STOCK PRICE FORECASTING

NZDUSD,H1 0.59074 0.59084 0.59034 0.59056
0.64915
0.64810
0.64705
0.64600
0.64495
0.64390
0.64285
0.64180
0.64075
0.63970
0.63865
0.63760
Stoch(5,3,3) 58.9606 71.7326
100
75
25
0
10 Feb 2020
10 Feb 19:00
11 Feb 03:00
11 Feb 11:00
11 Feb
2020.02.12 04:00
2020.02.12 14:00
12 Feb 19:00
13 Feb 03:00
13 Feb 11:00
13 Feb 19:00
14 Feb 03:00
14 Feb 11:00
14 Feb 19:00
17 Feb 03:00
17 Feb 11:00
17 Feb 19:00
18 Feb 03:00

Author of report: PIO MPAKANY

Email: piompakanyi@gmail.com

University of RWANDA

Year of publication 2024

ABSTRACT

The seminar on stock price forecasting aimed to predict stock prices using historical data. The research came to explore result on the question highlighted by using time series analysis and other statistical models, all of them will be performed in R-programming as statistical package. In brief stock price forecasting is essential content in financial market and academic modules where different scholars, stock exchange institutions, analyst, scientist, traders always putting your eyes on how the trends stock price fluctuating, this led them to ask yourself can historical data predictfuture stock movement? The result of this question will be outlining in the section of findings as part that will explain in detail answer of research question. Kindly, I advise you to continue to read this report because you will get more information on how future stock movement will be based on historical data.

INTRODUCTION

World economy today is modeled on how stock price exchange is held, so mainly stock price exchange is mechanism for exchange and trading of stock but that can’t be occur without liquidity and information on the stock market. As an important part of a country's economy, the stock market provides a financing and investment environment for the country's companies and investors. Stock price forecasting, also known as stock market prediction or equity forecasting, is the process of using various methods and models to predict the future movements of stock prices. Investors and traders seek to forecast stock prices to make informed decisions about buying, selling, or holding stocks, they are several approaches are used for stock price forecasting which includes technical analysts study historical price charts, trading volume, and various technical indicators to identify patterns that may indicate future price movements, fundamental analysts is other approach which evaluate a company's financial health, earnings, dividends, and other relevant factors to estimate its intrinsic value also quantitative analysts is other approach use mathematical models and statistical techniques to analyze historical data and identify patterns, other approach is Econometric models incorporate economic indicators and macroeconomic factors to forecast stock prices. Which deals with variables such as interest rates, inflation, and GDP growth are considered to understand how broader economic trends may influence stock markets. These approaches are one’s advanced approaches that can used to predict the future movement of stock prices but all of them we know that future is uncertainty, that leads stock prediction difficult due to the complexity of financial markets, which are influenced by numerous unpredictable factors, including economic data, geopolitical events, investor sentiment, and unforeseen news and thereby may not lead to an accurate prediction of the stock prices(Jishag et al., 2020) .Now let me focus on historical data as a valuable input for investment analysis, successful investing requires a comprehensive approach that considers current market conditions, fundamental analysis, and an awareness of potential future stock movement. Investors often use a combination of historical data, technical analysis, and other tools to make more informed decisions. In this research, my goal was to develop a model that accurately predicts future stock market trends with minimal error and the highest possible precision, based on historical data. The forecasting stock price model relies predominantly on many elements: sentiment analysis and historical data, autoregressive integrated moving average (ARIMA).

Sentiment analysis involves assessing text and assigning scores based on three categories: negative, neutral, and positive. Simultaneously, the historical data analysis component forecasts a function derived from the stock prices in preceding years, autoregressive integrated moving average (ARIMA) is models use a combination of autoregressive and moving average terms to capture the temporal dependencies in stock prices. (https://medium.com/swlh/stock-price-prediction-using-sentiment-analysis-and-historical-stock-data-587488db8576). For now, my emphasis in this research will be on delving into historical data analysis, which is one of the advanced methods used to predict future stock movements. Further details on this analysis will be covered in the upcoming sections.

RELATED LITERATURE

According to (Jishag et al., 2020) Stock market prediction involves analyzing forthcoming changes in a company's stock prices. Numerous studies have explored this field, with some concentrating on enhancing prediction accuracy through sentiment analysis of stock-related news, while others have targeted forecasting price differentials across various phases.
According to (Shynkevich et al., 2015)Capital market traders receive substantial textual data from trading systems, encompassing official announcements, analysts' recommendations, financial journals, discussion boards, and news feeds from news wire services. To amalgamate data from diverse sources into a novel dataset or feature set that imparts comprehensive insights into the factors impacting stock price fluctuations. This resultant dataset is subsequently employed as input for prediction models, enhancing the precision of predictions and enabling more informed investment decisions (Li et al., 2023), however I disagree with Li et al.'s assertion that certain data sources cannot be utilized in predicting stock price movements.

Research in finance has investigated the impact of multi-source and heterogeneous data on stock markets. This refers to the inclusion of diverse data from various sources like the stock market, foreign exchange market, and even the weather system. It encompasses structured data such as stock prices and trading volumes, along with unstructured data like stock news, announcements, and social network information. The efficient market hypothesis posits that information from diverse sources influences the stock market, while behavioral finance suggests that individual behaviors and motivations of traders explain, study, and predict financial markets' trends and the extent of price fluctuations(Guo, n.d.). For forecasting prices, the stock market has been a common application. Traditionally, researchers often looked into utilizing the Autoregressive Integrated Moving Average (ARIMA) model, a traditional time series model, to predict stock price changes. However, with advancements in science and technology, machine learning models have emerged as a new tool for stock price prediction. Similar to predictions for traditional financial investment products like stocks, there is a growing interest in applying machine learning models for predicting bitcoin prices instead of relying solely on traditional time series models.

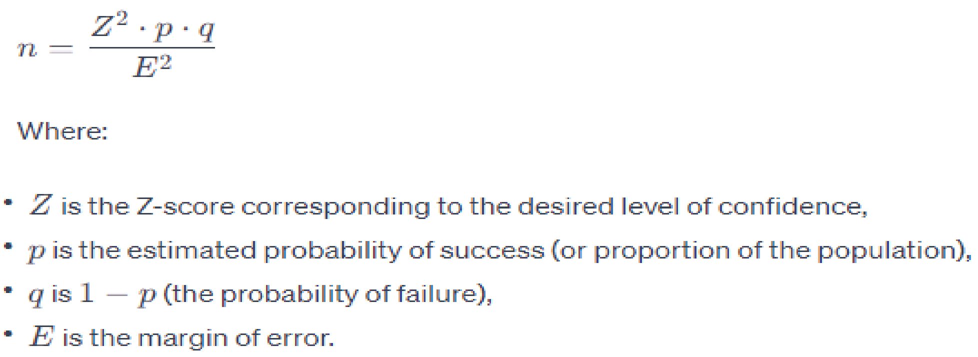
Researchers typically approach predictions of stock prices through machine learning from two perspectives. The first involves direct price prediction using regression models, while the second perspective utilizes classification models to predict changes in bitcoin prices. In contrast to direct price predictions, an increasing number of studies are concentrating on predicting changes in stock or bitcoin prices, aiming to enhance experimental performance(Wang & Yan, 2023). The stock market is characterized by its intricate time series nature with dynamic features. Following the market opening, there is significant dynamic trading in stocks, leading to corresponding changes in stock prices(Ji et al., 2021). To address the identified research gap, this researcer proposes a novel stock price prediction approach utilizing deep learning technology. The method integrates Doc2Vec, stacked auto-encoder (SAE), wavelet transform, and long short-term memory (LSTM) model. Text information from social media is crucial for capturing investor sentiment and enhancing stock price predictions.

The approach classifies prediction features into financial and text categories, employing established financial features and utilizing deep learning to extract text features from social media. Doc2Vec is applied to preserve semantic information and word relationships in social media documents, overcoming limitations of traditional methods. Additionally, SAE reduces the dimension of text feature vectors, achieving a balance with financial features. Wavelet transform is employed to transform the target variable stock price and eliminate random noise in time series data. The input features, consisting of stock finance features and extracted text features, are then used in conjunction with LSTM for stock price prediction(Ji et al., 2021).

METHODOLOGY

Quantitative research design

This study aims to support the analysis of a stock exchange dataset using R-programming. The design will involve selecting a sample of 384 variables from the entire population of 1285 variables. In this case, the probability of selecting variables (*p*^) and the probability of unselected variables (*q*^) is both 50%. The level of significance (*Zc*) is 1.96, and the margin of error (*E*)is 5%. Substituting these values into the formula,

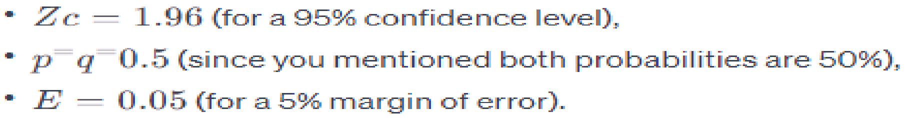


$$n = \frac{Z^2 \cdot p \cdot q}{E^2}$$

Where:

- Z is the Z-score corresponding to the desired level of confidence,
- p is the estimated probability of success (or proportion of the population),
- q is $1 - p$ (the probability of failure),
- E is the margin of error.

Given the information from the dataset:



- $Zc = 1.96$ (for a 95% confidence level),
- $p = q = 0.5$ (since you mentioned both probabilities are 50%),
- $E = 0.05$ (for a 5% margin of error).

Substituting these values into the formula, n= **384 *variable*.** This sample size will be utilized for the analysis of the stock exchange dataset.

ANALYSIS AND RESULTS

Summary of the central tendency and distribution

	Variable	Mean	Median	Mode
open	open	34.9902	10.08	10.1
high	high	35.656	10.11	10.1
low	low	34.3012	10.005	10.1
close	close	34.9644	10.08	10.1
adjclose	adjclose	34.4832	10.061	10.1
volume	volume	758602	84060	0

Interpretation:

- ✓ The volume variable has a mean of approximately 758,602.19 and a median of 84,060.00, suggesting a potentially skewed distribution with a few high values.
- ✓ The mode of volume is 0.0, indicating that this value occurs most frequently. This might suggest a substantial number of zero values in the dataset.

Overall, these statistics provide a summary of the central tendency and distribution characteristics of the variables in your dataset.

The correlation matrix

Matrix	**open**	**high**	**low**	**Close**	**adjclose**	**volume**
open	1	0.99988	0.9998775	0.999659875	0.99961	0.006
high	0.99988	1	0.9998751	0.999861818	0.99983	0.00661
low	0.99988	0.99988	1	0.999858982	0.99982	0.00571
close	0.99966	0.99986	0.999859	1	0.99997	0.00628
adjclose	0.99961	0.99983	0.9998156	0.999965348	1	0.00655
volume	0.006	0.00661	0.0057086	0.006284837	0.00655	1

Interpretation the correlation matrix

- ✓ The diagonal elements of the matrix (from top-left to bottom-right) represent the correlation of each variable with itself, which is always 1.
- ✓ The off-diagonal elements represent the pairwise correlations between different variables.

Correlation between Variables

- ✓ The correlation between open and high is very close to 1 (0.99988), indicating an extremely strong positive correlation between these two variables.
- ✓ Similar strong positive correlations are observed between open and low, open and close, open and adjclose, all of which are very close to 1.
- ✓ The same patterns of strong positive correlation are observed between high and the other variables (low, close, adjclose).
- ✓ Volume has very low correlation values with open, high, low, close, andadjclose (all around 0.006), indicating weak linear relationships.

Interpretation of Correlation with Volume

- ✓ The low correlation values between 'volume' and the price-related variables (open, high, low, close, adjclose) suggest that there is no strong linear relationship between trading volume and these price variables.

Analysis is based on the following mathematical regression model:

$$\text{Open } (Y_t) = \mathbf{b0} + \mathbf{b1}\text{high}(\mathbf{x}) + \mathbf{b2}\text{low}(\mathbf{x2}) + \mathbf{b}3\text{close}(\mathbf{x3}) + \text{b4adjclose}(\mathbf{x4}) + \text{b}\mathbf{5}\text{volume(x5)} + \epsilon t$$

Where t = date

b0 = intercept point

b_t = coefficients of different independent variable mentioned above.

ϵt = residuals

lm(formula = open ~ high + low + close + adjclose + volume, data = stock.e

Residuals				
Min	1Q	Median	3Q	Max
-16.8734	-0.093	0.0218	0.0979	15.824
Coefficients	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-2.82E-02	1.19E-02	-2.368	0.0179 *
high	8.12E-01	7.81E-03	103.896	< 2e-16 ***
low	8.33E-01	8.02E-03	103.918	< 2e-16 ***
close	-5.92E-01	1.54E-02	-38.381	< 2e-16 ***
adjclose	-5.30E-02	1.34E-02	-3.958	7.62e-05 ***
volume	-1.57E-09	2.79E-09	-0.563	0.5734

Residual standard error: 0.9669 on 7775

Degrees of freedom Multiple R-squared: 0.9999,

Adjusted R-squared: 0.9999

F-statistic: 1.659e+07 on 5 and 7775 DF,

p-value: < 2.2e-16

Open (Yt) = -2.816+8.11high(x) +8.33low(x2) -5.917close(x3) -5.304adjclose(x4) - 1.572volume(x5) +ϵt

Interpretation:

Coefficients:

- ✓ **Intercept:** The intercept (***b*0**) is approximately -0.02816. It represents the estimated 'open' value when all other predictor variables are zero.
- ✓ **high, low, close, adjclose:** The coefficients for these variables (***b*1, *b*2, *b*3, *b*4**) indicate the estimated change in open for a one-unit increase in the respective predictor, holding other predictors constant.
- ✓ **volume:** The coefficient for volume (***b*5**) is very small (-1.572e-09) and not statistically significant (p-value = **0.5734**). This suggests that 'volume' does not have a significant impact on predicting 'open' in this model.

Significance:

- ✓ high, low, close, and adjclose are highly significant (p<0.001), suggesting that these variables are strong predictors of 'open.'
- ✓ volume is not significant (p=0.5734), supporting the idea that it doesn't contribute significantly to predicting open in this model.

Residuals:

- ✓ The residuals (differences between actual and predicted values) have a minimum of -16.8734 and a maximum of 15.8240.
- ✓ The residual standard error is 0.9669, indicating the typical magnitude of the residuals.

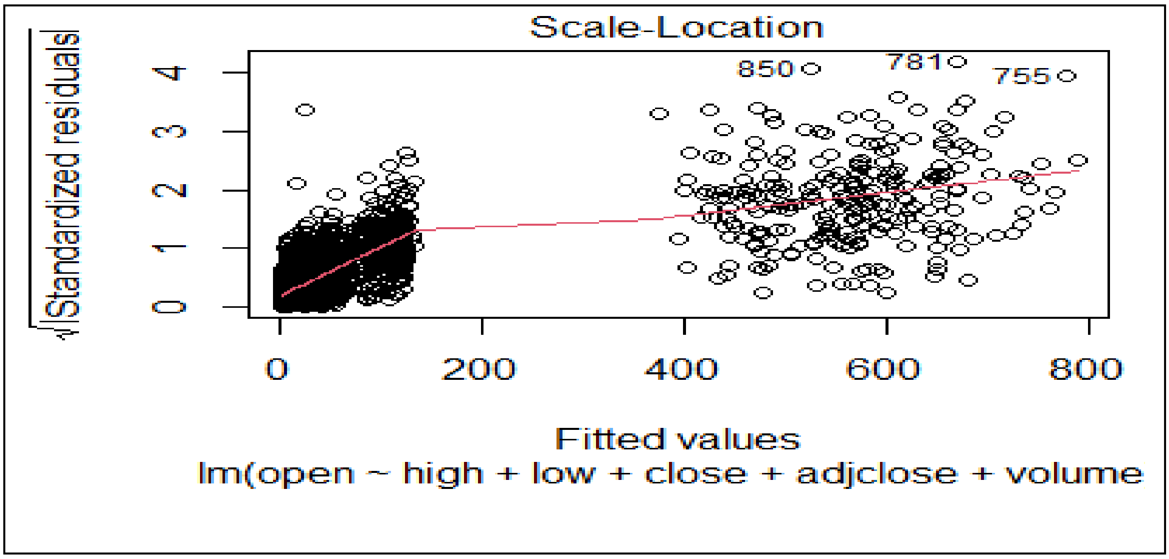
Model Fit:

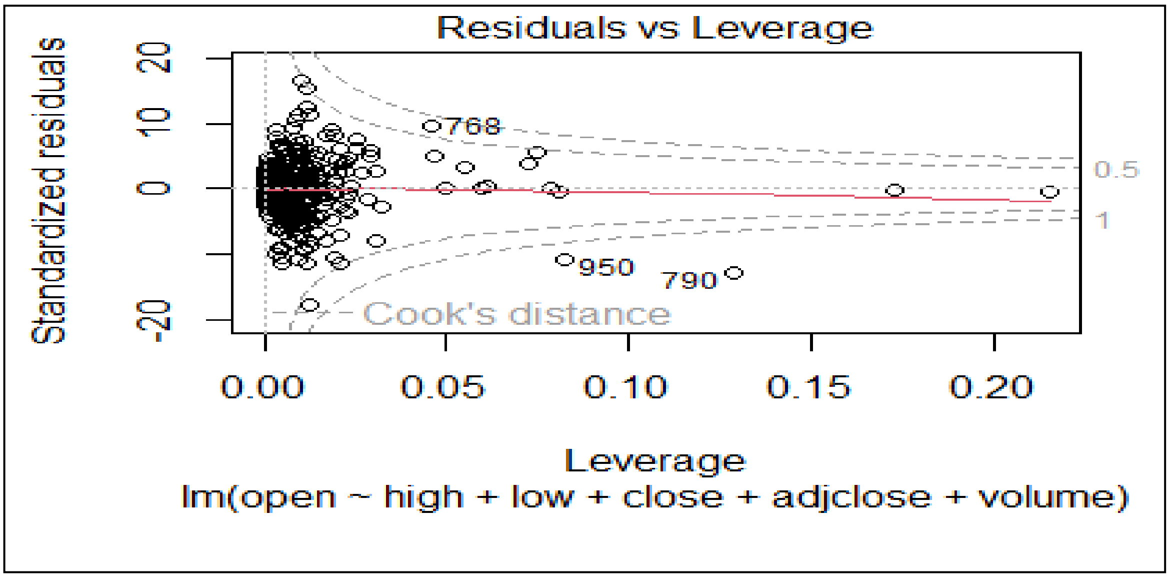
- ✓ *R*2 (Multiple R-squared): This is very close to 1 (0.9999), suggesting that the model explains a very high proportion of the variance in the open variable.
- ✓ Adjusted *R*2: It adjusts *R*2 for the number of predictors, and it is also very close to 1

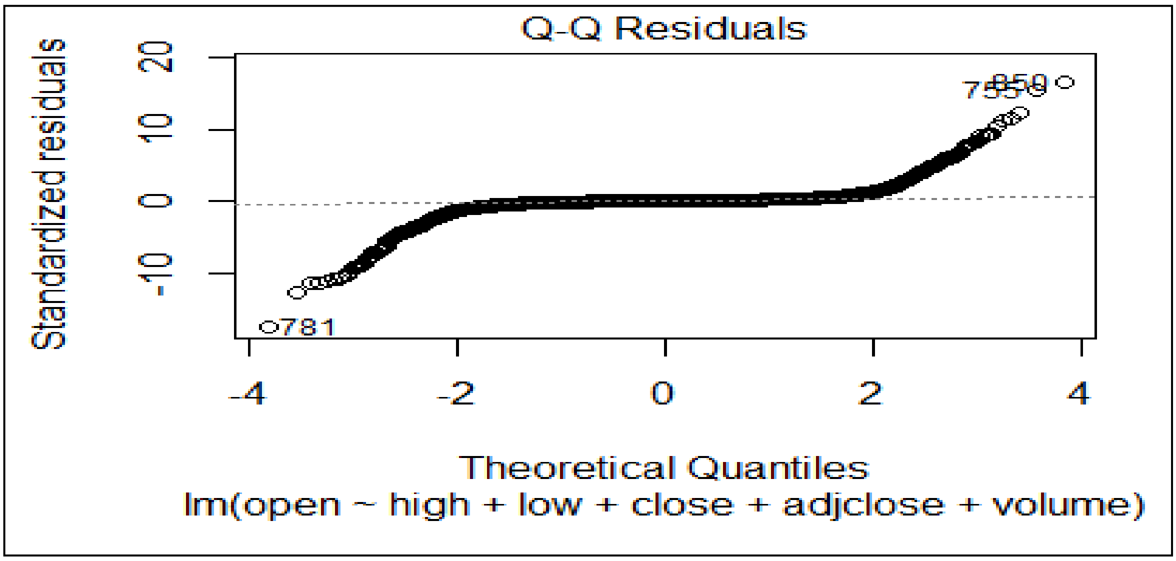
(0.9999).

✓ F-statistic: This is a test of the overall significance of the model. The very lowp-value ($<2.2e-16$) indicates that the model is statistically significant.

GRAPHICAL REPRESENTATION







CONCLUSION

The quantitative research design employed in this study focused on analyzing a stock exchange dataset using R-programming. The sampling strategy involved selecting a representative sample of 384 variables from the entire population of 1285 variables, with a probability of selection (p^) and unselected variables (q^) set at 50%. The statistical significance level (Zc) was established at 1.96, and the margin of error (E) was 5%, resulting in a sample size of 384 variables for the subsequent analysis. The analysis unveiled insightful findings about the central tendency and distribution of variables within the stock exchange dataset. Notably, the volume variable demonstrated a potentially skewed distribution, with a mean of approximately 758,602.19 and a median of 84,060.00. The mode of volume being 0.0 suggests a notable frequency of zero values in the dataset. Furthermore, the correlation matrix illustrated strong positive correlations between price-related variables (open, high, low, close, adjclose), while the correlation with volume was consistently low, indicating a weak linear relationship between trading volume and these price variables. The subsequent regression model reinforced the importance of historical data in predicting stock prices, with coefficients for high, low, close, and adjclose being highly significant. However, the volume variable was found to be statistically insignificant, implying a limited impact on predicting open prices. The model demonstrates a remarkable fit with the data, elucidating nearly all the variance in open stock prices. Notably, the inclusion of volume does not yield substantial additional information for predicting open prices within this model. This underscores the significance of historical data as a predominant contributor in forecasting stock prices, as highlighted by the aforementioned model. In essence, the historical trends play a more pivotal role in the accuracy of stock price predictions compared to the supplementary information provided by trading volume.

RECOMMENDATIONS:

By incorporating these recommendations, future analyses can enhance the robustness and effectiveness of stock price forecasting models, providing more accurate insights into market trends.

- ✓ Emphasize Historical Data: Given the significance of historical data in predicting stock prices, continue to prioritize the inclusion and analysis of past trends and patterns in forecasting models.
- ✓ Explore Additional Predictors: While high, low, close, and adjclose proved to be

strong predictors, consider exploring additional relevant variables that might enhance the predictive power of the model.

- ✓ Continuous Model Validation: Establish a routine for regularly updating and validating the forecasting model against actual stock prices to ensure its ongoing accuracy and reliability.
- ✓ Investigate Skewed Distribution in Volume: Explore the presence of a skewed distribution and frequent zero values in the volume variable. Investigate the potential impact on model performance and consider appropriate adjustments.
- ✓ Diversify Analysis Techniques: Explore advanced statistical techniques or alternative regression models to capture complex relationships within the dataset and improve the overall accuracy of stock price predictions.
- ✓ Collaborate with Domain Experts: Collaborate with financial analysts or domain experts to gain valuable insights into market dynamics and potential variables that may impact stock prices.
- ✓ Document Assumptions: Clearly document the assumptions underlying the regression model, acknowledging the significance of historical data and the limitations associated with the volume variable.

REFERENCES

Guo, Y. (n.d.). *Stock Price Prediction Using Machine Learning*.

Ji, X., Wang, J., & Yan, Z. (2021). A stock price prediction method based on deep learning technology. *International Journal of Crowd Science*, *5*(1), 55–72. https://doi.org/10.1108/IJCS-05-2020-0012

Jishag, A. C., Athira, A. P., Shailaja, M., & Thara, S. (2020). Predicting the stock market behavior using historic data analysis and news sentiment analysis in r. *Advances in Intelligent Systems and Computing*, *1045*, 717–728. https://doi.org/10.1007/978-981-15-0029-9_56

Li, A., Wei, Q., Shi, Y., & Liu, Z. (2023). Research on stock price prediction from a data fusion perspective. *Data Science in Finance and Economics*, *3*(3), 230–250. https://doi.org/10.3934/dsfe.2023014

Shynkevich, Y., McGinnity, T. M., Coleman, S., & Belatreche, A. (2015). Predicting stock price movements based on different categories of news articles. *Proceedings - 2015 IEEE Symposium Series on Computational Intelligence, SSCI 2015*, 703–710. https://doi.org/10.1109/SSCI.2015.107

Wang, Y., & Yan, K. (2023). Application of Traditional Machine Learning Models for Quantitative Trading of Bitcoin. *Artificial Intelligence Evolution*, 34–48. https://doi.org/10.37256/aie.4120232226

APPENDICES

```
#Load necessary
librarieslibrary(dplyr)
library(tidyr)
library(ggplot2)

# Check summary statistics and structure of the dataset
summary(stock.exchang)
str(stock.exchange)

# Correlation matrix
cor_matrix <- cor(stock.exchange[, c("open", "high", "low", "close", "adjclose", "volume")])

# Visualize correlation matrix using a heatmap
heatmap(cor_matrix, annot = TRUE)
# Create a scatterplot matrix
scatterplotMatrix(stock.exchang[, c("open", "high", "low", "close", "adjclose", "volume")])

# Multiple linear regression
lm_model <- lm(open ~ high + low + close + adjclose + volume, data =
stock.exchange)summary(lm_model)

# Visualize the regression results
plot(lm_model)
# Central tendencies: Mean, Median, Mode
means <- colMeans(stock.exchange[, c("open", "high", "low", "close", "adjclose",
"volume")])
medians <- apply(stock.exchange[, c("open", "high", "low", "close", "adjclose", "volume")],
2, median)
modes <- apply(stock.exchange[, c("open", "high", "low", "close", "adjclose", "volume")], 2,
function(x) { tbl <- table(x) mode_val <- as.numeric(names(tbl[tbl == max(tbl)]))
return(mode_val) })

# Display central tendencies
central_tendencies <- data.frame(Variable = names(means), Mean = means, Median =
medians, Mode = modes)
central_tendencies

# Histogram for 'open'
hist(stock.exchange$open, main = "Histogram of Open Prices", xlab = "Open Prices", col =
"lightblue", border = "black")

# Boxplot for 'open'
boxplot(stock.exchange$open, main = "Boxplot of Open Prices", ylab = "Open Prices", col =
"lightblue", border = "black")
```