DOI: https://doi.org/10.4038/sljas.v25i1.8131

Sri Lankan Journal of Applied Statistics, Vol (25-1)

Effectiveness of Using Candlestick Charts to Forecast Ethereum Price Direction: A Machine Learning Approach

N.I.M.B. Senanayaka * and H.A. Pathberiya

Department of Statistics, Faculty of Applied Sciences, University of Sri Jayewardenepura, Colombo, Sri Lanka

*Corresponding author: menashabrayani@gmail.com

Received: 9th June 2024/ Revised: 30th June 2024/ Published: 10th July 2024

ABSTRACT

Cryptocurrency is a form of decentralized digital currency. Ethereum is the second-largest cryptocurrency by market capitalization and the largest altcoin. Cryptocurrencies including Ethereum are highly volatile. Hence, short-term directional forecasts in the cryptocurrency market have become a widely discussing topic. Candlestick charts are useful visualizations of the open, high, low and close prices which can identify patterns and gauge the near-term direction of prices. This research explores the effectiveness of forecasting hourly Ethereum closing price direction based on candlestick charts within a short time horizon. The proposed forecasting algorithm incorporates clustering methods such as fuzzy K-means, K-means and partition around medoids clustering to cluster candlestick chart properties namely upper shadow length, body length and lower shadow length. Classification methods such as random forest, support vector machine and K-nearest neighbour were used to forecast closing price direction using 16 different predictor variable sets including open, high, low and close prices, candlestick chart price direction, USL, BL and LSL. The accuracy for all considered cases was around 50%. Clustering improved the accuracy slightly and including the CPD with the predictor variable sets under consideration can increase the accuracy slightly. However, this approach is performing better in predicting the Down cases to the total number of actual Down cases because there is a higher sensitivity of 81.20% based on the SVM with Open, High, Low and Close at t in the clustering ignored method.

Keywords: Machine Learning, Cryptocurrency, Candlestick Charts, Directional Forecast, Clustering

1 Introduction

The digitalization is considered as the industrial revolution or the digital revolution which started in the 20^{th} century. Digitalization can be referred to as adopting digital means to enhance the existing process. The wave of digitalization has resulted in a revolutionary change in the lifestyle of human beings. Digital currency is created as a result of digitalization.

Cryptocurrency is a form of decentralized digital currency that uses cryptography to secure and verify transactions in a network. This is an alternative form of payment created using encrypted algorithms. Most cryptocurrencies depend on the blockchain technology. Blockchain technology is a distributed and advanced database mechanism or ledger that allows transparent and secure information sharing among a computer network's nodes. Cryptocurrencies are not backed by any of the public or private entities so it is difficult to make a case for their legal status throughout the world.

In recent years, there has been considerable growth in cryptocurrencies. According to CoinMarketCap (2023), there are more than 20000 cryptocurrencies by January 2023. The increasing popularity and market acceptance of cryptocurrencies have attracted both investors and the academic community. As of January 2023, Bitcoin (BTC) is the largest cryptocurrency by market capitalization and Ethereum (ETH) is the second largest cryptocurrency by market capitalization, it is the top-performing altcoin where altcoins can be described as all the cryptocurrencies other than Bitcoin.

Ethereum (ETH) is the second largest cryptocurrency and the largest altcoin by market capitalization as on January 2023. Ethereum is also called Ether. According to Chen et al. (2019), Ethereum is a highly volatile cryptocurrency. Ethereum is a native cryptocurrency of the Ethereum network. According to Ethereum (2023), one major advantage of Ethereum over Bitcoin is the unlimited supply of Ethereum. The Bitcoin system will not release new Bitcoins when the limit of 21 million coins is reached. However, Ethereum has no limit and the total supply is programmed to increase gradually providing an infinite supply of coins.

According to Nison (2001), the concept of candlestick charts was developed in the 18^{th} century by a Japanese rice trader named Munehisa Homma. This technique was developed as a technical analysis method to analyze the price of rice contracts. The candlestick analysis is a tool that offers a glimmer into the psychology of short-term trading activities. Nison (1991) introduced candlestick charting techniques to the Western world and thereafter this technique became popular in many fields. Candlestick charts are useful visualizations of open, high, low and close prices. Traders use these charts and identify the patterns and gauge the near-term direction of prices. A Candlestick chart is created using the open, high, low and close prices. The colour of the candlestick chart is based on the candlestick chart price direction which is found by comparing the open and close prices.

Several research studies (Parikh et al., 2022; Iqbal et al., 2021; Liu and Serletis, 2019) have been done with a focus on cryptocurrencies using various methods and techniques. However, the most frequently explored area was cryptocurrency price prediction and most of the researchers have focused on long-term cryptocurrency price prediction using time series models (Derbentsev et al., 2019; Yamak et al., 2019). Even though short-term forecasting (Derbentsev et al., 2019) was explored, the studies done with the focus on short-term forecasting were relatively low. Several research studies (Ortu et al., 2022; Barnwal et al., 2019) have been done regarding cryptocurrency price direction forecasting but those have mainly focused on Bitcoin short-term price direction forecasting (Pabuccu et al., 2023) and the other cryptocurrencies which are known as altcoins are kept unexplored compared to that of Bitcoin. Hence, it is important to explore other cryptocurrencies. Candlestick charts are commonly used in technical analysis. There are several well-known candlestick behaviours and the traders and investors use those standard patterns. Several research studies (Cohen, 2021; Ho et al., 2021) were found in literature regarding those standard patterns and pattern classification. Candlestick chart analysis is commonly used in other financial markets such as the forex market and the stock market. The use of candlestick charts in cryptocurrency market is not very common. Research studies done with a focus on candlestick chart properties and cryptocurrencies are comparatively low.

Based on the identified research gap which states about the importance of exploring other cryptocurrencies commonly known as altcoins based on new methods and techniques that are not explored yet, Ethereum short-term directional forecasting using candlestick charts was taken into consideration and motivated in researching it.

In this study, we explored the effectiveness of forecasting Ethereum short-term closing price direction using the properties of candlestick charts and the possibility of identifying clusters based on candlestick properties. Also we explored about the classification models that can provide accurate forecasts of the short-term price direction of Ethereum using the properties of candlestick charts. Furthermore, this paper investigates whether clustering help to improve the forecasting accuracy.

2 Methodology

This study used timely data with one hour frequency for the year 2022. First, the data preprocessing was done. New variables that include the candlestick chart properties and the target variable which is the closing price direction at the next immediate time point (FCPD) were created and added to the dataset. Then a preliminary analysis was done. Next, clustering was done to identify similar candlestick chart patterns. Sixteen different predictor variable sets were created and classification was done considering the clustering based method and clustering ignored method. Confusion matrix was derived and evaluation metrics were obtained. Finally, the effectiveness of the used predictors and methodologies in predicting the target variable was discussed.

2.1 Data Sources

This research study was carried out using secondary data and the data set was downloaded from the cryptocurrency exchange named Gemini using Cryptodatadownload Website.

Original variables were open, high, low and close price values of Ethereum with hourly frequency from 2022-01-01 00:00 to 2022-12-31 23:00 in Eastern Standard Time (EST).

2.2 Data Preprocessing

This study focused on using candlestick chart properties to forecast the price direction. The values of open, high, low and close prices, the candlestick chart price direction (CPD) and the values of upper shadow length (USL), body length (BL) and lower shadow length (LSL) were used in this study.

The method of finding the candlestick chart price direction and the methods of calculating the candlestick properties can be shown using Fig. 1.

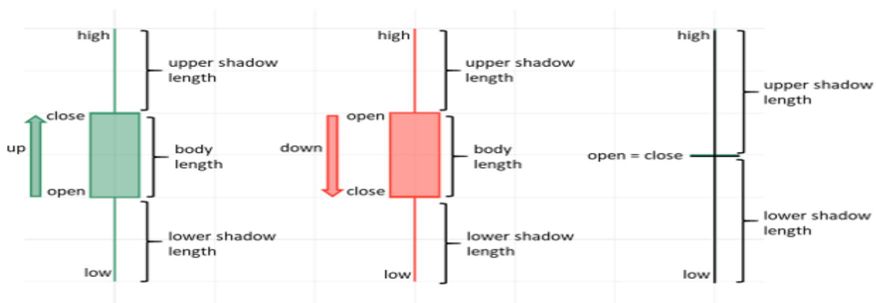

Fig. 1: Basic candlestick charts

If the open price is less than the close price then the candlestick chart price direction is Up and it is considered as a Green Candlestick.

When the candlestick chart price direction is Up, the USL, BL and LSL can be defined as in Equation 1, Equation 2 and Equation 3,

$$\text{USL} = \text{High Price - Close Price} \tag{1}$$
$$\text{BL} = \text{Close Price - Open Price} \tag{2}$$
$$\text{LSL} = \text{Open Price - Low Price} \tag{3}$$

If the open price is greater than the close price then the candlestick chart price direction is Down and it is considered as a Red Candlestick.

When the candlestick chart price direction is Down, the USL, BL and LSL can be defined as in Equation 4, Equation 5 and Equation 6,

$$\text{USL} = \text{High Price - Open Price} \tag{4}$$

$$\text{BL} = \text{Open Price - Close Price} \tag{5}$$
$$\text{LSL} = \text{Close Price - Low Price} \tag{6}$$

If the open price is the same as the close price then the candlestick chart price direction is Same and it is considered as a Doji Candlestick.

When the candlestick chart price direction is Same, the USL, BL and LSL can be defined as in Equation 7, Equation 8 and Equation 9,

$$\text{USL} = \text{High Price - Close Price (or Open Price)} \tag{7}$$
$$\text{BL} = 0 \text{ (Open price} = \text{Close price)} \tag{8}$$
$$\text{LSL} = \text{Open Price (or Close Price) - Low Price} \tag{9}$$

The past values up to three lags t-1, t-2, and t-3 for all the variables namely open price, high price, low price, close price, CPD, USL, BL and LSL were computed and included in the data set. This research study explored the candlestick chart behaviour only up to three lags because this study is focusing on a short time horizon.

The target variable which is the closing price direction at the next immediate time point (FCPD) was computed and included in the data set. The target variable can be defined as in Equation 10,

$$\text{FCPD} = \begin{cases} up & \text{if Close Price at } t < \text{Close Price at } t+1 \\ down & \text{if Close Price at } t > \text{Close Price at } t+1 \end{cases} \tag{10}$$

Here, t is the considered time point.
The situations where the close price at $t =$ close price at $t+1$ which gives the level as same were not considered in the analysis because the data for this situation was insufficient. It can be considered as a constraint under this research study.

2.3 Clustering

Clustering was done to identify similar candlestick chart patterns. The candlestick chart properties namely USL, BL and LSL were clustered using three clustering methods namely fuzzy K-means clustering, K-means clustering and partition around medoids clustering.

The initial number of desired clusters was not known at the beginning so a function was created using R programming language in order to find the optimal number of clusters.

Under this research study, the desired number of clusters under each of the above mentioned three methods was found by considering the highest silhouette index. Then the most appropriate clustering method was decided by comparing the highest silhouette indexes under each method and selecting the method with the highest silhouette index.

Silhouette index (Equation 11) is a well-known cluster validity index and this is used to evaluate the goodness of the clustering structure. This measures how well an observation is clustered by considering how close the objects are within the cluster and how far the objects are in between clusters.

$$\text{Silhouette Index} = \frac{b-a}{max(a,b)} \tag{11}$$

where a is the mean distance within the cluster and b is the mean distance to the nearest cluster.

2.3.1 Fuzzy K-means Clustering

Fuzzy K-means clustering is a method which comes under soft clustering. Under this clustering method, the data point belongs to more than one cluster with different degrees of membership. The probability score or likelihood that a data point belongs to a particular cluster is known as the degree of membership. The degree of membership ranges between zero and one and the data point is assigned to the cluster which has the highest degree of membership.

2.3.2 K-means Clustering

K-means clustering is a method which comes under hard clustering. Under this clustering method, the data point belongs to only one cluster. K-means clustering is one of the most used clustering methods which comes under partitioning clustering. The K-means clustering method finds the similarity between the observations and groups them into clusters.

2.3.3 Partition Around Medoids Clustering

Partition around medoids clustering is a method which comes under hard clustering. Under this clustering method, the data point belongs to only one cluster. This is a clustering method which is based on the medoids.

2.4 Predictor Variable Sets

The 16 different predictor variable set (X_t)s used for classification:

- USL, BL, LSL and CPD at t
- USL, BL and LSL at t
- Open, High, Low, Close and CPD at t
- Open, High, Low and Close at t
- USL, BL, LSL and CPD at t, t-1
- USL, BL and LSL at t, t-1
- Open, High, Low, Close and CPD at t, t-1
- Open, High, Low and Close at t, t-1
- USL, BL, LSL and CPD at t, t-1, t-2
- USL, BL and LSL at t, t-1, t-2
- Open, High, Low, Close and CPD at t, t-1, t-2
- Open, High, Low and Close at t, t-1, t-2
- USL, BL, LSL and CPD at t, t-1, t-2, t-3
- USL, BL and LSL at t, t-1, t-2, t-3
- Open, High, Low, Close and CPD at t, t-1, t-2, t-3
- Open, High, Low and Close at t, t-1, t-2, t-3

2.5 Classification

The predictor variable sets were used as input variables to forecast the closing price direction at the next immediate time point. Classification methods such as random forest (RF), support vector machine (SVM) and K-nearest neighbour (KNN) were used. Classification was done under two approaches namely clustering based method and clustering ignored method. The data set was trained considering 10-fold cross validation. Predictions were obtained for the test set and confusion matrix and evaluation metrics were derived. The forecast comparison was done considering the Accuracy values.

Accuracy (Equation 12) is the ratio of correct predictions to the total number of predictions made by the classification model.

Also, the sensitivity values were obtained. Sensitivity (Equation 13) is the the ratio of actual Down cases that got predicted as Down according to the classification model to the total number of actual Down cases.

$$\text{Accuracy} = \frac{TD + TU}{TD + FD + TU + FU} \tag{12}$$

$$\text{Sensitivity} = \frac{TD}{TD + FU} \tag{13}$$

where TD is the situation where model has predicted Down and the actual is also Down, FD is the situation where model has predicted Down but the actual is Up, FU is the situation where model has predicted Up but the actual is Down and TU is the situation where model has predicted Up and the actual is also Up.

2.5.1 Random Forest

Random forest is a classification technique which is based on the concept of ensemble learning. The random forest grows multiple decision trees which are merged together for a more accurate prediction.

2.5.2 Support Vector Machine

Support vector machines are supervised machine learning algorithms. The classifier is formally defined by a separating hyperplane. Given the labelled training data this algorithm outputs an optimal hyperplane that categorizes new observations.

2.5.3 K-Nearest Neighbour

K-nearest neighbour is one of the simplest machine learning algorithms that come under supervised learning. It is considered a lazy learner algorithm because it does not learn from the training set immediately but it stores the data set and at the time of classification, it performs an action on the data set.

The analysis of this research study was done using R programming language and RStudio. Functions were created for clustering and classification using R programming language.

2.6 Forecasting Algorithm

The overall methodology used can be presented as an algorithm.

2.6.1 Algorithm 1: Forecasting the Closing Price Direction based on Candlestick Chart Properties

Input:
Hourly OHLC prices for the time points which have a trading volume greater than zero
Data Preprocessing:
Predictor Variables (X_t):
Create variables to represent CPD, USL, BL and LSL
Create variables to represent the OHLC prices, CPD, USL, BL and LSL upto three lags
Target Variable:
Closing price direction at the next immediate time point (FCPD)
Preliminary Analysis:
Data Visualizations and Contingency Tables
Clustering the Candlestick Properties (USL, BL, LSL):
Write functions using R programming language to the optimal number of clusters
Redo
For each method fuzzy K-means, K-means and partition around medoids clustering
    Identify the highest silhouette index and the corresponding number of clusters
End
Identification of the best clustering method:
Compare the highest silhouette indexes under each of the above three clustering methods
Select the highest silhouette index among those three methods
Consider the corresponding clustering method as the most appropriate clustering method
Perform clustering based on the selected method and the optimal number of clusters
Classification of the Future Close Price Direction in hour t+1:
Create functions using R programming language for each of the three methods
Redo
Separately for each cluster and also for the whole data set without considering clustering
    Redo
    For each predictor set (X_t)
        Perform the classification using RF, SVM and KNN
        Use k-fold cross-validation technique
        Obtain the predicted values for the test set
    End
End
Derive confusion matrix and get evaluation metrics
Discuss the effectiveness of the used predictors in predicting the target variable

3 Results and Discussion

The data preprocessing and creation of the dataset was done as discussed in section 2, Methodology. Then, the preliminary analysis was done.

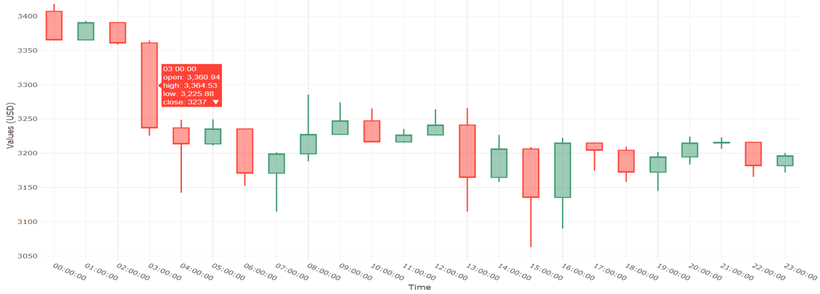

Fig. 2: The hourly candlestick charts for a single day (2022.01.07)

Fig. 2 shows the hourly candlestick charts for a single day. There are red candlesticks which indicate that the open price is higher than the close price in that hour and green candlesticks which indicate that the open price is lower than the close price in that hour. The candlesticks do not show any systematic pattern over time.

Fig. 3: The hourly candlestick charts for the considered date and time range

Fig. 3 shows the hourly candlestick charts from 2022.01.01 00:00 to 2022.12.31 23:00. The candlestick charts show higher values between January 2022 to June 2022 compared to other time ranges.

3.1 Results of Clustering

The optimal number of clusters under each of the fuzzy K-means clustering, K-means clustering and partition around medoids clustering were obtained. The best clustering method was found by comparing the values of silhouette indexes at the optimal number of clusters under each method.

Table 1: The table of the optimal number of clusters and respective silhouette index under each considered clustering method

Method	Optimal number of clusters	Silhouette index
Fuzzy K-means clustering	2	0.787
K-means clustering	2	0.603
Partition around medoids clustering	2	0.684

Based on Table 1, it was identified that the fuzzy K-means clustering method gave the highest silhouette index while the K-means clustering gave the lowest silhouette index. So, it was decided to use the fuzzy K-means clustering method as the best clustering method because it gave the highest silhouette index at its optimal number of clusters.

The variables USL, BL and LSL of hourly candlestick charts were clustered based on the fuzzy K-means clustering method and its optimal number of clusters which was identified as two. The behaviour of the two clusters can be shown as in Fig. 4.

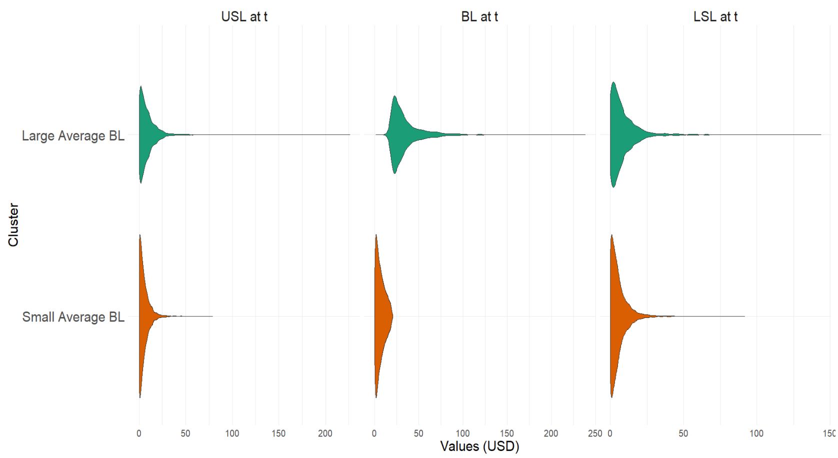

Fig. 4: The cluster-wise violin plots for the variables used in clustering

Fig. 4 shows the distribution of USL, BL and LSL inside the two clusters. The USL and LSL inside the two clusters behave in a similar manner while the BL behaves differently. So the two clusters were named as Large Average BL cluster and Small Average BL cluster.

3.2 Results of Classification

Parameter tuning was done using 10-fold cross-validation with each predictor variable set on each train data set based on clustering and ignoring clustering. The classification was done using RF, SVM and KNN methods by tuning the parameters.

Table 2: The table of accuracy values for each considered predictor variable set and classification method under clustering based method and clustering ignored method

Predictor variables	Approach	Classification method		
		RF	SVM	KNN
USL, BL, LSL and CPD at t	Clustering based	53.44%	53.49%	51.32%
	Clustering ignored	53.77%	52.50%	52.45%
USL, BL and LSL at t	Clustering based	49.94%	49.08%	51.55%
	Clustering ignored	50.37%	47.27%	49.74%
Open, High, Low, Close and CPD at t	Clustering based	53.95%	53.15%	50.29%
	Clustering ignored	**53.89%**	53.31%	52.56%
Open, High, Low and Close at t	Clustering based	50.74%	50.74%	51.15%
	Clustering ignored	50.89%	50.20%	50.09%
USL, BL, LSL and CPD at t, t-1	Clustering based	52.75%	53.72%	51.15%
	Clustering ignored	52.91%	52.56%	50.89%
USL, BL and LSL at t, t-1	Clustering based	49.94%	48.22%	50.52%
	Clustering ignored	50.95%	47.84%	51.81%
Open, High, Low, Close and CPD at t, t-1	Clustering based	52.12%	52.81%	51.89%
	Clustering ignored	52.50%	51.81%	52.45%
Open, High, Low and Close at t, t-1	Clustering based	51.49%	50.74%	52.23%
	Clustering ignored	50.26%	49.91%	52.10%
USL, BL, LSL and CPD at t, t-1, t-2	Clustering based	53.38%	53.72%	51.95%
	Clustering ignored	52.96%	52.27%	51.76%
USL, BL and LSL at t, t-1, t-2	Clustering based	51.66%	49.71%	50.06%
	Clustering ignored	48.59%	48.65%	52.27%
Open, High, Low, Close and CPD at t, t-1, t-2	Clustering based	52.41%	52.92%	51.26%
	Clustering ignored	52.04%	51.99%	49.45%
Open, High, Low and Close at t, t-1, t-2	Clustering based	50.17%	50.17%	52.63%
	Clustering ignored	51.64%	49.86%	49.63%
USL, BL, LSL and CPD at t, t-1, t-2, t-3	Clustering based	53.09%	**54.24%**	50.80%
	Clustering ignored	53.66%	52.22%	52.33%
USL, BL and LSL at t, t-1, t-2, t-3	Clustering based	52.06%	49.60%	50.29%
	Clustering ignored	48.01%	48.13%	50.55%
Open, High, Low, Close and CPD at t, t-1, t-2, t-3	Clustering based	52.06%	53.61%	52.29%
	Clustering ignored	51.70%	52.85%	50.78%
Open, High, Low and Close at t, t-1, t-2, t-3	Clustering based	50.00%	49.60%	50.97%
	Clustering ignored	50.89%	50.66%	50.83%

Table 3: The table of sensitivity values for each considered predictor variable set and classification method under clustering based method and clustering ignored method

Predictor variables	Approach	Classification method		
		RF	SVM	KNN
USL, BL, LSL and CPD at t	Clustering based	53.05%	56.55%	51.92%
	Clustering ignored	54.58%	54.22%	53.37%
USL, BL and LSL at t	Clustering based	51.69%	63.43%	50.00%
	Clustering ignored	52.17%	66.99%	51.93%
Open, High, Low, Close and CPD at t	Clustering based	51.13%	48.53%	49.55%
	Clustering ignored	53.98%	56.75%	51.45%
Open, High, Low and Close at t	Clustering based	51.35%	68.28%	51.02%
	Clustering ignored	52.29%	**81.20%**	48.19%
USL, BL, LSL and CPD at t, t-1	Clustering based	66.25%	59.26%	52.82%
	Clustering ignored	63.37%	65.90%	53.37%
USL, BL and LSL at t, t-1	Clustering based	51.13%	66.48%	52.14%
	Clustering ignored	52.53%	77.71%	54.58%
Open, High, Low, Close and CPD at t, t-1	Clustering based	59.71%	**69.07%**	51.69%
	Clustering ignored	60.00%	63.61%	49.16%
Open, High, Low and Close at t, t-1	Clustering based	51.58%	67.16%	51.24%
	Clustering ignored	50.00%	79.52%	50.60%
USL, BL, LSL and CPD at t, t-1, t-2	Clustering based	61.74%	57.90%	52.60%
	Clustering ignored	61.69%	63.01%	54.46%
USL, BL and LSL at t, t-1, t-2	Clustering based	53.05%	68.96%	48.76%
	Clustering ignored	52.53%	64.82%	51.81%
Open, High, Low, Close and CPD at t, t-1, t-2	Clustering based	59.71%	65.01%	51.81%
	Clustering ignored	59.52%	66.39%	49.28%
Open, High, Low and Close at t, t-1, t-2	Clustering based	50.34%	62.19%	50.90%
	Clustering ignored	52.77%	79.52%	48.19%
USL, BL, LSL and CPD at t, t-1, t-2, t-3	Clustering based	60.27%	63.88%	48.53%
	Clustering ignored	63.01%	62.29%	55.06%
USL, BL and LSL at t, t-1, t-2, t-3	Clustering based	53.16%	59.48%	49.66%
	Clustering ignored	52.77%	71.45%	54.70%
Open, High, Low, Close and CPD at t, t-1, t-2, t-3	Clustering based	55.76%	68.62%	53.39%
	Clustering ignored	57.83%	64.46%	50.96%
Open, High, Low and Close at t, t-1, t-2, t-3	Clustering based	49.10%	60.84%	48.08%
	Clustering ignored	53.61%	71.57%	47.11%

According to Table 2, it can be identified that the accuracy for the considered classification methods namely RF, SVM and KNN under clustering based method and clustering ignored method for all the predictor variable sets provided similar results for the accuracy value and it was around 50%.

According to Table 3, it can be identified that the sensitivity for the considered classification methods namely RF, SVM and KNN for all the predictor variable sets were around 50% to 70% in the clustering based method and around 50% to 80% in the clustering ignored method. The highest sensitivity is recorded under clustering ignored method and it was recorded as 81.20%.

The highest accuracy values and sensitivity values obtained under the clustering based method and clustering ignored method are stated as follows,

Clustering Based Method

Highest accuracy: 54.24%
Predictor variable set: USL, BL, LSL and CPD at t, t-1, t-2, t-3
Classification method: SVM

Highest sensitivity: 69.07%
Predictor variable set: Open, High, Low, Close and CPD at t, t-1
Classification method: SVM

Clustering Ignored Method

Highest Accuracy: 53.89%
Predictor Variable Set: Open, High, Low, Close and CPD at t
Classification Method: RF

Highest sensitivity: 81.20%
Predictor variable set: Open, High, Low and Close at t
Classification method: SVM

Based on the findings of this research study it can be stated that the proposed predictor variable sets which were created based on the candlestick chart properties including open, high, low and close prices, CPD, USL, BL and LSL within a short time horizon and the proposed methodology which was based on clustering and ignoring clustering is not much successful in forecasting the short term hourly closing price direction.

However, the accuracy could be increased slightly by clustering. Also, the accuracy was slightly higher in the situations where CPD was included in the predictor variable set than the situation where CPD was not included in the predictor variable set under consideration.

Considering the sensitivity values, it can be stated that there is a possibility to gain a higher sensitivity based on the SVM with Open, High, Low and Close at t in the clustering ignored method. Hence, this approach is performing better in predicting the Down cases to the total number of actual Down cases.

4 Conclusion

The proposed algorithm and the predictor variable sets which were created considering the hourly candlestick charts within a short time horizon are not considerably successful in accurately forecasting the Ethereum short term price direction. The accuracy was around 50% for all the considered predictor variable sets under RF, SVM and KNN in the clustering-based method and clustering ignored method. However, this algorithm was successful and was able to increase the accuracy at some situations. Based on the findings it can be stated that clustering improves the accuracy slightly and including the CPD with the predictor variable sets under consideration can increase the accuracy slightly. Furthermore, the proposed approach is performing better in predicting the Down cases to the total number of actual Down cases when clustering is ignored and SMV is used with Open, High, Low and Close at t.

References

Barnwal, A., Bharti, H. P., Ali, A., and Singh, V. (2019) Stacking with neural network for cryptocurrency investment, *New York scientific data summit (NYSDS)*.

Chen, M., Narwal, N., and Schultz, M. (2019) Predicting price changes in ethereum. *International Journal on Computer Science and Engineering (IJCSE)*.

Cohen, G. (2021) Optimizing candlesticks patterns for bitcoin's trading systems. *Review of Quantitative Finance and Accounting,* 57(3):1155–1167.

CoinMarketCap, *Today's cryptocurrency prices by market cap (2023)*, Available at: https://coinmarketcap.com/(Assessed on: 12 February 2023).

Derbentsev, V., Datsenko, N., Stepanenko, O., and Bezkorovainyi, V. (2019). Forecasting cryptocurrency prices time series using machine learning approach, *SHS Web of Conferences 2019*, 65:02001.

Ethereum, *Ethereum: What is ethereum?, the foundation for our digital future*, Available at: https://ethereum.org/en/ (Assessed on: 12 February 2023).

Ho, K.H., Chan, T.T., Pan, H., and Li, C. (2021). Do candlestick patterns work in cryptocurrency trading? , *IEEE International Conference on Big Data (Big Data) 2021*, 4566–4569.

Iqbal, M., Iqbal, M., Jaskani, F., Iqbal, K., and Hassan, A. (2021). Timeseries prediction of cryptocurrency market using machine learning techniques, *EAI Endorsed Transactions on Creative Technologies*, 8(28).

Liu, J. and Serletis, A. (2019). Volatility in the cryptocurrency market, *Open Economies Review*, 30:779–811.

Nison, S. (1991). *Japanese candlestick charting techniques*, New York Institute of Finance.

Nison, S. (2001). *Japanese candlestick charting techniques: a contemporary guide to the ancient investment techniques of the Far East.*, Penguin.

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

Pabuccu, H., Ongan, S., and Ongan, A. (2023). Forecasting the movements of bitcoin prices: an application of machine learning algorithms, *arXiv*.

Parikh, H., Panchal, N., and Sharma, A. (2022). Cryptocurrency price prediction using machine learning, *Proceedings of 6th International Conference on Advance Computing and Intelligent Engineering: ICACIE 2021*, 275–285.

Yamak, P. T., Yujian, L., and Gadosey, P. K. (2019). A comparison between arima, lstm, and gru for time series forecasting, *Proceedings of the 2019 2nd International Conference on Algorithms, Computing and Artificial Intelligence*, 49–55.