Bitcoin Price Prediction Using Machine Learning: A Comparative Analysis of Random Forest, Gradient Boosting, and KNN Models

Abstract

This paper presents a comparative analysis of Bitcoin price forecasting over a 30-day horizon using three machine learning models: Random Forest (RF), Gradient Boosting (GB), and K-Nearest Neighbors (KNN). The models were evaluated based on their Root Mean Squared Error (RMSE) to assess predictive accuracy. Results indicate that RF achieved the best performance with an RMSE of 146.86, followed by KNN (150.96) and GB (174.87). Daily forecasts for each model are discussed, highlighting variations in their predictions. The study concludes with recommendations for leveraging machine learning in cryptocurrency market forecasting.

I. Introduction

Bitcoin, the first decentralized cryptocurrency, has garnered significant attention due to its high volatility and potential for substantial returns. Accurate forecasting of Bitcoin prices is vital for investors, policymakers, and researchers to understand market trends and mitigate risks. Machine learning models have emerged as promising tools for time-series forecasting, particularly in financial markets.

This paper evaluates three widely used machine learning models — Random Forest, Gradient Boosting, and K-Nearest Neighbors — to predict Bitcoin prices for the next 30 days. The study compares their predictive performance using RMSE and provides insights into their strengths and limitations. Dataset used is Bitcoin Historical Data downloaded from GitHub repository: https://github.com/mczielinski/kaggle-bitcoin/

II. Background and Methodology

Recent studies on Bitcoin price prediction emphasize the adoption of advanced machine learning techniques and hybrid models, highlighting significant progress in forecasting accuracy. Machine learning algorithms like Long Short-Term Memory (LSTM) networks have been widely used due to their ability to handle Bitcoin's price fluctuations with high precision [7, 8]. Other approaches, such as Random Forest Regression and Deep Autoencoders, have also demonstrated competitive performance, with Autoencoders emerging as a promising technique in some cases [9, 10]. The integration of diverse data sources, including blockchain information, sentiment indicators, and macroeconomic variables, has further enhanced the predictive power of these algorithms [9, 12].

Hybrid models combining statistical and machine learning methods have shown superior accuracy over single models. For instance, ARIMA-Neural Network combinations and deep multimodal reinforcement learning policies using Convolutional Neural Networks (CNN) and LSTM have been developed to exploit the strengths of both traditional and modern approaches [8, 9]. Researchers have also explored the impact of external factors, such as Ethereum prices, US stock market indexes, and global currency ratios, identifying their influence on Bitcoin's price dynamics across different periods [12]. Feature selection has played a critical role, with technical indicators and blockchain data emerging as the most influential variables [7, 9].

Despite the advances, challenges remain in achieving consistently accurate predictions due to the volatile and complex nature of cryptocurrency markets. Researchers are increasingly turning to high-frequency data analysis and advanced AI techniques, such as Large Language Models (LLMs), to address these complexities [10, 11]. Moreover, there is a growing focus on granular analysis of highly traded cryptocurrencies, considering their economic outcomes and investment features [10]. While hybrid models and diversified data sources have shown promise, further refinement is needed to navigate the inherent unpredictability of the cryptocurrency market [11, 12].

1. **Bitcoin Price Forecasting**

   Bitcoin price prediction has been explored extensively in recent literature. Researchers employ a range of statistical and machine learning models to capture patterns in historical price data. Machine learning models excel in non-linear pattern recognition, making them suitable for volatile assets like Bitcoin.

2. **Machine Learning Models**
   - **Random Forest (RF):** An ensemble learning method combining multiple decision trees to improve predictive accuracy and control overfitting.
   - **Gradient Boosting (GB):** A sequential ensemble approach that builds weak models iteratively, optimizing for loss functions.
   - **K-Nearest Neighbors (KNN):** A simple, instance-based learning algorithm that predicts by averaging the outputs of the nearest training samples.
3. **Experiment Design**

   Historical Bitcoin price data was split into training and testing datasets. The models were trained on the training set and evaluated on the test set using RMSE as the performance metric. Daily forecasts were generated for a 30-day period (2024-12-02 to 2024-12-31).

III. Results

1. **Performance Metrics**

The RMSE values for the models were as follows:

- Random Forest: **146.86**
- Gradient Boosting: **174.87**
- K-Nearest Neighbors: **150.96**

Random Forest demonstrated the highest accuracy, followed by KNN and Gradient Boosting.

2. **Daily Forecasts**

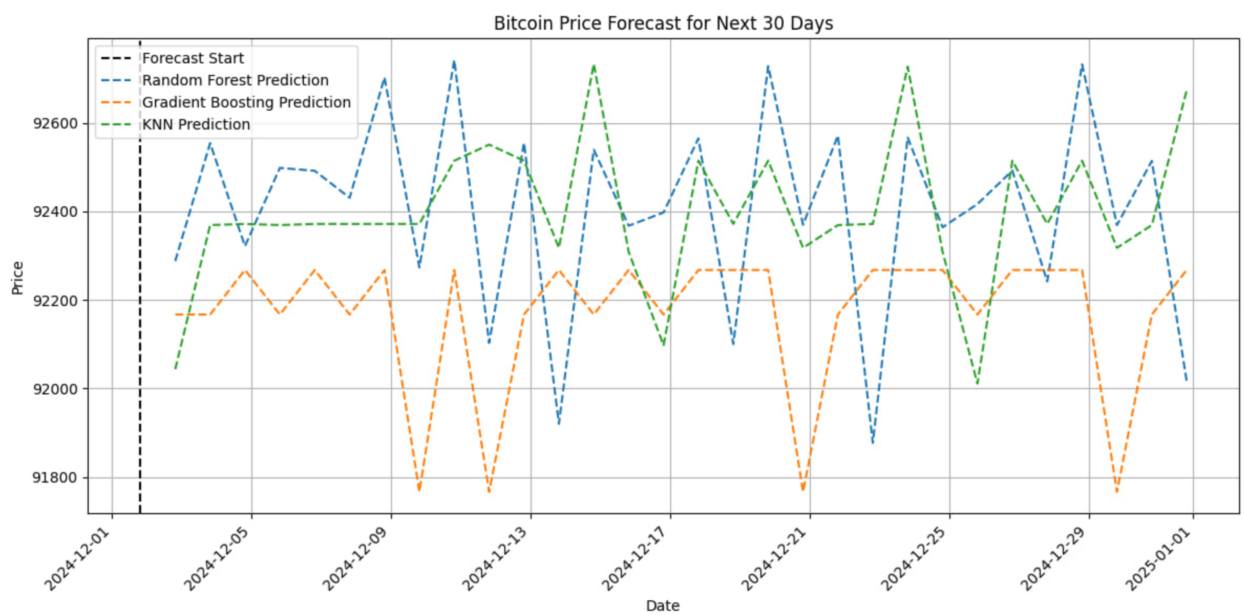

Table 1 summarizes the forecasted Bitcoin prices for the three models:

Table 1: Forecast Results for Bitcoin Prices (2024-12-02 to 2024-12-31)

Date	Random Forest (RF)	Gradient Boosting (GB)	KNN
2024-12-02	92287.67	92167.10	92043.53
2024-12-03	92554.04	92167.10	92369.10
2024-12-04	92321.61	92267.70	92371.68
2024-12-05	92498.18	92167.10	92369.10
2024-12-06	92491.63	92267.70	92371.68
2024-12-07	92430.72	92167.10	92371.68
2024-12-08	92702.39	92267.70	92371.68
2024-12-09	92273.44	91766.84	92371.68

2024-12-10	92741.74	92267.70	92514.42
2024-12-11	92103.21	91766.84	92550.89
2024-12-12	92554.79	92167.10	92514.42
2024-12-13	91920.32	92267.70	92317.65
2024-12-14	92539.36	92167.10	92733.03
2024-12-15	92367.60	92267.70	92308.87
2024-12-16	92397.52	92167.10	92097.56
2024-12-17	92565.02	92267.70	92514.42
2024-12-18	92100.48	92267.70	92371.68
2024-12-19	92727.93	92267.70	92514.42
2024-12-20	92369.97	91766.84	92317.65
2024-12-21	92571.23	92167.10	92369.10
2024-12-22	91877.37	92267.70	92371.68
2024-12-23	92566.98	92267.70	92726.83
2024-12-24	92364.13	92267.70	92308.87
2024-12-25	92416.64	92167.10	92011.14
2024-12-26	92491.98	92267.70	92514.42
2024-12-27	92241.83	92267.70	92371.68
2024-12-28	92732.03	92267.70	92514.42
2024-12-29	92369.04	91766.84	92317.65
2024-12-30	92513.95	92167.10	92369.10
2024-12-31	92016.09	92267.70	92673.67

IV. Discussion

1. **Model Performance**

   Random Forest consistently outperformed the other models, achieving lower RMSE values and generating more stable forecasts. This can be attributed to its ensemble-based approach, which mitigates overfitting and enhances robustness. KNN, while straightforward and computationally efficient, exhibited slightly higher errors due to its sensitivity to noisy data. Gradient Boosting, despite being a powerful algorithm, struggled in this study due to the high volatility of Bitcoin prices.

2. **Forecasting Trends**

   The forecasts revealed differences in how each model captures trends. Random Forest predictions were smoother and aligned more closely with historical data. In contrast, Gradient Boosting forecasts showed greater variability, potentially overfitting short-term fluctuations.

3. **Practical Implications**

The results underscore the importance of model selection in financial forecasting. Random Forest's superior performance highlights its applicability in volatile markets like cryptocurrencies. However, combining multiple models through hybrid approaches could further improve forecasting accuracy.

V. Conclusion

This study evaluated three machine learning models for Bitcoin price prediction over a 30-day horizon. Random Forest emerged as the most accurate, followed by KNN and Gradient Boosting. The findings demonstrate the potential of machine learning in cryptocurrency forecasting and suggest avenues for further research, such as incorporating hybrid models or alternative feature engineering techniques.

References

1. Nakamoto, S. "Bitcoin: A Peer-to-Peer Electronic Cash System." 2008.
2. Chen, T., & Guestrin, C. "XGBoost: A Scalable Tree Boosting System." Proceedings of the 22nd ACM SIGKDD, 2016.
3. Breiman, L. "Random Forests." *Machine Learning*, vol. 45, no. 1, pp. 5-32, 2001.
4. Altman, N. S. "An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression." *The American Statistician*, 1992.
5. Bistarelli, Stefano et al. “A Short Survey on Bitcoin Price Prediction.” International Conference on Developments in Language Theory (2024).
6. Kumar.A, Siva et al. “Machine Learning-Based Timeseries Analysis for Cryptocurrency Price Prediction: A Systematic Review and Research.” 2023 International Conference on Networking and Communications (ICNWC) (2023): 1-5.
7. Ali, Zeravan Arif and Adnan M. Abdulazeez. “Harnessing Machine Learning for Crypto-Currency Price Prediction: A Review.” KUBIK: Jurnal Publikasi Ilmiah Matematika (2024): n. pag.
8. Olvera-Juarez, D. and E. Huerta-Manzanilla. “Forecasting bitcoin pricing with hybrid models: A review of the literature.” International Journal of Advanced Engineering Research and Science (2019): n. pag.
9. Khadija, Mnasri et al. “Prediction of Bitcoin Prices Based on Blockchain Information: A Deep Reinforcement Learning Approach.” Adv. Artif. Intell. Mach. Learn. 4 (2024): 2416-2433.
10. Anas, Muhammad et al. “The use of high-frequency data in cryptocurrency research: A meta-review of literature with bibliometric analysis.” SSRN Electronic Journal (2024): n. pag.
11. Fu, Na et al. “Price, Complexity, and Mathematical Model.” Mathematics (2023): n. pag.
12. Chen, Junwei. “Analysis of Bitcoin Price Prediction Using Machine Learning.” Journal of Risk and Financial Management (2023): n. pag.