CryptoForecast: Leveraging Data Science for Predicting Markets

Gwal Buddhadev[1], Tanisha Ghosh[2], Sneha Das[3], and Stephen Baraik[4]

[1,2,3,4]School of Mathematics, Applied Statistics and Analytics, NMIMS Deemed to be University, Navi Mumbai

[1]gwalbuddhadev@gmail.com

[2]Ghoshtanisha@gmail.com

[3]Snehadas1593@gmail.com

[4]stephenbaraik@gmail.com

Abstract

This study is devoted to the development and evaluation of various predictive models, including artificial neural networks and random forests, with the aim of forecasting the closing prices of four major cryptocurrencies: Bitcoin, Ethereum, Binance Coin, and Solana. The central objective is to identify the most accurate, reliable, and efficient model for predicting cryptocurrency price fluctuations. This study employs linear regression, artificial neural networks, random forests, and lightGBM to predict the closing price of four key cryptocurrencies: Bitcoin (BTC), Ethereum (SOL), Binance (BNB), and Ethereum (ETH) the next day. The findings indicate a notable trend: Random Forest and LightGBM generally demonstrate higher accuracy and reliability compared to ANN and Linear Regression approaches for all cryptocurrencies analyzed. In addition, this study emphasizes that the Random Forest approach is far ahead of the artificial neural network and linear regression approaches in predicting the closing values of all main cryptocurrencies. In turn, LightGBMA improved the efficiency of Ethereum, BNB, and Solana, highlighting its efficiency and strength in relation to predictive modeling. Such models, tested here with many other features and approaches, could be considered in further studies for possible improvements in the predictive accuracy. This study will be useful to traders and investors seeking more precise predictions in the cryptocurrency sphere to find value in employing these ensemble methods, particularly random forests.

Keywords: Machine Learning Models, Cryptocurrency Price Prediction, Data Science in Cryptocurrencies, Financial Time Series Forecasting, Ensemble Learning, Feature Engineering for Price Prediction.

1 Introduction

Liaw and Wiener, 2002 indicates that Random Forest is particularly suited for data classification and prediction, with promising results when working with complex data sets. Hamzaçebi et al., 2009 examined various applications of neural networks on time-series data comparing the effectiveness of each in terms of forecast accuracy and explaining pros and cons of each method for forecasting purposes. Alternatively, Pandey et al., 2023 employed machine-learning-based algorithms to improve financial time-series forecasting and subsequently predict stock values. Zhang, 2003 innovatively hybridized the ARIMA models with neural networks within a framework that proves to be more accurate than usual schemes. Murkute and Sarode, 2015 discussed the reliability of artificial neural networks in predicting stock prices, which also shows results of higher precision in forecasting. Wang, 2022 highlighted that deep learning is well suited for the prediction of share prices of the tech companies underlining its role in financial prediction. Dagur et al., 2023 have made people aware of state-of-the-art advancements in AI, blockchain, and computing, forging new paths toward security and technology. This is an attempt to use the most complex data science methodologies to predict closing prices for a large variety of cryptocurrencies. This research specifically deals with the application of Artificial Neural Networks and Random Forest algorithms to the Open, High, Low, and Close values of several different cryptocurrencies. Predictability performance in terms of high-accuracy value forecasts of cryptocurrencies is measured through strategic metrics such as RMSE, MAPE, and MBE. Therefore, this study contributes new knowledge by extending AI applications in financial markets, particularly in the volatile, dynamic, and fast-changing cryptocurrency sectors.

2 Preliminary

2.1 Linear Regression (LR)

According to S., 2023 As a kind of statistical analysis, linear regression establishes the ability to predict continuous responses in the presence of one or more independent variables. This method assumes a linear relationship between variables and forms an optimal line or hyperplane such that the sum of the squared residuals is at its minimum. The residual is the difference between the actual and predicted values. Linear regression was used to forecast the closing prices of four major cryptocurrencies (BTC), (SOL), (BNB), and (ETH). A linear regression model was trained to determine the best coefficients for relevant features to ensure the precise prediction of closing cryptocurrency prices. The basic equation for the LR is as follows:

$$Y = a + bX + \epsilon \tag{1}$$

2.2 Artificial Neural Networks (ANN)

According to Vasilj et al., 2024 ANNs are robust computational frameworks modeled upon the complex neural configurations of the human intellect. It is comprised of neurons that process complex information through a multilayer system of interconnected neurons. These networks are highly effective for dealing with complex data. The neurons perform a linear transformation followed by a nonlinear activation function, meaning that the model is very good at identifying intricate patterns with great accuracy. A conventional ANN design starts with an information intake layer, with one or more concealed layers, and ends with a final layer. Activation functions, such as the ReLU and sigmoid, provide model nonlinearity, which improves its performance. The backpropagation technique helps reduce errors further from the model by adjusting some connections between neurons using gradient descent, and the use of an optimizer such as Adam optimizes the weight update for maximum performance. Their flexibility makes ANNs suitable for a wide variety of applications including regression, classification, and time-series prediction. This makes them extremely useful for applications such as the accurate prediction of the closing prices of some major cryptocurrencies, such as Bitcoin, Solana, Binance Coin, and Ethereum.

2.3 Random Forest (RF)

As stated by Marsland, 2014 Regression with Random Forests is a popular and efficient ensemble algorithm based on the principle of many decision trees as a single model. Essentially, many decision trees are formed, and each is created based on a random sample of the training set and a random sample of features used for a particular tree. This is called bootstrapping, and this idea helps curb overfitting and improve the generalization ability of the model. Finally, the mean of all tree predictions yielded the overall prediction, increasing precision and stability. Random forests can comprehend and analyze complex patterns, absent data, and the importance of each feature; hence, they can be employed in many regression tasks.

2.4 LightGBM (LGB)

According to Koch, 2021 emphasizes that LightGBM is a highly regarded gradient boosting framework, celebrated for its impressive speed and efficiency in processing large datasets, which contributes to its widespread adoption. This framework is characterized by several noteworthy features: it employs gradient boosting, an ensemble methodology that sequentially constructs decision trees, allowing each tree to address the errors made by its predecessors. Unique to LightGBM is its leaf-wise tree growth, which generally enhances accuracy in comparison to the level-wise approach utilized by numerous other algorithms. Furthermore, the histogram-based algorithm facilitates faster training while minimizing memory usage through efficient

split proposals. The framework also includes regularization techniques to mitigate the risk of overfitting and can handle categorical features. In this study, LightGBM Regression was applied to forecast the closing prices of four prominent cryptocurrencies: Bitcoin (BTC), Solana (SOL), Binance Coin (BNB), and Ethereum (ETH).

3 Model Evaluation Metric

3.1 Mean Bias Error

Indicating whether the forecasts are often higher or lower than the actual values, MBE calculates the average bias in the model's predictions.

$$\frac{1}{n}\sum_{i=1}^{n}(z_i - \hat{z}_i) \tag{2}$$

3.2 Root Mean Square Error

Larger deviations are given more weight by RMSE, a statistic that measures the average size of the discrepancy between the anticipated and actual values.

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(z_i - \hat{z}_i)^2} \tag{3}$$

3.3 Mean Absolute Percentage Error

MAPE provides an indication of the accuracy of percentage mistakes by calculating the average absolute percentage difference between expected and actual numbers.

$$\frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{4}$$

4 Problem Formulation

4.1 Introduction to the Problem Area

As cryptocurrencies have been more widely accepted, it has been quite hard to establish their prices, especially considering their crazy fluctuation. This finding implies that an accurate price forecast is vital for investors, traders, and financial institutions. Nevertheless, despite the use of different models, the unpredictable nature of the cryptocurrency market renders these predictions unreliable.

4.2 Identification of the Research Gap

In order to predict cryptocurrency prices, researchers have put into practice many different machine learning models involving artificial neural networks and random forests. However, few comparisons are available regarding their ability to forecast the closing prices of cryptocurrencies the following day. Moreover, we do not examine the accuracy or reliability of these models when the market becomes highly volatile.

4.3 Problem Statement

The present paper is devoted to the construction and testing of Artificial Neural Networks and Random Forest predictive models, focusing on forecasting the closing price of cryptocurrencies the next day. This means that attention is focused on determining which model is more accurate, reliable, and efficient in handling the complexities of cryptocurrency price fluctuations.

4.4 Research Objectives

4.4.1 ***Primary Objective***

In this work, the performance of LR,ANN,RF,LGB models with respect to the prediction of the next day's closing prices of some selected cryptocurrencies was developed and checked.

4.4.2 ***Secondary Objectives***

- To check the effectiveness of the models using some of the important performance indicators, such as RMSE, MAPE and MBE.

- This flows directly into the examination of the adequacy of the models in different market conditions, especially with regard to high volatility, and the reliability of the models.
- This will identify which model is more time and resource economic while fulfilling the accuracy constraints on the prediction of results.

4.5 Significance of Study

It will be useful to the cryptocurrency price prediction field, as it compares in detail LR,ANN,RF,LGB models. This study will be useful to traders and investors in their decision-making processes by providing them with more accurate forecasts.

5 Solution Methodology

5.1 Data Collection

*5.1.1 **CryptoCurrency***

This research will focus on the four largest digital currencies, where in the discussion of their leader will revolve around Bitcoin, followed by the others on the list: Ethereum, Binance Coin, and Solana. These digital currencies are the most traded and considered the best-known and exchanged assets in the digital currency market.

*5.1.2 **Data Sources***

For this project, we are going to extract historical price data for these cryptocurrency tokens via Yahoo Finance API. These data consist of the historical OHLC prices and trading volumes.

*5.1.3 **Data Pre-processing***

Data preprocessing steps for all models include feature scaling. This ensures proper numerical stability as well as an identical contribution by the features. For the LR, RF, and LGB models, the features were normalized using a standard scalar. This improved the strength and performance of the model and prevented skewed results. The features of the ANN model were scaled between 0 and 1 to improve the convergence of the ANN during training, and the MinMaxScaler was used. To evaluate the ability of the model to generalize to unseen data, the dataset was divided

into two components: 80% was allocated for training and 20% was set aside for testing, allowing for a comprehensive evaluation of the model's capabilities.

5.2 Feature Engineering

Common variables amongst those that have been used with the models for feature engineering are H-L, the difference between the highest and lowest prices; O-C, the difference between the opening and closing prices; moving averages over 7, 14, and 21 days; and the 7-day rolling standard deviation. More specifically, RF and LGB also use technical indicators such as the 14-day RSI, EMA 12, EMA 26, Bollinger bands, MACD, together with its signal line, and 10-day momentum. The ANN model used the same fundamental features as those used in the linear regression model.

5.3 Model Evaluation Performance Metrics

The evaluation of the accuracy of the model will be based on the predictions calculated by these three major performance metrics (RMSE), (MAPE), and (MBE). These models should be tested against market conditions such as those that can occasionally be expected when the market becomes volatile. Computational Efficiency: The computational efficiency of both models was quantified using the duration of training and resource consumption.

Table 1: Performance Metrics for Different Models

Crypto	Metric	Linear Regression	ANN	Random Forest	LightGBM
Bitcoin	RMSE	400.94	0.04	229.45	278.13
	MAPE	0.68%	3.99%	**0.35%**	0.58%
	MBE	**-45.45**	-1.75	-7.21	-6.18
Ethereum	RMSE	28.50	0.02	13.91	**7.65**
	MAPE	0.75%	1.16%	0.35%	**0.25%**
	MBE	-2.72	-0.28	-0.53	-0.11
Binance	RMSE	5.04	0.02	2.50	**0.88**
	MAPE	0.74%	2.17%	0.36%	**0.13%**
	MBE	-0.71	-0.78	-0.12	0.03
Solana	RMSE	2.65	0.02	1.53	**0.62**
	MAPE	1.73%	3.96%	0.99%	**0.37%**
	MBE	-0.60	**2.86**	-0.08	-0.01

6 Results

Performance testing of the model is carried out with forecast evaluation approaches such as RMSE, MAPE, and MBE of linear regression, artificial neural network, random forest, and lightGBM. This study discusses four key cryptocurrencies: Bitcoin, Ethereum, Binance Coin, and Solana. In general, the Random Forest and LightGBM models consistently delivered the most accurate predictions across all cryptocurrencies. LightGBM, in particular, proved highly effective in reducing prediction errors compared with ANN and Linear Regression models.

7 Model Comparisons for Cryptocurrencies

7.1 Bitcoin

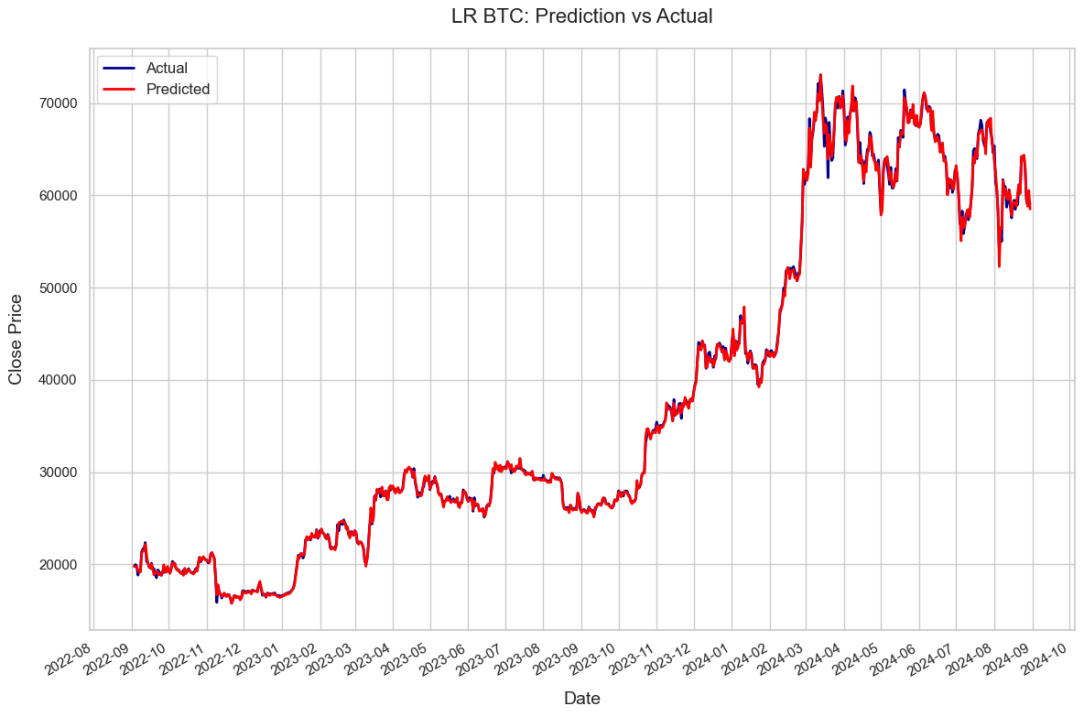

Figure 1: Predicted vs. Actual Closing Prices for Bitcoin using the LR Model.

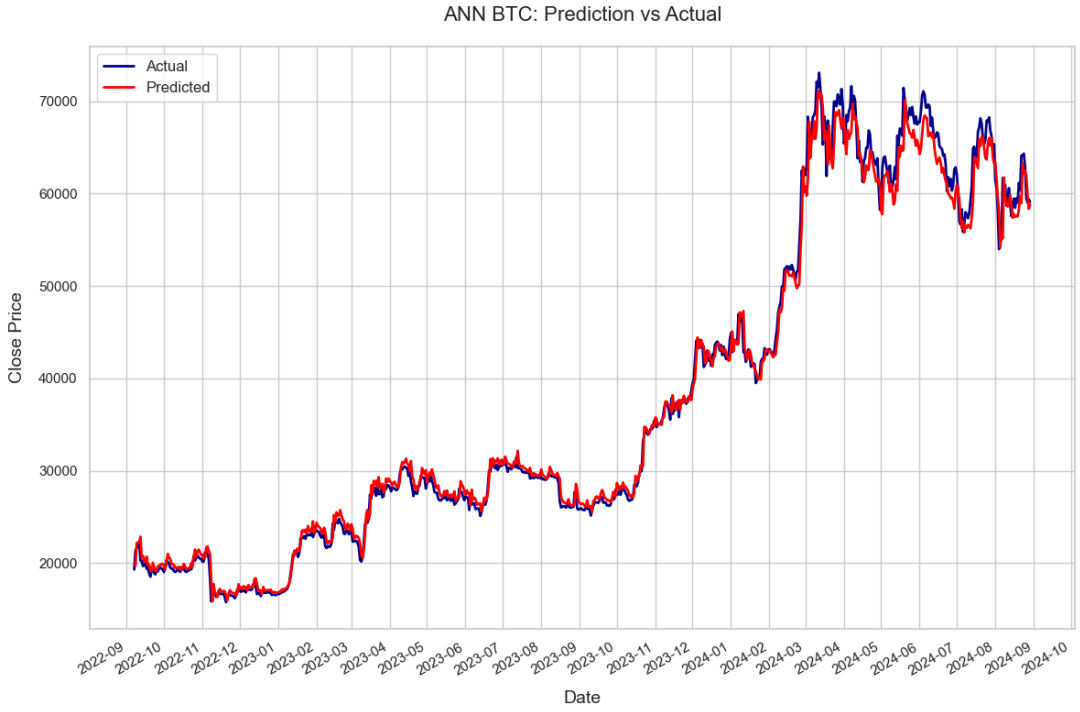

Figure 2: Predicted vs. Actual Closing Prices for Bitcoin using the ANN Model.

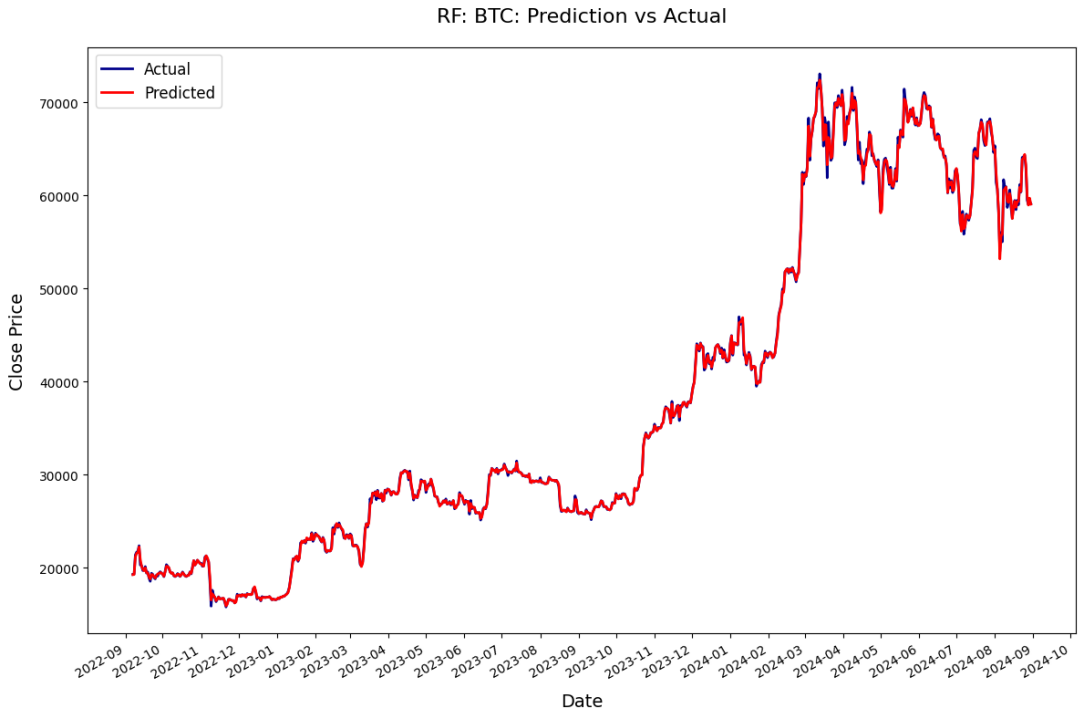

Figure 3: Predicted vs. Actual Closing Prices for Bitcoin using the RF Model.

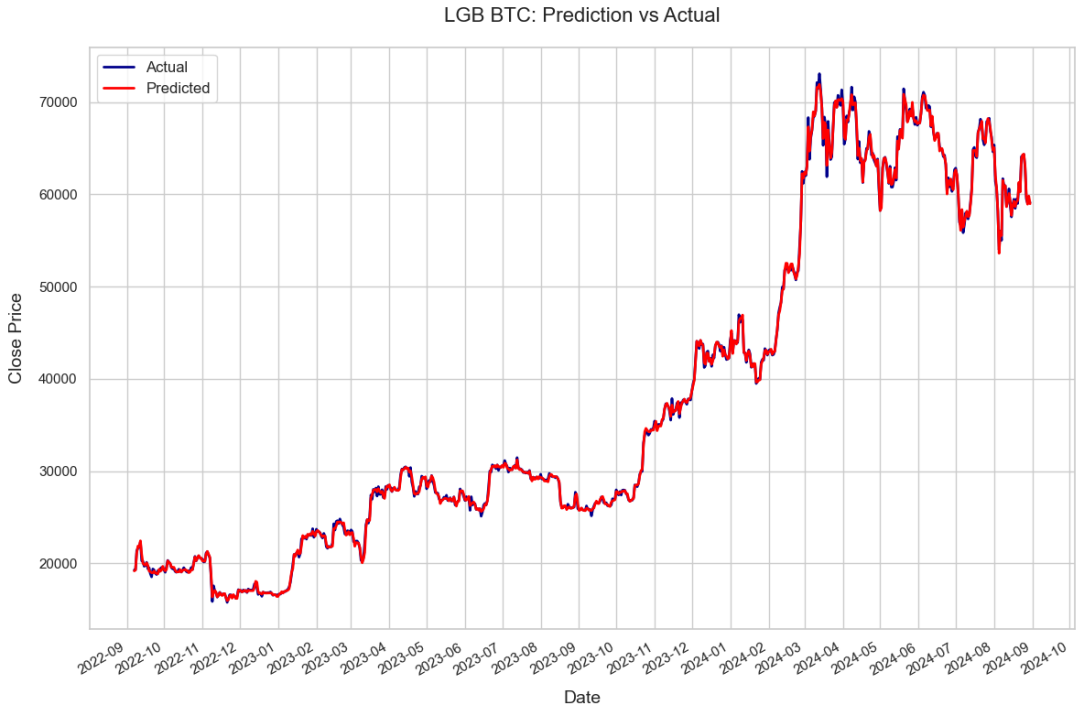

Figure 4: Predicted vs. Actual Closing Prices for Bitcoin using the LGB Model.

7.2 Ethereum

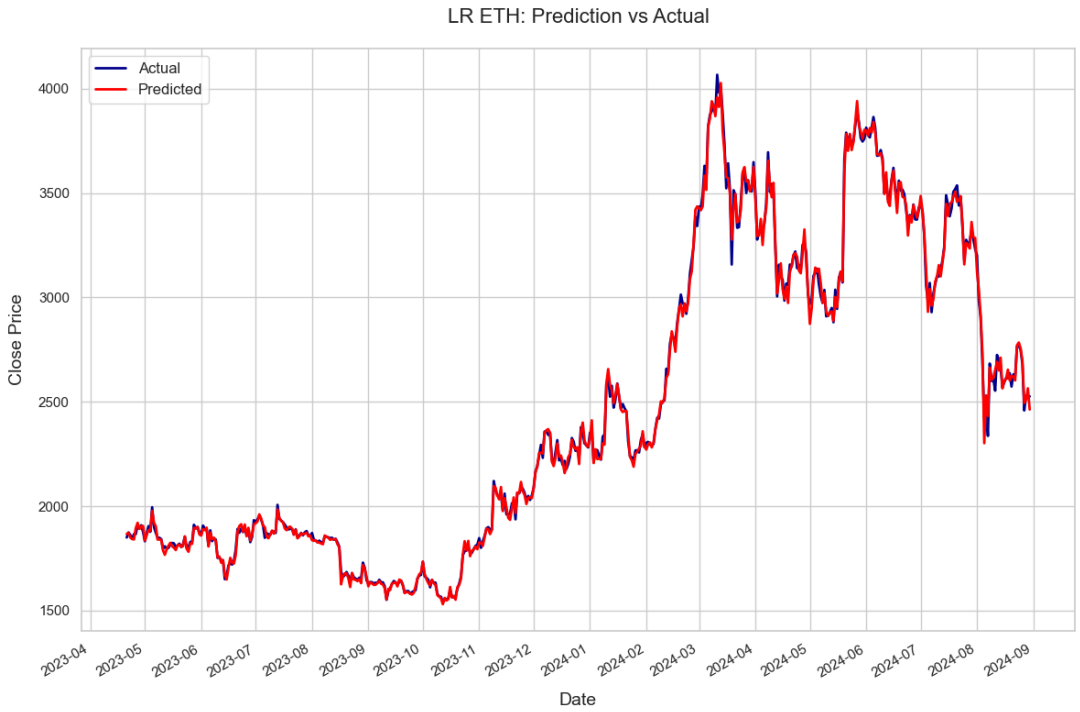

Figure 5: Predicted vs. Actual Closing Prices for Ethereum using the LR Model.

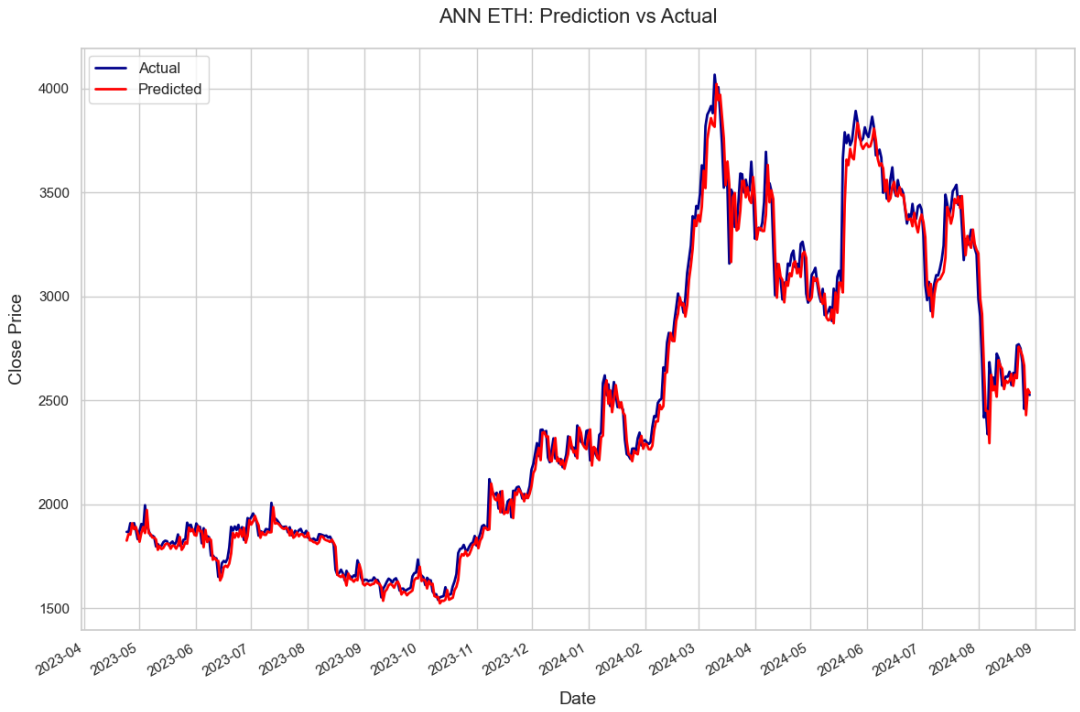

Figure 6: Predicted vs. Actual Closing Prices for Ethereum using the ANN Model.

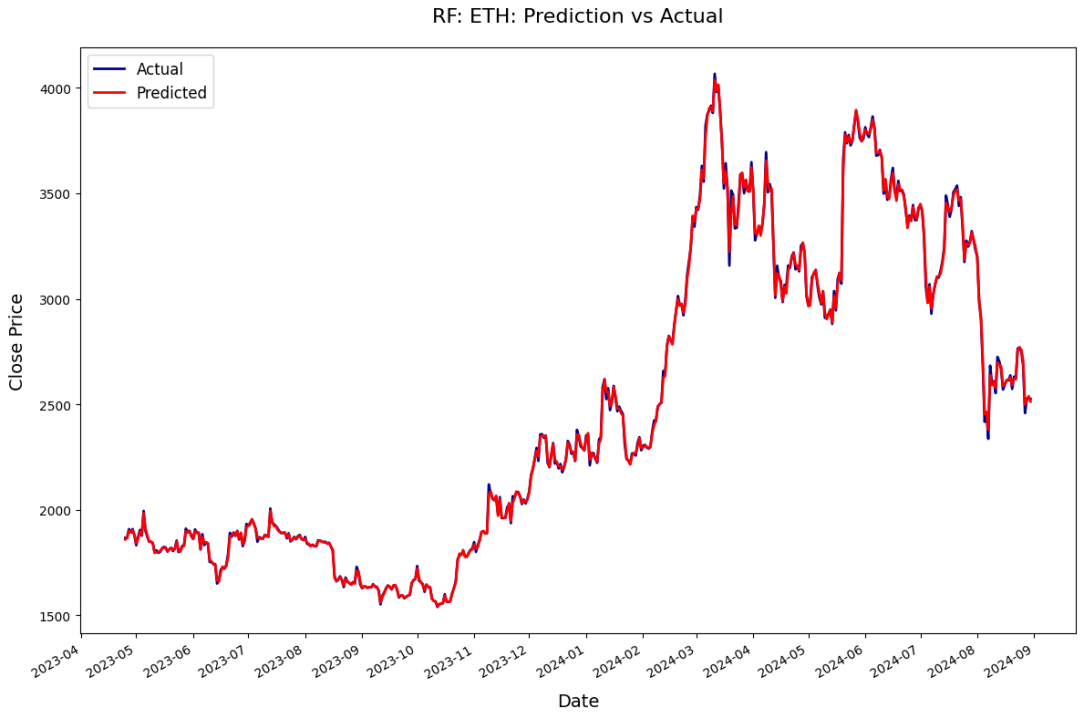

Figure 7: Predicted vs. Actual Closing Prices for Ethereum using the RF Model.

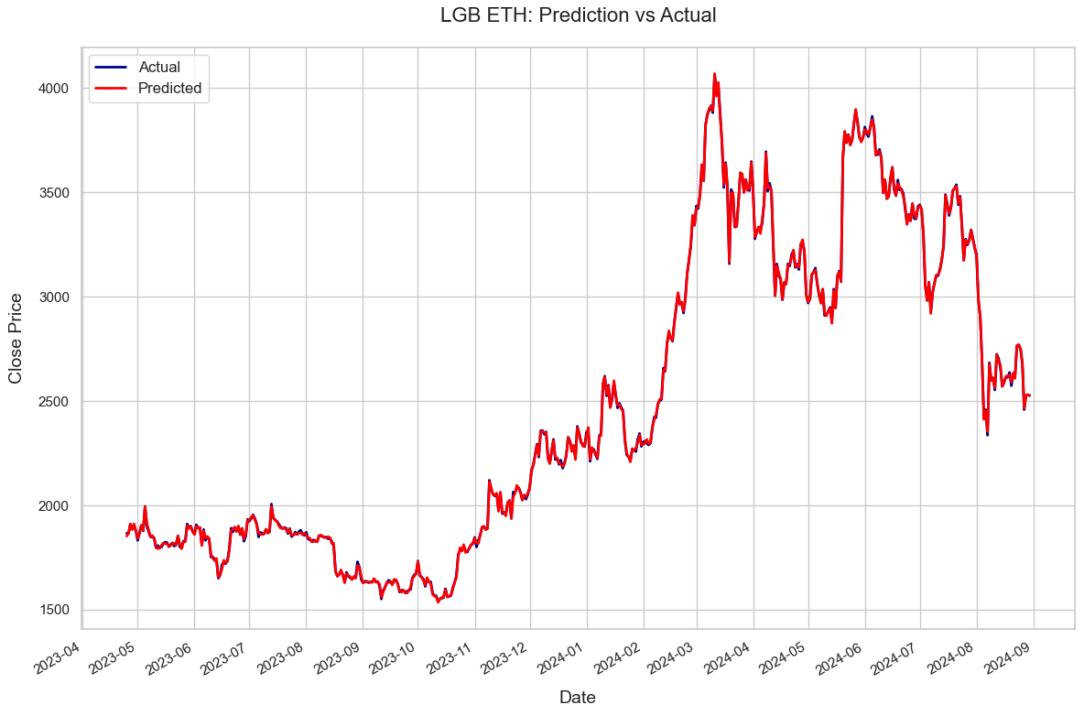

Figure 8: Predicted vs. Actual Closing Prices for Ethereum using the LGB Model.

7.3 Binance Coin

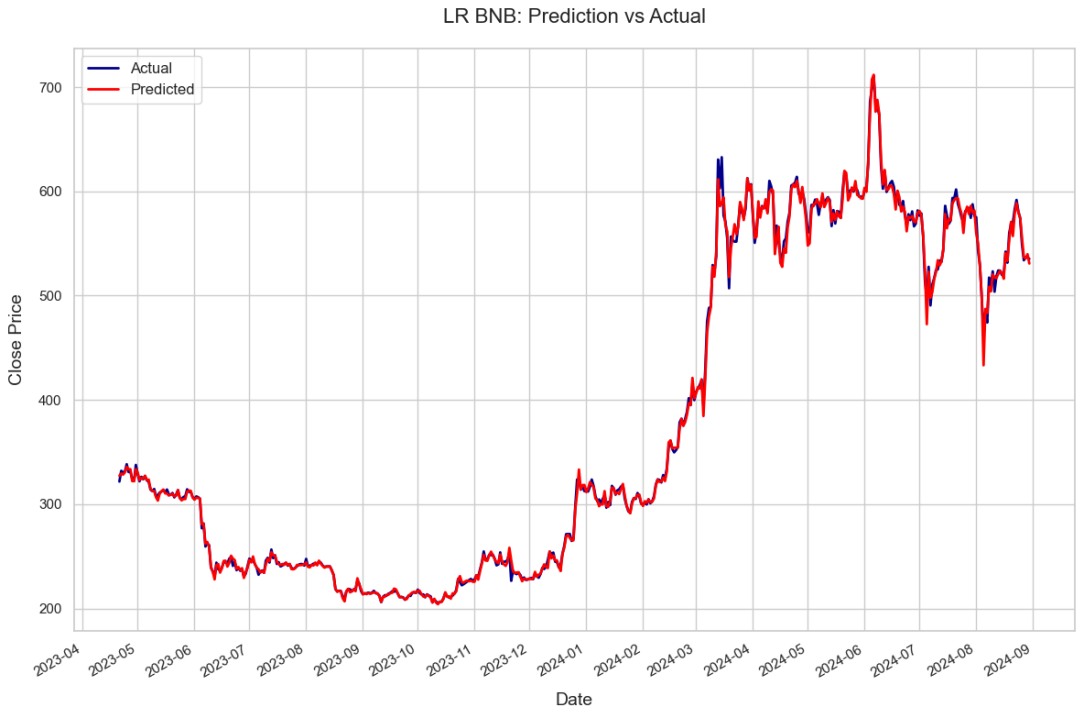

Figure 9: Predicted vs. Actual Closing Prices for Binance Coin using the LR Model.

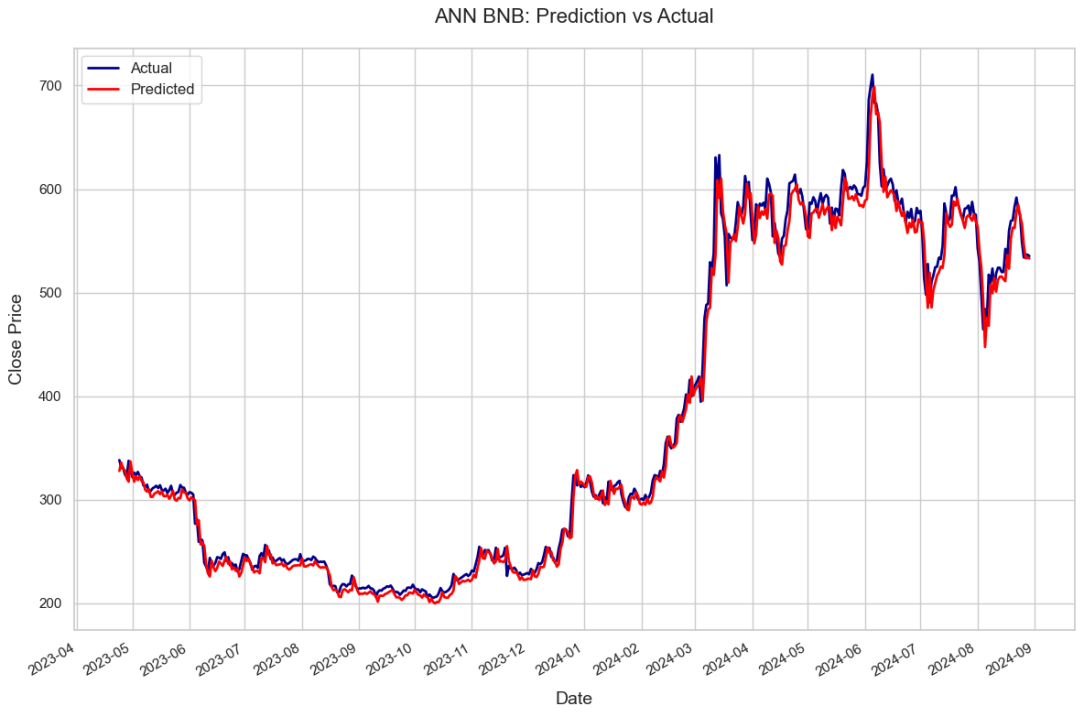

Figure 10: Predicted vs. Actual Closing Prices for Binance Coin using the ANN Model.

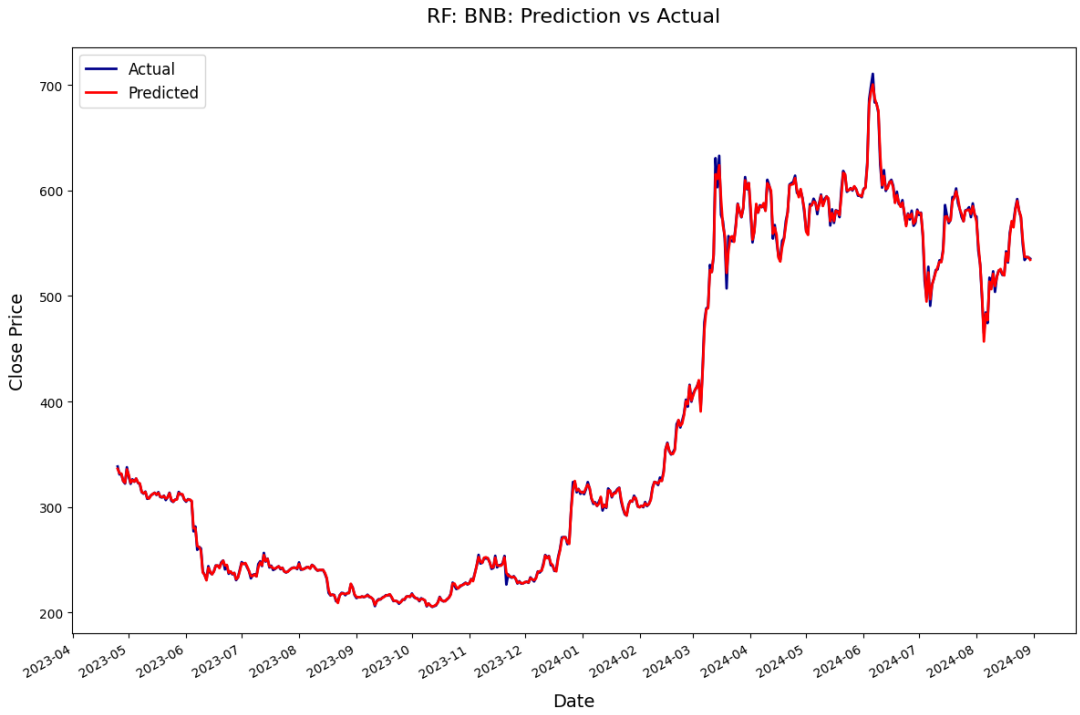

Figure 11: Predicted vs. Actual Closing Prices for Binance Coin using the RF Model.

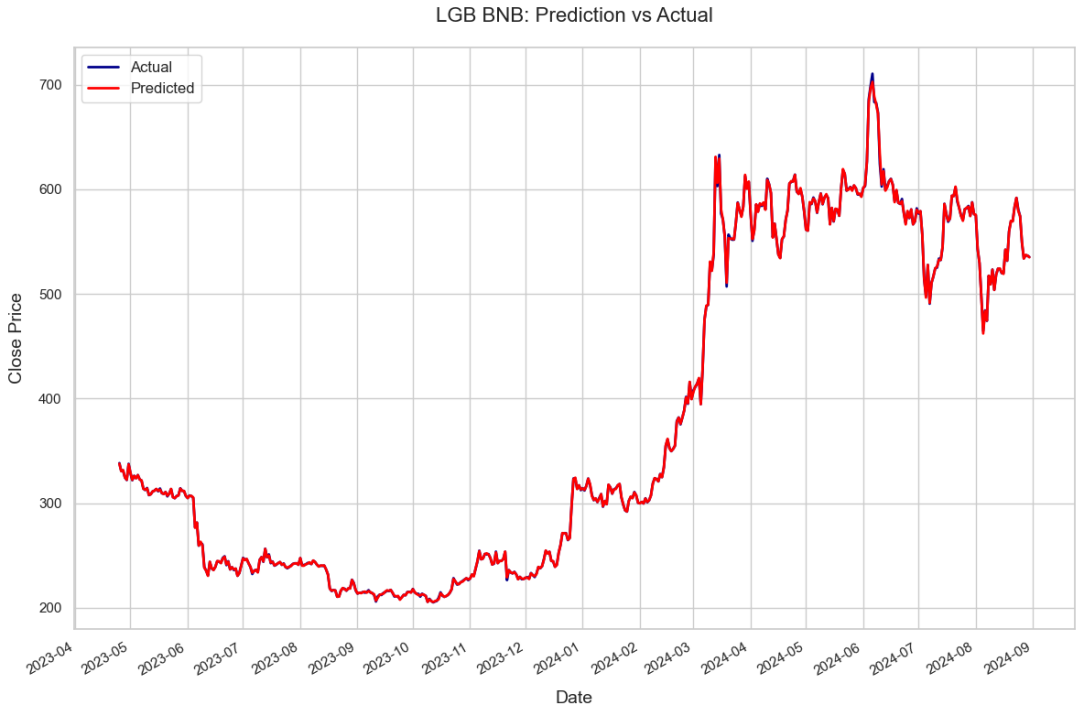

Figure 12: Predicted vs. Actual Closing Prices for Binance Coin using the LGB Model.

7.4 Solana

Figure 13: Predicted vs. Actual Closing Prices for Solana using the LR Model.

Figure 14: Predicted vs. Actual Closing Prices for Solana using the ANN Model.

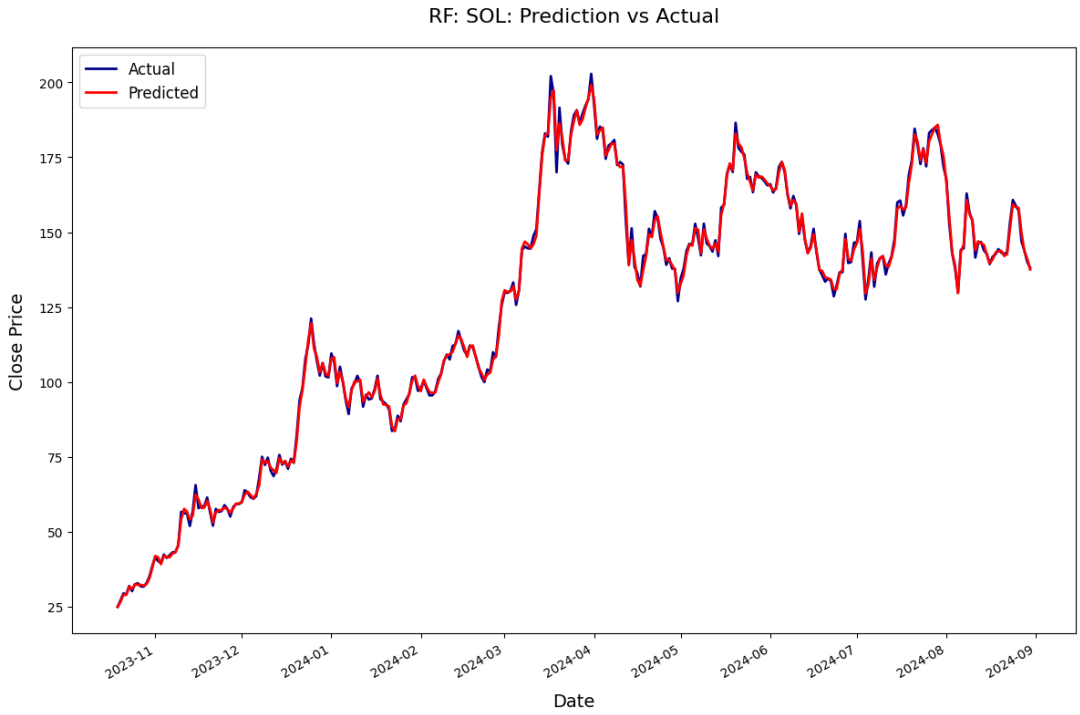

Figure 15: Predicted vs. Actual Closing Prices for Solana using the RF Model.

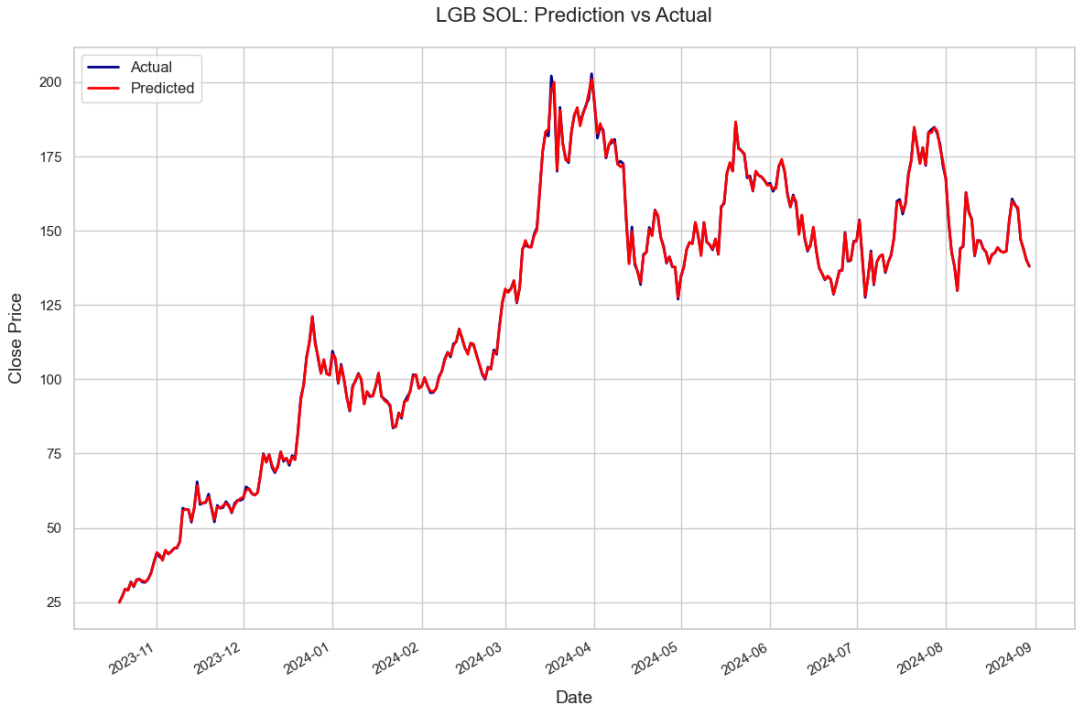

Figure 16: Predicted vs. Actual Closing Prices for Solana using the LGB Model.

8 Conclusion

This study emphasizes that the Random Forest and LightGBM approaches are far ahead of the Artificial Neural Network and Linear Regression approaches in predicting the closing values of all main cryptocurrencies the next day. Very good predictions were made for the almost uncontrollable Random Forest, particularly for very volatile assets such as Bitcoin and Ethereum, while mastering the intrinsic complexities of the data. In turn, LightGBM improved the efficiency of Ethereum, Binance Coin, and Solana in predicting prices, highlighting their efficiency and strength in relation to predictive modeling. This is in contrast with the ANN model, which exhibited the worst performance for all metrics and cryptocurrencies. Hence, although ANN has potential, it requires more fine-tuning and optimization to be as accurate as ensemble methods such as Random Forest and LightGBM, in particular Random Forest and LightGBM, which are best suited to cryptocurrency prices as they capture volatility, thus enhancing sureness. Such models, tested here with many other features and approaches, could be considered in further studies for possible improvements in predictive accuracy. In this research, the focus is on the prediction of the closing prices of the next day for a specific set of cryptocurrencies; given their historical price data consisting of open, high, low, and closed prices in the future, we can include factors such as market sentiment, change in regulatory policies, or economic conditions that might have an impact on prices.

References

Liaw, A., & Wiener, M. (2002). Classification and regression by random forest. *R News*, *2*(3), 18–22.

Zhang, G. P. (2003). Time series forecasting using a hybrid arima and neural network model. *Neurocomputing*, *50*, 159–175.

Hamzaçebi, C., Akay, D., & Kutay, F. (2009). Comparison of direct and iterative artificial neural network forecast approaches in multi-periodic time series forecasting. *Expert Systems with Applications*, *36*(2), 3839–3844.

Marsland, S. (2014). *Machine learning: An algorithmic perspective* (2nd) [Accessed: 2024-10-19]. Chapman; Hall/CRC. http://14.139.161.31/OldFiles/Machine_Learning_An_Algorithmic_Perspective_(2nd_ed).pdf

Murkute, A., & Sarode, T. (2015). Forecasting market price of stock using artificial neural network. *International Journal of Computer Applications*, *124*(12), 11–15.

Koch, J. (2021). *Machine learning with lightgbm and python* [Accessed: 2024-10-19]. Packt Publishing. https://www.packtpub.com/en-us/product/machine-learning-with-lightgbm-and-python-9781800564749

Wang, M. (2022). Prediction of the technology company's stock price through the deep learning method. *Open Journal of Modelling and Simulation*, *10*(04), 428–440. https://doi.org/10.4236/ojmsi.2022.104024

Dagur, A., Singh, P., Mehra, D. K., & Shukla, K. S. (Eds.). (2023). *Artificial intelligence, blockchain, computing and security volume 1*. Taylor & Francis. https://www.taylorfrancis.com/books/edit/10.1201/9781003393580/artificial-intelligence-blockchain-computing-security-volume-1-arvind-dagur-pawan-singh-mehra-dhirendra-kumar-shukla-karan-singh

Pandey, A., Singh, A., & Singh, A. (2023). Stock price prediction using machine learning. *International Journal of Engineering Technology and Computer Research (IJETCR)*, *9*(1), 01–07. https://www.ijetcr.org/paper/stock-price-prediction-using-machine-learning

S., B. J. (2023). *Linear models in statistics* [Accessed: 2024-10-19]. University of Toronto. https://www.utstat.toronto.edu/~brunner/books/LinearModelsInStatistics.pdf

Vasilj, D., Vučić, F., & Matković, Ž. (2024). Application of artificial intelligence in the economic and legal affairs of companies. In *Communications in computer and information science* (pp. 229–240). Springer. https://doi.org/10.1007/978-3-031-62058-4_14