W., & Yeze, Z. (2019). Cryptocurrency Price Analysis with Artificial Intelligence. In *International Conference on Information Management*.

Appendix

In the appendix, we provide additional detailed information that complements the main text. Firstly, we outline the steps involved in our prediction process with a multi-level hierarchical graph, giving readers a clear understanding of our methodology. Next, we present detailed examples of the data used during the data collection phase. We also show the specific data formats resulting from our manual data processing, using Crypto1 as an example. Additionally, we offer a practical view of our research by sharing a sample of machine learning model predictions. To provide context, the appendix lists the real virtual currency types corresponding to Crypto1 through Crypto20. This supplementary information is meant to enhance the overall clarity and completeness of our research.

Figure 5A Steps of Forecasting Experiment



The hierarchical graph outlines the structured path we've undertaken to develop a cryptocurrency forecasting model, harnessed through Python programming and the sklearn library. Beginning with data preparation, we embark on data collection, bringing together a range of relevant cryptocurrency data. Then data exportation, where we seamlessly merge files through importing and consolidation, ensuring clarity and coherence. The process is bolstered by meticulous handling of importing files and data consolidation, streamlining the organizational aspects. Moving into Data Cleaning, a vital phase, we address missing values with a choice between 'Drop' or 'Fill' strategies, emphasizing data integrity. Further, we handle Outliers to prevent distortions. Repeat Value Processing adds finesse to data refinement, strengthening the dataset. In the Modeling realm, our predictive structure takes form. This involves crafting Training and Test Datasets through 'train_test_split' for robust validation. The selection of machine learning methods—Gaussian Naive Bayes, Logistic Regression, K-Nearest Neighbors, Decision Tree, and Support Vector Classification—through sklearn library. Training Models unfolds as we 'fit' algorithms, translating theoretical prowess into practical application. Evaluation gains prominence through Model Accuracy, gauged via 'accuracy_score', 'recall_score', and a confusion matrix that provides the specificity score.

The model can be formally used for prediction after all the above steps have been completed, and we have used an automated form of reading data files so that the 20 cryptocurrencies are predicted sequentially, which makes our manual work easier and automatically generates results that are easy to read.

Table 8A Cryptocurrency 1-20 Name Index

Index	Cryptocurrency
Crypto1	TRX
Crypto2	BCH
Crypto3	DOGE
Crypto4	ETC
Crypto5	ETH
Crypto6	LINK
Crypto7	USDT
Crypto8	XLM
Crypto9	XMR
Crypto10	XRP
Crypto11	XTZ
Crypto12	BAT
Crypto13	BTG
Crypto14	CVC
Crypto15	DASH
Crypto16	DCR
Crypto17	DGB
Crypto18	ENJ
Crypto19	ERG
Crypto20	GLM

Table 9A Sample Raw Data of Cryptocurrency Price

Day	Crypto1	Crypto2	Crypto3	Crypto4	Crypto5	...	Crypto15	Crypto16	Crypto17	Crypto18	Crypto19	Crypto20
1	0.002344	654.302979	0.001415	14.2095	320.884003		438.83362	38.94766	0.011462	0.025334	10.01528	0.251504
2	0.002013	1007.41998	0.001163	14.6031	299.252991		680.38483	37.34174	0.00967	0.023154	9.632342	0.240394
3	0.002003	1340.44995	0.001201	19.4209	314.681		544.34497	40.48812	0.010777	0.030556	12.19917	0.260737
4	0.001783	1388.85999	0.001038	15.1837	307.90799		544.38794	43.69431	0.011643	0.036529	11.94643	0.269225
5	0.002112	1353.98999	0.001211	16.1059	316.716003		541.84808	43.85643	0.011688	0.031517	10.28418	0.273176
6	0.002485	1273.53003	0.001184	17.865999	337.631012		533.5166	41.64148	0.012079	0.026801	9.813058	0.26776
7	0.002322	1212.40002	0.001339	17.548	333.356995		542.23584	40.87409	0.011692	0.024619	9.730766	0.26521
8	0.002209	900.776001	0.00139	16.880699	330.924011		580.20532	44.43181	0.011548	0.026644	9.138749	0.268553
9	0.001984	1185.47998	0.001313	17.244499	332.394012		565.84845	44.85264	0.011682	0.030163	12.77797	0.280568
10	0.002028	1254.53003	0.001373	17.7185	347.612		623.34863	43.61602	0.011207	0.029328	9.751252	0.280939
...												
1799	6.896555	1.000074	0.112451	140.50818	0.488966		57.88554	36.54181	0.01217	0.593647	3.030456	0.349202
1800	6.886903	1.000087	0.111889	139.52072	0.481507		56.947067	37.25486	0.012419	0.598065	3.063868	0.362993
1801	7.175411	1.000057	0.113487	143.16438	0.47701		55.901173	35.89361	0.012143	0.596547	3.008812	0.358017
1802	7.333412	1.000069	0.114267	143.612	0.479918		54.711063	35.50415	0.011641	0.578156	2.894637	0.36096
1803	7.117894	1.000106	0.112434	146.09383	0.465977		54.766109	35.19766	0.011555	0.571334	2.883029	0.351165
1804	6.778678	1.00006	0.110749	145.04485	0.451227		55.425442	36.36702	0.011572	0.565699	2.704482	0.349128
1805	6.666734	0.999994	0.110085	141.20764	0.448084		55.705006	35.73229	0.011591	0.565986	2.696064	0.347633
1806	6.813489	1.000123	0.110925	140.72948	0.461098		56.89362	36.33366	0.011761	0.576964	2.675471	0.353498
1807	6.870749	1.000085	0.111254	142.80187	0.46547		54.965897	35.97055	0.011563	0.569016	2.611335	0.355219
1808	7.065635	1.000077	0.111954	144.48808	0.469033		55.269115	36.07995	0.011607	0.588018	2.597597	0.3529

Table 10A Sample Raw Data of Cryptocurrency Trading Volume

Day	Crypto1	Crypto2	Crypto3	Crypto4	Crypto5	...	Crypto15	Crypto16	Crypto17	Crypto18	Crypto19	Crypto20
1	2924350	710387008	6259550	129201000	893249984		112418333	1424076	4034612	334718	3673	2871866
2	2193620	5195420160	4246520	299856992	885985984		604342902	1880400	5227778	517305	3927	5066971
3	1748460	5139769856	2231080	958982016	842300992		395975519	1516493	2941164	1530045	8132	2562124
4	2174370	8371319808	3288960	697452992	1613479936		157703298	761832	3573402	3770585	8034	2565559
5	2889150	4850570240	2481270	350880000	1041889984		131700542	1778201	3953404	2281897	4296	2731359
6	4040400	1697910016	2660340	449737984	1069680000		124162509	1186590	3369809	885655	656	3149351
7	5008060	1321779968	2840180	149567008	722665984		121337192	988079	2967144	511911	148	2982238
8	5008060	1321779968	2840180	149567008	722665984		185330865	1134821	3021765	438179	2	2751159
9	5082450	2034690048	3423010	182720992	797254016		106123820	1681857	3264658	1160168	132	3067851
10	5720510	3203429888	2787480	150956992	621732992		134142921	1285876	2150200	602698	252	2990705
...												
1799	698236582	211846099	242691429	496015066	13113767755		78587685	5767434	6400681	20173273	1057547	3474197
1800	476732078	167106067	182208164	282062942	6798512624		81548735	10967231	4764351	21055606	1244516	93754417
1801	300535408	177467027	162246280	267766808	7491625206		81475508	3272557	11664076	31577027	1190814	50670458
1802	350961967	185301371	174261450	318526984	9401189650		83190607	2377205	3628686	38211637	1464364	38769327
1803	334605707	216478780	239236218	337779781	10416747806		67410854	1775207	2702514	25221214	1631984	10055898
1804	384236864	161389501	287297160	282074315	8350692785		87398882	2780870	3282779	27474924	2116818	10555324
1805	309499843	170432867	241388629	292496953	9009111996		82049558	7718806	1919384	21278076	1767858	6061279
1806	335452050	204900553	224787600	297010750	10412565245		90198088	4101073	3205818	22536665	1778151	10540377
1807	255324499	181042453	159431212	390333675	7175324564		97993122	2151334	2791486	23373309	1681436	6814043
1808	275426762	185071846	182523575	352468259	9909510925		94755600	1654382	2837126	40578312	1564109	6580500

Table 9A and Table 10A

In the above two tables, we show the data required for the experiment. They are past price and daily volume data for crypto1 through crypto20, respectively. The table has a total of 1808 rows, representing the prices and volumes for the past 1808 days from the start date.

Table 11A Instances of Price Forecasting of 7 days' Time Horizon for Crypto1

Instance	Day1	Day2	Day3	Day4	Day5	Day6	Day7	Output
1	0.002344	0.002013	0.002003	0.001783	0.002112	0.002485	0.002322	0
2	0.002013	0.002003	0.001783	0.002112	0.002485	0.002322	0.002209	0
3	0.002003	0.001783	0.002112	0.002485	0.002322	0.002209	0.001984	1
4	0.001783	0.002112	0.002485	0.002322	0.002209	0.001984	0.002028	0
5	0.002112	0.002485	0.002322	0.002209	0.001984	0.002028	0.002002	1
6	0.002485	0.002322	0.002209	0.001984	0.002028	0.002002	0.002134	1
7	0.002322	0.002209	0.001984	0.002028	0.002002	0.002134	0.002143	1
8	0.002209	0.001984	0.002028	0.002002	0.002134	0.002143	0.002302	0
9	0.001984	0.002028	0.002002	0.002134	0.002143	0.002302	0.002105	0
…								
1793	0.062072	0.062427	0.062832	0.061627	0.061776	0.061009	0.063941	0
1794	0.062427	0.062832	0.061627	0.061776	0.061009	0.063941	0.061823	0
1795	0.062832	0.061627	0.061776	0.061009	0.063941	0.061823	0.061821	1
1796	0.061627	0.061776	0.061009	0.063941	0.061823	0.061821	0.062665	0
1797	0.061776	0.061009	0.063941	0.061823	0.061821	0.062665	0.06217	0
1798	0.061009	0.063941	0.061823	0.061821	0.062665	0.06217	0.061783	1
1799	0.063941	0.061823	0.061821	0.062665	0.06217	0.061783	0.062213	0
1800	0.061823	0.061821	0.062665	0.06217	0.061783	0.062213	0.061662	0
1801	0.061821	0.062665	0.06217	0.061783	0.062213	0.061662	0.06164	1

Table 12A Instances of Trading Volume Forecasting of 14 days' Time Horizon for Crypto1

Instance	Day1	Day2	Day3	Day4	Day5	...	Day10	Day11	Day12	Day13	Day14	Output
1	2924350	2193620	1748460	2174370	2889150		5709820	4289470	3909200	3417520	4119730	1
2	2193620	1748460	2174370	2889150	4040400		4289470	3909200	3417520	4119730	4402340	0
3	1748460	2174370	2889150	4040400	5008060		3909200	3417520	4119730	4402340	4204470	1
4	2174370	2889150	4040400	5008060	5082450		3417520	4119730	4402340	4204470	4704770	1
5	2889150	4040400	5008060	5082450	5720510		4119730	4402340	4204470	4704770	6245890	1
6	4040400	5008060	5082450	5720510	5709820		4402340	4204470	4704770	6245890	6405580	0
7	5008060	5082450	5720510	5709820	4289470		4204470	4704770	6245890	6405580	6065010	1
8	5082450	5720510	5709820	4289470	3909200		4704770	6245890	6405580	6065010	7852660	0
9	5720510	5709820	4289470	3909200	3417520		6245890	6405580	6065010	7852660	4155450	1
...												
1786	292352772	388404433	318156914	302584824	290676174		565904749	352174118	475164602	484368346	698236582	0
1787	388404433	318156914	302584824	290676174	344121613		352174118	475164602	484368346	698236582	476732078	0
1788	318156914	302584824	290676174	344121613	299943015		475164602	484368346	698236582	476732078	300535408	1
1789	302584824	290676174	344121613	299943015	252050650		484368346	698236582	476732078	300535408	350961967	0
1790	290676174	344121613	299943015	252050650	226469326		698236582	476732078	300535408	350961967	334605707	1
1791	344121613	299943015	252050650	226469326	565904749		476732078	300535408	350961967	334605707	384236864	0
1792	299943015	252050650	226469326	565904749	352174118		300535408	350961967	334605707	384236864	309499843	1
1793	252050650	226469326	565904749	352174118	475164602		350961967	334605707	384236864	309499843	335452050	0
1794	226469326	565904749	352174118	475164602	484368346		334605707	384236864	309499843	335452050	255324499	1

Table 13A Instances of Price Forecasting of 21 days' Time Horizon for Crypto1

Instance	Day1	Day2	Day3	Day4	Day5	...	Day17	Day18	Day19	Day20	Day21	Output
1	0.002344	0.002013	0.002003	0.001783	0.002112		0.002048	0.002138	0.002062	0.002308	0.002068	1
2	0.002013	0.002003	0.001783	0.002112	0.002485		0.002138	0.002062	0.002308	0.002068	0.002244	0
3	0.002003	0.001783	0.002112	0.002485	0.002322		0.002062	0.002308	0.002068	0.002244	0.002118	0
4	0.001783	0.002112	0.002485	0.002322	0.002209		0.002308	0.002068	0.002244	0.002118	0.002073	1
5	0.002112	0.002485	0.002322	0.002209	0.001984		0.002068	0.002244	0.002118	0.002073	0.002116	1
6	0.002485	0.002322	0.002209	0.001984	0.002028		0.002244	0.002118	0.002073	0.002116	0.002144	0
7	0.002322	0.002209	0.001984	0.002028	0.002002		0.002118	0.002073	0.002116	0.002144	0.002085	1
8	0.002209	0.001984	0.002028	0.002002	0.002134		0.002073	0.002116	0.002144	0.002085	0.003055	1
9	0.001984	0.002028	0.002002	0.002134	0.002143		0.002116	0.002144	0.002085	0.003055	0.004275	1
...												
1779	0.059871	0.059727	0.059548	0.05941	0.059595		0.062832	0.061627	0.061776	0.061009	0.063941	0
1780	0.059727	0.059548	0.05941	0.059595	0.061032		0.061627	0.061776	0.061009	0.063941	0.061823	0
1781	0.059548	0.05941	0.059595	0.061032	0.061014		0.061776	0.061009	0.063941	0.061823	0.061821	1
1782	0.05941	0.059595	0.061032	0.061014	0.060475		0.061009	0.063941	0.061823	0.061821	0.062665	0
1783	0.059595	0.061032	0.061014	0.060475	0.060879		0.063941	0.061823	0.061821	0.062665	0.06217	0
1784	0.061032	0.061014	0.060475	0.060879	0.061394		0.061823	0.061821	0.062665	0.06217	0.061783	1
1785	0.061014	0.060475	0.060879	0.061394	0.06232		0.061821	0.062665	0.06217	0.061783	0.062213	0
1786	0.060475	0.060879	0.061394	0.06232	0.062426		0.062665	0.06217	0.061783	0.062213	0.061662	0
1787	0.060879	0.061394	0.06232	0.062426	0.062784		0.06217	0.061783	0.062213	0.061662	0.06164	1

Table 14A Instances of Trading Volume Forecasting of 30 days' Time Horizon for Crypto1

Instance	Day1	Day2	Day3	Day4	Day5	...	Day26	Day27	Day28	Day29	Day30	Output
1	2924350	2193620	1748460	2174370	2889150		10447300	8749920	33314700	44215300	21605600	0
2	2193620	1748460	2174370	2889150	4040400		8749920	33314700	44215300	21605600	17520500	1
3	1748460	2174370	2889150	4040400	5008060		33314700	44215300	21605600	17520500	18145200	1
4	2174370	2889150	4040400	5008060	5082450		44215300	21605600	17520500	18145200	20137200	1
5	2889150	4040400	5008060	5082450	5720510		21605600	17520500	18145200	20137200	48699100	1
6	4040400	5008060	5082450	5720510	5709820		17520500	18145200	20137200	48699100	174319008	1
7	5008060	5082450	5720510	5709820	4289470		18145200	20137200	48699100	174319008	192816992	0
8	5082450	5720510	5709820	4289470	3909200		20137200	48699100	174319008	192816992	139710000	1
9	5720510	5709820	4289470	3909200	3417520		48699100	174319008	192816992	139710000	336496000	1
...												
1770	397542060	351830909	262000923	313019265	403942833		565904749	352174118	475164602	484368346	698236582	0
1771	351830909	262000923	313019265	403942833	353093809		352174118	475164602	484368346	698236582	476732078	0
1772	262000923	313019265	403942833	353093809	431504689		475164602	484368346	698236582	476732078	300535408	1
1773	313019265	403942833	353093809	431504689	359264471		484368346	698236582	476732078	300535408	350961967	0
1774	403942833	353093809	431504689	359264471	376538249		698236582	476732078	300535408	350961967	334605707	1
1775	353093809	431504689	359264471	376538249	300528950		476732078	300535408	350961967	334605707	384236864	0
1776	431504689	359264471	376538249	300528950	267696551		300535408	350961967	334605707	384236864	309499843	1
1777	359264471	376538249	300528950	267696551	378865177		350961967	334605707	384236864	309499843	335452050	0
1778	376538249	300528950	267696551	378865177	347069034		334605707	384236864	309499843	335452050	255324499	1

Table 11A, Table 12A, Table 13A and Table 14A

In Tables 11A, 12A, 13A, and 14A, we provide four different sets of data samples for prediction using Crypto1 as an example. These tables contain examples used in the prediction process. Noted that, because the time horizon, that is the time tracing back are different, the times windows in each table are different. In addition, the number of instances generated varies accordingly due to the time window. For example, during the 7-day time window, 1801 instances were created. In the 14-day time window, 1794 instances were created; in the 21-day time horizon, 1787 instances were created; and in the 30-day time horizon, 1778 instances were created.

Table 15A Sample Results of volume prediction of 7 days' time horizon

	NB			LR			k-NN			RT			SVM		
file	**accuracy**	**recall**	**specificity**	**accuracy**	**recall**	**specificity**	**accuracy**	**recall**	**specificity**	**accuracy**	**recall**	**specificity**	**accuracy**	**recall**	**specificity**
crypto1	0.4958	0.8555	0.1649	0.5623	0.2370	0.8617	0.4931	0.4624	0.5213	0.4765	0.4913	0.4628	0.5180	0.6012	0.4415
crypto2	0.4820	0.7337	0.2604	0.5485	0.1893	0.8646	0.5180	0.4852	0.5469	0.4765	0.4438	0.5052	0.5706	0.3254	0.7865
crypto3	0.4903	0.9535	0.0688	0.5429	0.1453	0.9048	0.5540	0.5698	0.5397	0.4958	0.4767	0.5132	0.5263	0.0058	1.0000
crypto4	0.4848	0.9181	0.0947	0.5623	0.1345	0.9474	0.4931	0.4444	0.5368	0.5429	0.5614	0.5263	0.5346	0.0819	0.9421
crypto5	0.5485	0.8814	0.2283	0.6260	0.6441	0.6087	0.5291	0.5085	0.5489	0.5319	0.5819	0.4837	0.6011	0.5424	0.6576
crypto6	0.4986	0.2364	0.7194	0.5789	0.5333	0.6173	0.5956	0.5879	0.6020	0.5208	0.5697	0.4796	0.5346	0.0182	0.9694
crypto7	0.5208	0.7771	0.2796	0.6066	0.5657	0.6452	0.5291	0.5714	0.4892	0.5180	0.6000	0.4409	0.5706	0.4514	0.6828
crypto8	0.4958	0.9294	0.1099	0.5983	0.3235	0.8429	0.5485	0.5353	0.5602	0.5069	0.5353	0.4817	0.5429	0.1941	0.8534
crypto9	0.4986	0.9213	0.0874	0.5457	0.4213	0.6667	0.5706	0.5225	0.6175	0.5679	0.5337	0.6011	0.5069	0.0506	0.9508
crypto10	0.4626	0.9136	0.0955	0.6066	0.2778	0.8744	0.5568	0.5556	0.5578	0.4931	0.4753	0.5075	0.6039	0.4136	0.7588
crypto11	0.5208	0.8824	0.1322	0.5374	0.2513	0.8448	0.5540	0.5615	0.5460	0.4792	0.4599	0.5000	0.5152	0.1872	0.8678
crypto12	0.5319	0.8632	0.1637	0.4931	0.1000	0.9298	0.5817	0.5684	0.5965	0.5319	0.4632	0.6082	0.5152	0.1947	0.8713
crypto13	0.4958	0.9231	0.1198	0.5374	0.0651	0.9531	0.5014	0.4734	0.5260	0.5429	0.4438	0.6302	0.5263	0.0237	0.9688
crypto14	0.5042	0.9198	0.1658	0.5623	0.0370	0.9899	0.5429	0.4753	0.5980	0.5125	0.5000	0.5226	0.5596	0.0802	0.9497
crypto15	0.4820	0.9586	0.0625	0.5291	0.0947	0.9115	0.5540	0.6095	0.5052	0.5014	0.5444	0.4635	0.5402	0.8521	0.2656
crypto16	0.5042	0.0556	0.9503	0.6039	0.5944	0.6133	0.5568	0.5333	0.5801	0.5152	0.4611	0.5691	0.5125	0.1167	0.9061
crypto17	0.4765	0.9363	0.1225	0.5623	0.0318	0.9706	0.5346	0.5350	0.5343	0.4931	0.4777	0.5049	0.5374	0.8089	0.3284
crypto18	0.4958	0.8889	0.1421	0.5651	0.2865	0.8158	0.5319	0.4561	0.6000	0.5069	0.4678	0.5421	0.5346	0.1345	0.8947
crypto19	0.5457	0.9301	0.1371	0.6233	0.6613	0.5829	0.5928	0.6452	0.5371	0.5900	0.6129	0.5657	0.5817	0.9247	0.2171
crypto20	0.5152	0.9434	0.1782	0.5817	0.1321	0.9356	0.5789	0.5283	0.6188	0.5651	0.5220	0.5990	0.6150	0.7736	0.4901

Table 15A Sample Results of volume prediction of 7 days' time horizon

The table above shows a raw result sheet we generated after one single forecasting process. There are eight forecasting loops in total, and there are eight corresponding result sheets like this obtained. They are results of trading volume forecasting of 7 days, 14 days, 21 days and 30 days; results of price forecasting of 7 days, 14 days, 21 days and 30 days, respectively. Table 8A below shows a sample of results of volume prediction of 7 days' time horizon. For clarity and ease of comprehension, the results are presented in a tabular format with rows corresponding to individual cryptocurrencies (labeled as "crypto1" to "crypto20") and columns representing different performance measures. These measures include accuracy, recall, and specificity for each algorithm and prediction category.

List of Abbreviations

Abbreviations	Explanations
ALSTM	Attention-based Long Short Term Memory
ANN	Artificial Neural Network
AR	Auto regressive
BO	Bayesian Optimization
CNN	Convolutional Neural Network
DFFNN	Deep Feed-forward Neural Network
FFNN	Feed-forward Artificial Neural Network
GBC	Gradient Boosting Classifiers
GBT	Gradient Boosting Trees
GLM	Generalized Linear Mode
GRNN	Generalized Regression Neural Networks
GRU	Gated Recurrent Unit
HARRV	Heterogeneous Auto-Regressive Realized Volatility
K-NN	K-Nearest Neighbors
LDA	Attention-based LSTM
LR	Logistic Regression/Linear Regression
LSTM	Long Short Term Memory
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MLP	Multilayer Perceptron
MSE	Mean Square Error
NB	Naïve Bayes
RBFNN	Radial basis function neural networks
RF	Random Forest
RMSE	Root Mean Square Error
RSM	Random Sampling Method
RT	Regression Trees
RV	Realized volatility
SVM	Support Vector Machines
SVR	Support Vector Regression
TAN	Tree Augmented Naive
TCN	Temporal Convolutional Network