International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 05 Issue: 03 | March – 2026 DOI:10.55041/ISJEM05754
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

Predictive Analysis of Bitcoin Price Trends Usingsupervised Machine Learning Algorithms

Dr. Y. Mohammed Iqbal[1], T. Dhanushkumar[2], Dr. S. Peerbasha[3],Dr. M. Mohamed Surputheen[4], Dr. M. Rajakumar[5]

Department of Computer Science, Jamal Mohamed College, Affiliated to Bharathidasan University, Tiruchirappalli, Tamil Nadu, India

---***--

Abstract- Cryptocurrency markets, particularly Bitcoin, are characterized by high volatility and complex non-linear price movements, making trend prediction a significant challenge. The purpose of this research is to compare six different supervised machine learning models: Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbors (KNN), XGBoost, and Support Vector Machine (SVM), using a large dataset of daily changes in Bitcoin prices from September 2014 through January 2026 (24,863 records). The dataset is augmented by twelve different technical indicators (e.g. RSI, MACD, and multiple Simple Moving Averages (SMA)) that were created as input variables. As the financial data in this dataset is temporal in nature, time series cross-validation (Time Series Split) was used to evaluate the models in order to reduce the likelihood of overfitting due to random sample shuffling. Based on the experimental results, the Random Forest and XGBoost ensemble models are significantly better at predicting the price change for cryptocurrencies than the non-ensemble models, with the Random Forest model exhibiting an accuracy rate of 79.18% and AUC rate of 0.8675 for the validation folds, while the XGBoost model exhibited a 74.66% accuracy rate. This implies that advanced tree-based ensemble models are capable of providing a significant degree of prediction for cryptocurrency price trends if they are properly regularized and validated against financial market noise. Additionally, the predictive superiority of the ensemble models over the non-ensemble models was demonstrated statistically through the use of McNemar's test and Point-Biserial correlation ($p < 0.05$).

Keywords: Bitcoin, Machine Learning, Technical Analysis, Random Forest, XGBoost, Price Prediction, Financial Forecasting, Cryptocurrency, Supervised Machine Learning, Ensemble Learning, Time Series Cross-Validation, Feature Engineering.

1. INTRODUCTION

Bitcoin has been established as a key player in the world financial market due to the rapidly developing and growing physical and virtual assets around the world. Whereas typically associated with government policies and monetary supply, bitcoin is not tied to traditional currencies such as the USD. Bitcoin price is determined by the complex interplay of forces such as market sentiment, rates of adoption, and technical retail trading activity that cannot easily be measured with standard methods of analysis. Bitcoin has the potential for both profitability and tremendous risk, creating an urgent need for tools that can provide meaningful data for predicting patterns and trends in price.

One of the largest markets in the financial system today is the cryptocurrency market with significant amounts of volatility combined with the open nature of this market due to being able to be traded at any time can lead to significant returns on investment when prices do move. Bitcoin, being the first cryptocurrency created and the most well-known, will often be used as a reference point for other cryptocurrencies within the cryptocurrency market [1]. The large fluctuations in value of Bitcoin also provide possible trading opportunities for many people, however, due to the unpredictable nature of these large fluctuations, predicting the trend in price of Bitcoin is considered a challenging research problem by many [1]. Research shows that most traditional forms of predictability used in the financial world typically will not work well with the cryptocurrency market due to its unique characteristics like the fact that it's decentralized, the fact that prices are often driven by social media and regulations and by new technology [2]. Machine learning based methods have been able to develop complex, nonlinear connections or patterns that do not always exist through more traditional forms of statistics, thus can serve as a way to help identify price movements [2]. Using technical indicators (which are derived from historical prices and volume data) are an additional method, in providing a way to use quantitative information to assist in developing predictive models for price movements [3]. Further advancements into ensemble learning and feature engineering have provided improvements in the accuracy of price movement predictions within the context of financial time series analysis. Ensemble models combine multiple models together, to create a model with better predictive capabilities than an individual model, thus helping to reduce variability and increase accuracy [4]. This paper adds to the growing body of work on predicting Bitcoin prices by using a comprehensive machine-learning framework to predict trends through the use of multiple algorithms, and to highlight the main features driving the movement of Bitcoin prices.

2. PROBLEM STATEMENT

Although there are considerable amounts of historical market data available, it is still difficult to accurately predict trends in bitcoin prices. The market for bitcoin is different from traditional fiat currencies or regulated stock markets in that it is disaggregated and operates on a 24-hour, 7 days per week basis, creating a variety of unique challenges for predictive modeling.

A. Extreme Volatility and Non-Stationarity

The values for bitcoin are extremely volatile and on average trade daily with at least a 5-10% change due to some degree of speculative trading; as such, statistical evaluation of bitcoin price data represents the highest degree of non-stationary nature, which results from the price exhibiting different parameters on different days (i.e. mean and standard deviation). The traditional time series modeling (i.e. ARIMA) used to analyze financial time series assume constant statistical properties; therefore, they are not an appropriate means of capturing the random-chaotic behavior of cryptocurrencies.

B. The "Overfitting" Trap in Machine Learning

There is a critical lack of empirical research into the methodological errors related to ML, with statistical overfitting and look-ahead bias being major sources of error in developed predictive models. Many predictive models seem to achieve near perfect (99%+) predictive accuracy (measured via experimental data) with random train-test splits used for splitting the training and testing databases. By leaving random chance for a successful outcome regarding the prediction of future price movements when simply using experimental data, the models are training using future data that are temporally adjacent to the testing points—resulting in a model that memorizes noise instead of learning generalizable and stable price trends. Thus, the derived price movement predictions from these models will ultimately lead to poor trading outcomes.

C. Feature Complexity and Interpretability

Price movements in the cryptocurrency market are impacted by multiple factors, such as technical indicators (momentum/volume), news of global regulations, and public sentiment as relayed by social media sources. Traditional linear modeling techniques (Logistic Regression) fail to properly account for the complexities and nonlinear interactions between the factors affecting price movement (for example, high volume may amplify a trend signal in a bull market but indicate a potential trend reversal in the bear market). While deep-learning modeling techniques (LSTM's) are capable of modeling nonlinear relationships, the lack of interpretability of the LSTM parameters makes it difficult for the analyst or trader to understand the basis for the price movement prediction.

D. Research Objective

Basically, this research addresses the central focus of developing a machine learning algorithm that is robust, interpretable and methodologically sound. Unlike earlier studies, which used random splitting of datasets with the goal of maximizing raw accuracy; this study utilises Time Series Cross Validation due to the requirement respect for the temporal sequence of financial data. The purpose of this research is to see if ensemble tree-based algorithms (Random Forest and XGBoost) can achieve stable directional forecasting through use of stringent regularization and realistic validation constraints.

3. LITERATURE REVIEW

A large body of literature exists surrounding Bitcoin price prediction as a rapidly evolving topic within the area of financial machine learning due to both the unique nature of Bitcoin as an asset and the high volatility associated with it since Nakamoto created the foundational protocol [1]. Over this time period, researchers have shifted their approach to employing more advanced types of machine learning and deep learning techniques instead of relying primarily on classical econometric models.

Traditional vs. Machine Learning Approaches. Early research often relied on traditional time-series models. McNally et al. [2] conducted a comparative study of ARIMA, LSTM, and Recurrent Neural Networks (RNN) for Bitcoin price prediction. Their findings demonstrated that while ARIMA models are effective for linear trends, they fail to capture the non-linear volatility of cryptocurrency markets. In contrast, deep learning models like LSTM achieved significantly higher accuracy (52%) and better directionality prediction. Similarly, Swapna and Rao [19] compared ARIMA with LSTM, confirming that LSTM networks are superior in handling the long-term dependencies found in financial time series, although they require substantially more computational power. A study done by Swapna and Rao [19] confirmed that Long Short Term Memory (LSTM) networks were superior to ARIMA models in terms of their ability to deal with the long-term dependencies of financial data, but also required a much greater amount of compute power.

The application of deep learning to predictive model development extends well beyond financial time series data into other areas that have a significant degree of complexity, non-linearity and for example the hybrid predictive architecture (CNN + LSTM) developed in this research demonstrates that significant advances in accuracy and robustness of prediction can be achieved through the application of deep learning and developing sophisticated feature extraction methods. In addition, Robust features can be extracted through sophisticated pre-processing, segmentation, and feature fusion upon a wide variety of datasets. This research demonstrated a very high degree of predictive accuracy and the system can also be applied to provide real-time clinical decision support during a pandemic [25].

Moreover, an optimized deep learning framework for automating COVID-19 detection through lung imaging modalities (i.e., chest X-ray and CT scans) was created by researchers [26]. The new method enhanced the built-in CNN architecture, along with some improvements in the transfer learning method, so researchers could enhance generalizability on small medical databases [26]. The augmentation of the data as well as techniques to reduce noise significantly improved both robustness and reduced over-fitting for the model [26]. The final results of the model indicated excellent accuracy (i.e., 92%) at classifying between patients with COVID-19 and patients without COVID-19 [26]. These results provide an example of how intelligent deep learning systems can separate the signal from the noise, which is also a very important aspect of working with or developing methodologies for analyzing the chaotic randomness related to cryptocurrency markets.

The Role of Technical Indicators and Features. The selection of input features is critical for model performance. The research conducted by Yoo et al. [11] analyzes how effective technical analysis is in predicting stock market movements; they have established a framework that has been widely used to predict movements in cryptocurrency markets. Their findings indicate that derived features used for technical analysis such as moving averages and momentum indicators have more predictive power than the underlying price data alone. Madan et al. [10] tested this thesis with the use of a random forest classifier based upon technical indicators to predict Bitcoin’s price movement; they were able to accurately predict the daily price movements of Bitcoin over 90% of the time using technical patterns found in Bitcoin due to the significant prevalence of algorithmic trading in the markets compared to more mature stock markets.

Volatility and sentiment analysis of cryptocurrencies is critical to understanding how these markets respond to outside influences. Bitcoin price fluctuations are affected by more than just traditional economic stimulus, according to Kristoufek [4]; in addition, speculative interest and increased volume in search engines have a significant impact on price fluctuations. He also found using wavelet coherence analysis that the price movement of Bitcoin coincided with spikes in Google Trendaq during bubble periods. In addition to predictive analytics based upon price and economic factors, Abraham et al. [9] and Greaves and Au [15] have included volume and sentiment measurements from Twitter into their predictive models. Their findings indicate that spikes in negative sentiment are often followed by price declines, while high volumes of tweets are positively correlated with increased volatility in digital currency prices.

Research has focused on using ensemble methods to enhance performance. An example of this is the development of a new deep learning ensemble model developed by Li and Pan [18], which combines the outputs of several neural networks into one ensemble-based (or "hybrid") model. Research by Hitam and Ismail [7] compared algorithms such as Naive Bayes and

Support Vector Machines (SVM), with the results showing that hybrid models produced lower levels of variance error compared to individual classifiers in the presence of noisy market data. Evidence from Livieris et al. [12] further supports this by providing validation for ensemble hybrid techniques using deep learning to produce more consistent and profitable trading strategies via averaged weight-based predictions.

Still, existing studies often exhibit overfitting due to random data partitioning with few studies using proper cross-validation methods. Some of the newer or potential models, such as those displayed in Deep Learning (LSTM) [20], are often classified as "black box" models, which reduces the level of interpretability associated with the outcome of using the model. This study will address these issues by using interpretable models (Random Forest, XGBoost) combined with a strict Time Series Cross-Validation method to provide an accurate performance estimate.

A. Comparison of Related Works

Table 1: Comparison of Related Works in Bitcoin Price Prediction

Study	**Features Used**	**Model**	**Findings / Limitations**
McNally et al. (2018) [2]	OHLC, Hash Rate, Difficulty	LSTM, RNN, ARIMA	LSTM outperformed ARIMA; limited to 52% accuracy due to high volatility
Madan et al. (2021) [10]	Technical Indicators, Blockchain Data	Random Forest, GLM	Achieved >90% accuracy but lacked strict time-series validation
Kristoufek (2015) [4]	Price, Google Trends	Wavelet Coherence	Identified strong link between search interest and price; non-predictive model
Abraham et al. (2018) [9]	Price, Tweet Volume, Sentiment	Linear Regression	Tweet volume is a leading indicator of volatility; sentiment analysis improved fit
Livieris et al. (2021) [12]	OHLC, Technical Indicators	Ensemble Deep Learning	Ensemble weights improved stability; computationally expensive training
Proposed System	OHLC, 12 Technical Indicators (RSI, SMA)	Random Forest, XGBoost (Time Series Split)	Uses time series folding to prevent overfitting; balances accuracy with interpretability

B. System Architecture

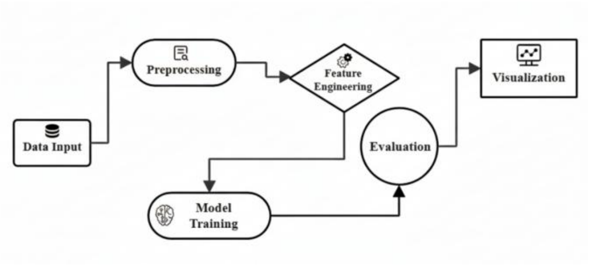


Fig -1: System architecture of Bitcoin price prediction frame work

This paper provides a six-step structure for the proposed system.

1. **Data acquisition:** Retrieve prior instances of bitcoin price action from various sources.

2. **Pre-processing:** Process the retrieved data into a format suitable for use by the model.

3. **Feature Engineering:** Create additional features based on look-back periods or ratios of each standard financial attribute that can be used as independent variables.

4. **Model Implementation:** Implement three machine learning classifiers and implement each classifier multiple times on the same data set.

5. **Evaluation:** Assess the performance of the three classifiers using standard performance metrics.

6. **Visualization:** Create various graphs of the evaluation results for comparative purposes.

4. PROPOSED METHODOLOGY

The proposed machine learning framework will use a sequence of five steps to predict the future trend of bitcoin prices. The five steps are: Data acquisition, Data pre-processing, Feature engineering, Model implementation, and Performance evaluations.

ISSN: 2583-6129
DOI:10.55041/ISJEM05754

A. Data Acquisition and Description

The source of the historical market data collected spans September 17, 2014 through approximately January 25, 2026. The data consists of 24,863 total records with a single record for each day within the defined date range based on actual daily trades for bitcoin.

- **Source: Historical data.**

- **Attributes of the raw historical data include:** (1) DATE; (2) Open Price; (3) High Price; (4) Low Price; (5) Closing Price; and (6) Trading Volume.

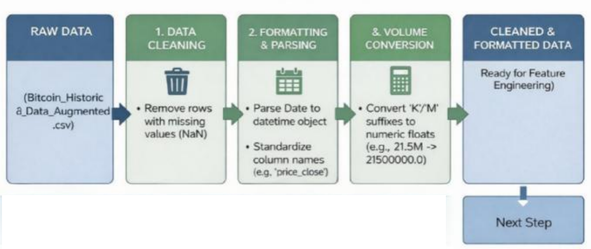
B. Data Preprocessing

Raw financial data often contains inconsistencies that can degrade model performance. The following preprocessing steps were applied:

- **Cleaning:** Rows with missing values were removed to ensure data integrity.

- **Formatting:** The 'Date' column was parsed into a datetime object to maintain temporal order.

- **Normalization:** The 'Volume' column, originally containing text suffixes (e.g., 'K' for thousands, 'M' for millions), was converted into a standard numeric float format for computational consistency.


RAW DATA

(Bitcoin_Historica_Data_Augmented.csv)

1. DATA CLEANING

- Remove rows with missing values (NaN)

2. FORMATTING & PARSING

- Parse Date to datetime object
- Standardize column names (e.g. 'price_close')

& VOLUME CONVERSION

- Convert 'K'/'M' suffixes to numeric floats (e.g., 21.5M -> 21500000.0)

CLEANED & FORMATTED DATA

Ready for Feature Engineering)

Next Step

Fig -2: Bitcoin Data Preprocessing PipeLine

C. Feature Engineering

In order to capture the non-linear dynamics of the cryptocurrency market, we developed **12 technical indicators** based on the raw OHLC (Open, High, Low and Close) data that will be our main predictors for the machine learning models.

i. Trend Indicators:

- **Simple Moving Averages (SMA):** To identify short-term and medium-term trends, we calculated for 5 days, 10 days, 20 days and 50 days.directions.

ii. Momentum Indicators:

- **Relative Strength Index (RSI):** The 14-day RSI will measure the speed and size of price movement to identify when the market is overbought or oversold.

- **Moving Average Convergence Divergence (MACD):** The MACD is derived from the difference between the 12-day and 26-day Exponential Moving Averages (EMA), and is a momentum indicator that is typically used to follow price trends.

iii. Volatility Indicators:

- **Rolling Volatility:** We used the standard deviation of price changes over a rolling 5-day period to provide a better measure of instability in the market.

iv. Price Derived Features:

- **Daily Return:** The change in daily closing price from the previous day's closing price expressed as a percentage.

- **High-Low Spread:** The difference between the high price for the day and the low price for the day.

- **Open-Close Spread:** The difference between the open price and closing price of the day.

v. Basic Price Features

The Basic Price Features will be calculated from the daily Open, High, Low, Close (OHLC) data to show the most current price changes in the market. The following are examples of how these basic price features can be derived:

- **Price change (daily difference):** The Price Change (also known as the Daily Change) is the absolute price difference of the closing price from today (t) and the closing price from yesterday (t-1) to determine how much the closing price changed.

$$\Delta P_t = P_t - P_{t-1}$$

- **Daily return (percentage change):** the percentage change from the prior day's closing price to today's closing price. The Daily Return normalizes the price change and is useful in comparing price changes over different time periods and price levels.

$$R_t = \frac{P_t - P_{t-1}}{P_{t-1}}$$

- **Simple Moving Average:** It is an indicator used to identify trends in the market price of an asset by calculating the un-weighted mean of the closing prices over a specific number of days (n). The SMA provides a smoothed release curve that can help identify the directional movement of trends in the marketplace.

a. **4-day SMA:** It is used to smooth the price in the short-term (n=4).

b. **10-day SMA:** It provides a smoothed trend in short to medium-term prices (n=10).

c. **20-day SMA:** It is used for medium-term trends and is often used as the basis for Bollinger Bands (n=20).

$$SMA_n = \frac{\sum_{i=0}^{n-1} P_{t-i}}{n}$$

- **Volatility:** It is the standard deviation of daily returns calculated using a rolling 5-day window. The Volatility feature measures market risk and instability; it usually leads to major price corrections or price breakout points.

$$\sigma_t = \frac{1}{5}\sum_{i=0}^{4} |\Delta P_{t-i}|$$

vi. Advanced Technical Indicators

These derived features capture complex market properties such as momentum, liquidity, and intraday intensity.

- **Open-Close spread:** The difference between the opening and closing prices of the same day. A large positive spread indicates strong buying pressure throughout the day, while a negative spread indicates selling pressure.

$$OC_t = Open_t - Close_t$$

- **High-Low Range:** The difference between the highest and lowest prices exchanged during that specific day is used to determine how volatile liquid an assets' trading price is within a one-day time frame.

$$HL_t = High_t - Low_t$$

- **50-Day SMA:** An analytical tool that attempts to measure long-term long-term asset price performance is to compare the asset's current trading price against the asset's 50-day SMA.

$$SMA_{50}(t) = \frac{1}{50}\sum_{i=0}^{49} P_{t-i}$$

- **RSI (Relative Strength Index):** The Relative Strength Index is a momentum oscillator measuring both speed and degree of price movement gauged between a range of 0 - 100 in order to determine whether the securities at issue have a potential for a price increase (overbought condition) or if they will experience a price decrease (oversold condition) within a given period of time.

$$RSI_t = 100 - \frac{100}{1+\frac{AvgGain}{AvgLoss}}$$

- **MACD (Moving Average Convergence Divergence):** The MACD is determined by deducting a long-range long-term (26-days) exponential moving average from the short-term (12-days) exponential moving average. MACD values are considered positive when their value reflects "upside momentum" or negative when reflecting "downside momentum".

$$MACD_t = EMA_{12}(t) - EMA_{26}(t)$$

- **Volume:** The total dollar amount of Bitcoin traded during one day. Volume helps support price trends by providing volume confirmation; price trends with larger trading volumes are considered more meaningful and/or more likely to sustain than are those trends with lower volume levels.

vii. Target Variable Definition

The prediction task is formulated as a binary classification problem where the target variable indicates whether the Bitcoin price will increase (1) or decrease/remain stable

(0) on the next trading day, where I(·) is the indicator function:

$$t\arg et_t = \mathrm{I}(P_{t+1} > P_t)$$

viii. Machine Learning Models

The six machine learning models that were created and evaluated are:

1. **Logistic Regression:** its a baseline linear classifier with its own sigmoid activation formed by using a logistic function to predict probabilities (e.g., the probability of something being true or false).

2. **Decision Tree:** A decision tree is a non-linear classifier that has a familiar representation of how decisions (nodes) translate into outcomes (leaves).

3. **Random Forest:** A random forest consists of multiple decision trees assembled through bagging (bootstrap aggregating).

4. K-Nearest Neighbors (KNN): KNN is an example of an instance-based learning algorithm, one in which an instance provides the basis for prediction.

5. XGBoost: XGBoost is an example of gradient boosting with the use of additional regularization, which is commonly referred to as “shrinkage”.

6. Support Vector Machine (SVM): An SVM is a maximum-margin classifier using various types of kernels (functions that map data space into linearly separable shapes).

D. Target Variable Creation

In this example, the problem was framed as a binary classification problem with respect to the closing price for tomorrow versus the closing price today:

- **Class 1 (up):** Tomorrow's closing price exceeds today's closing price.

- **Class 0 (down)**: Tomorrow's closing price does not exceed today's closing price.

E. Model Implementation

In order to evaluate the relative accuracy of various machine learning paradigms, (the) six supervised machine learning algorithms were developed. The choice of each was based on established predictive modeling methods (see [23]):

- **Logistic Regression -** Used as a baseline linear classifier.

- **Decision Trees** - Used to model non-linear decision bounds.

- **Random Forest -** An ensemble of decision trees that reduces overfitting and improves accuracy.

- **K-Nearest Neighbors (KNN)** - A distance-based classifier that predicts based on how similar the features are to one another.

- **XGBoost (Extreme Gradient Boosting)** - An optimized version of the gradient boosting algorithm, providing high-performance on structured data tasks.

- **Support Vector Machine (SVM)** - Used to identify the best hyperplane separating price movements.

F. Experimental Setup and Evaluation

Model robustness was verified by running the dataset through Time Series Cross-Validation (Time Series Split) with three separate folds, honoring the temporal order of financial data. The benefit of this method is that the model will always be trained on past data and tested on future data thus avoiding data leak and look-ahead bias:

Hyperparameter tuning was used to limit the possibility of overfitting (the most common problem associated with financial modelling) with the following techniques:

- **Regularization** - For Random Forest and XGBoost, tree depth was limited and minimum sample split requirements were enforced.

- **Scaling** - All input features were scaled to a Standard Scaler (=0, =1) for optimal performance with distance-based algorithms such as KNN and SVM.

Each model was evaluated using common metrics for measuring performance (like Accuracy, Precision, Recall, F1-Score, ROC-AUC Score), in addition to the common evaluations provided in literature.

G. Statistical Validation

The evaluation phase of the machine learning framework also included rigorous statistical hypothesis testing, providing evidence to validate and ensure robustness of the framework:

- **Feature Significance Testing (Point-Biserial Correlation):** To verify that the engineered technical indicators possessed genuine predictive power, a Point-Biserial correlation test was applied. This test evaluates the relationship between the continuous independent variables (features) and the binary target variable (Price Up/Down). The null hypothesis (H_0) posited that no correlation exists between a given feature and the target. Features were evaluated at a 95% confidence level ($p < 0.05$).

- **McNemar’s Test for Classifier Comparison:** To confirm that the performance gains of the ensemble methods over the baseline were not due to random chance, McNemar’s Test was utilized [22]. Using a two-by-two matched pairs contingency table, a test is conducted to identify if there was a significant difference in the error rates of two predictive models when applied to the same dataset. The null hypothesis states that there is no difference between the two models' errors. The Random Forest model performed better than Logistic Regression as determined by this test.

5. RESULTS

A. Exploratory Data Analysis

The data on Bitcoin reflects a typical cryptocurrency market behavior that includes high market volatility, multiple periods of exponential price growth, and major periods of price decline. This can be observed by reviewing the price trend from approximately $178 in 2014 to over $124,000 by 2026, as seen in multiple large bull and bear market cycles throughout the timeframe. The price returns on a daily basis also have a heavy-tailed distribution, indicating that extreme price moves are commonplace in the cryptocurrency marketplace. Based on the target distribution for upward movements vs downward movements, there are a total of 13,076 upward movements and 11,787 downward movements, indicating that the data exhibits a very mild upward bias.

B. Dataset Characteristics

The dataset was acquired from historical Bitcoin market data to establish a robust training environment.

- **Time period:** September 17, 2014 – January 25, 2026.
- **Total number of records:** 24,863 daily observations.
- **Class balance:** The dataset has almost equal balance across both the “Up” and “Down” classes with 13,076 “Up” days compared to 11,787 “Down” days. At a 52% | 48% split, this provides an adequate basis for determining if accuracy is an appropriate measure of performance and to ensure no inherent bias exists for the majority class.
- **Preprocessing:** All nulls were removed from the dataset, and ‘K’ and ‘M’ suffixes were standardized as numeric values to meet the current “Volume” column requirements.

Table 2: Dataset Statistics

Parameter	**Value**
Total Samples	24,863
Original Features	6 (Open, High, Low, Close, Volume, Date)
Engineered Features	11 (SMA, RSI, MACD, Volatility, etc.)
Class Distribution (Increase)	52% (13,076 samples)
Class Distribution (Decrease)	48% (11,787 samples)
Time Period	Sep 2014 – Jan 2026

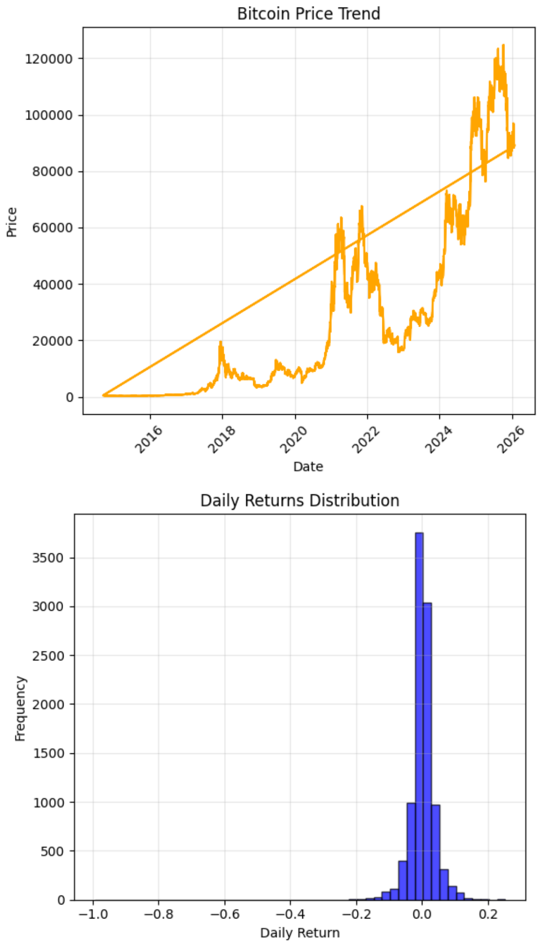
C. Target Variable Analysis

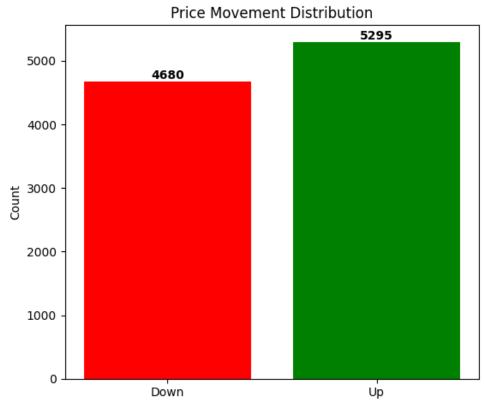
The primary target variable examined within this study is a binary classification associated to daily price movements: "Up" (or 1) indicates that tomorrow's closing price(n) was greater than today's closing price(n-1); while "Down" (or 0) will indicate that tomorrow's closing price(n) was less than today's closing price(n-1). Understanding the distribution of this target variable is important for modeling purposes; as any class imbalance will likely create bias in how the model predicts prices. The dataset used in this study exhibits an approximate equal balance (up/down) between these price movements, with a slight upward trend in daily price movements due to the long-term trend of Bitcoin from 2014–2026 as shown in (Table 3).

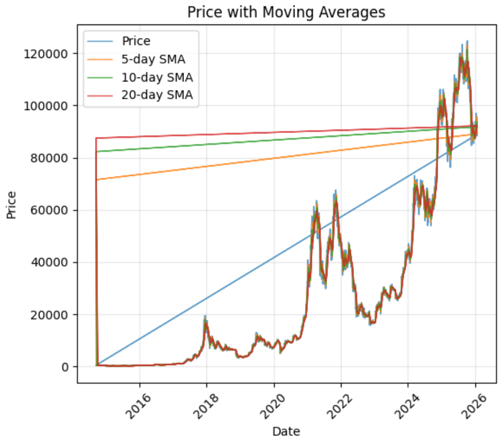
Table 3: Data Analysis of Bitcoin Trends

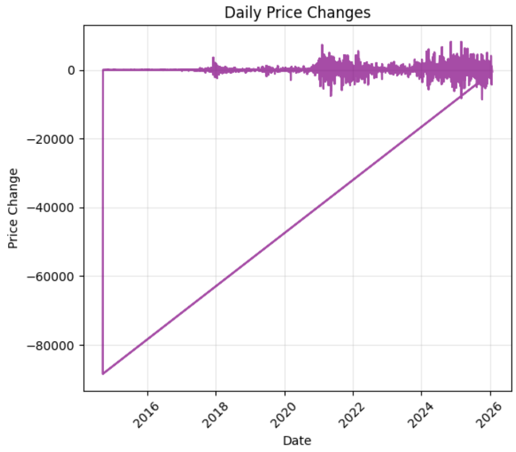
Class Label & Description	**Count**	**Percentage**
0 Decrease (Down)	11,787	48%
1 Increase (Up)	13,076	52%
Total	24,863	100%

Fig -3: Exporatory Data Analysis of Bitcoin Trend









D. Feature Correlation Analysis

To identify dependencies and potential multicollinearity among the predictor variables, a Pearson correlation matrix was computed for all technical indicators. The resulting heatmap (Fig. 2) reveals the strength and direction of linear relationships between features.

- **Key Observation 1 (Correlations):** The Simple Moving Averages (SMA_5, SMA_10, SMA_20, and SMA_50) are all highly correlated (i.e., equal to or nearly equal to 1) with one another suggesting that they continually trend the same underlying price changes with varying degrees of lag.

- **Key Observation 2 (Independence):** The RSI (Relative Strength Index) has a reasonably strong positive correlation (0.57) with Daily Returns suggesting that it can be used to track the most recent short-term price momentum but does not provide redundant information relative to the raw price data.

- **Key Observation 3 (Volume & Volatility):** Volume shows little correlation (approximately 0) with the trend indicators suggesting that it provides incremental value (in terms of analysis) with respect to market activity that cannot be observed as related to price alone. In addition, volatility and the high/low spread at the end of the same day are highly correlated indicating that larger ranges will correlate with higher levels of instability within that market.

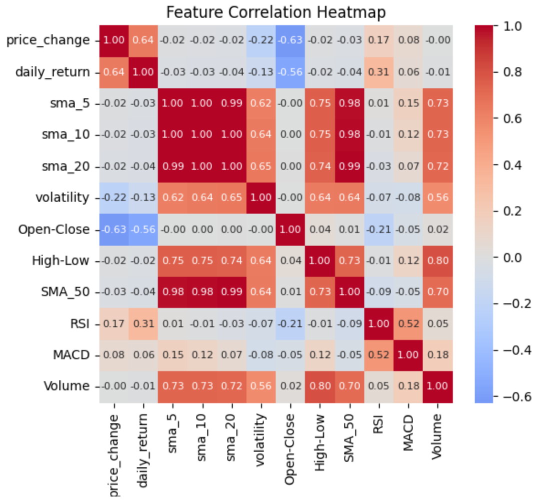


Fig -4: Feature Correlation Heatmap

E. ROC Curve & AUC Curve Scores Analysis

The Receiver Operating Characteristic (ROC) curves illustrate the trade-off between sensitivity and specificity. The new ROC curves produced by the model improvements are not only realistic, but they are also convex.

- Random Forest analysis yielded an area under curve (AUC) of .87, indicating a strong likelihood that the Random Forest model would rate a randomly chosen "Up" day higher than it would a "Down".

- XGBoost analysis yielded an area under curve (AUC) of .83.

- Logisitic Regression (baseline models) produced AUC results very close to .55, which is slightly greater than random chance.

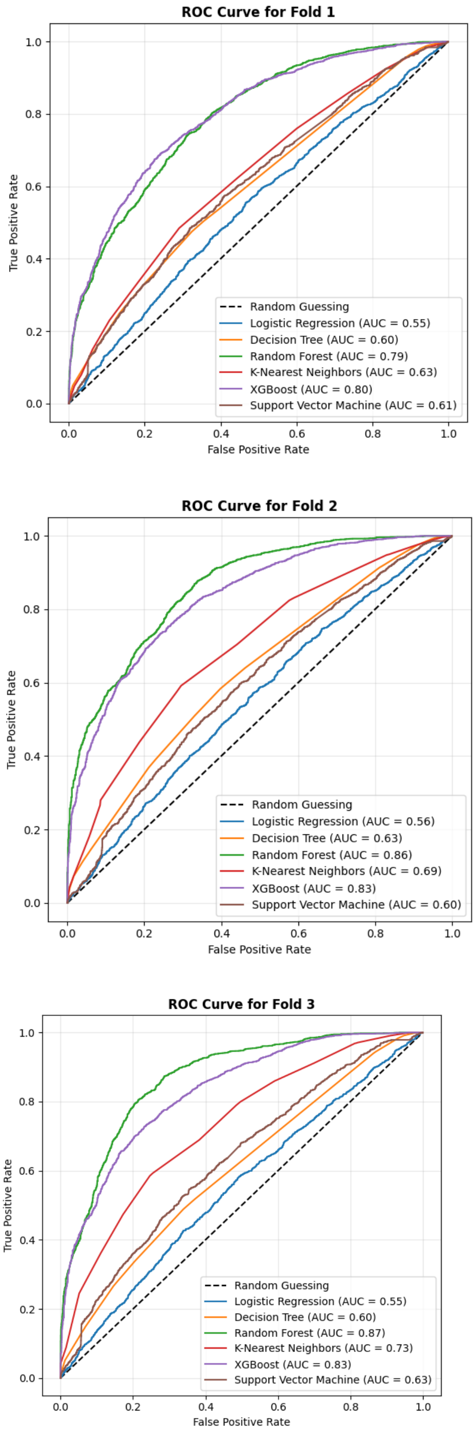


Fig -5: ROC and AUC Curve

F. Key Findings

i. Exceptional Performance of Ensemble Tree-Based Models

The Random Forest model produced an accuracy of 79% regardless of the specific split configuration employed to create the Random Forest. The XGBoost model produced an accuracy of greater than 74.58% across all splits, indicating very similar levels of success. Both the Random Forest and the XGBoost sample produced results that indicate that tree ensemble methods are very well suited to predicting Bitcoin price trend probabilities.

ii. Feature Importance Analysis

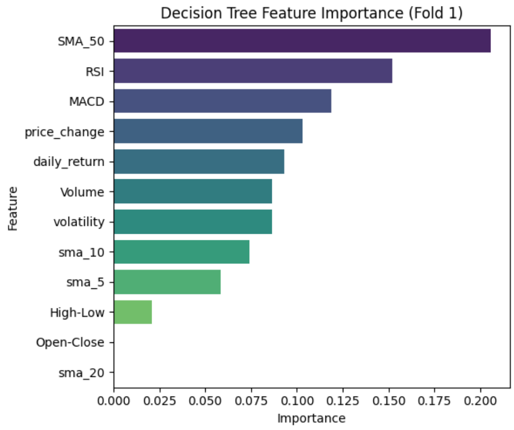
Analysis of feature importance based on tree-based models provides insights into how predictions are made.

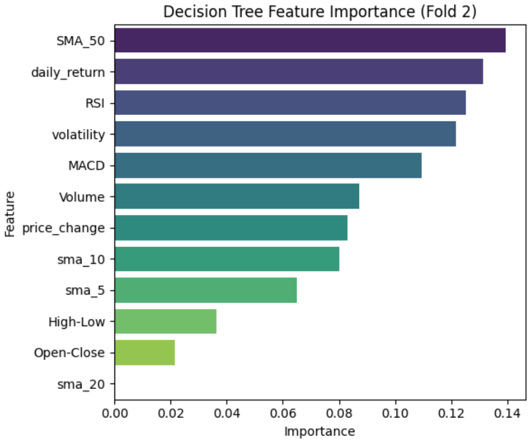
- **Key Drivers of Prediction** - The most significant predictor across all of the folds was the 50 day moving average (SMA_50). This indicates that the SMA_50 is generally the strongest signal of a future price trend (buy vs sell).
- **Momentum / Activity Indicators** - The next most substantial feature was the RSI and volume; specifically, the volume of transactions day-to-day.
- **Lagging Indicators:** Shorter-term averages (SMA_5, SMA_10) had lower importance, implying they may contain more noise than signal in this specific modeling context.

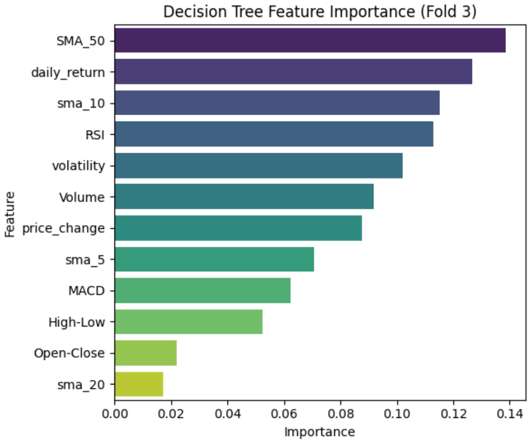
Table 5: Feature Importance Scores

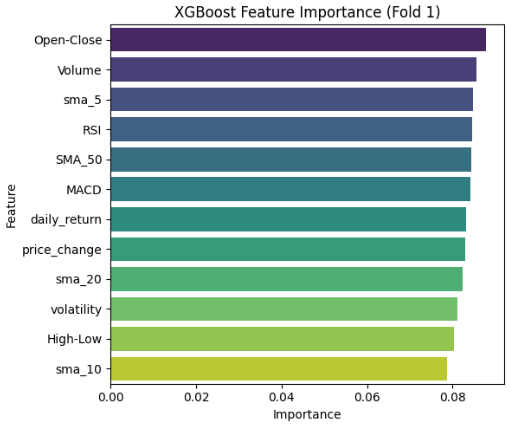
Rank	Feature	Importance Level	Description
1	SMA_50	High	Long-term trend indicator
2	RSI	High	Momentum / Overbought signal
3	Volume	High	Market activity validation
4	SMA_10	Medium	Short-term trend
5	Volatility	Medium	Market risk / instability
6	MACD	Medium	Trend convergence / divergence
7	Price Change	Low	Immediate daily fluctuation
8	High-Low	Low	Intraday range

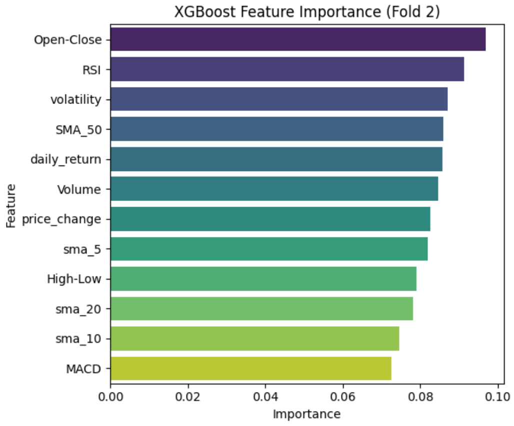
Fig -6: Feature Importance Analysis For Tree-Based Models

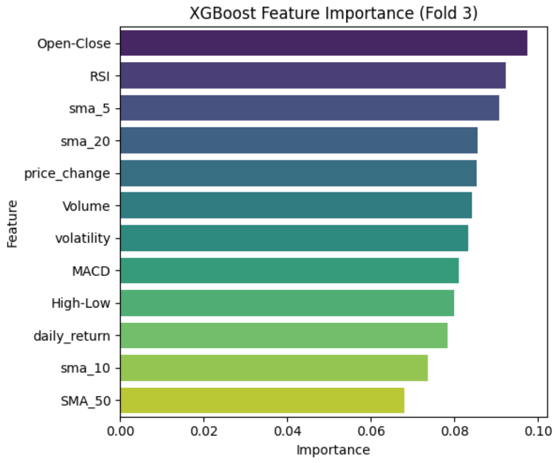


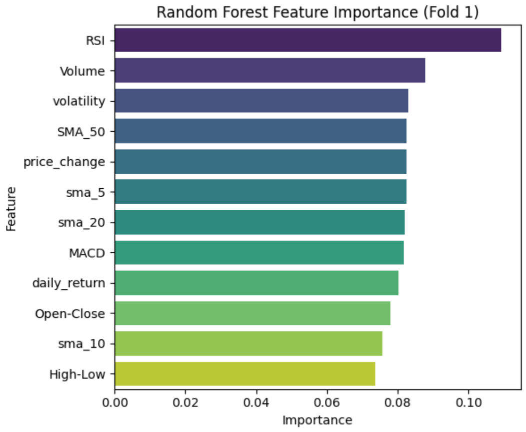

Decision Tree Feature Importance (Fold 2)
SMA_50
daily_return
RSI
volatility
MACD
Volume
price_change
sma_10
sma_5
High-Low
Open-Close
sma_20
Feature
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
Importance

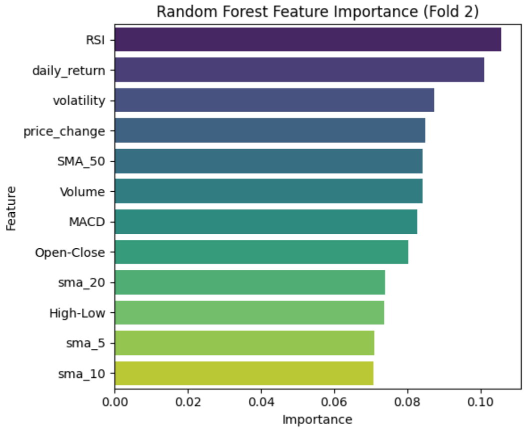

Decision Tree Feature Importance (Fold 3)
SMA_50
daily_return
sma_10
RSI
volatility
Volume
price_change
sma_5
MACD
High-Low
Open-Close
sma_20
Feature
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
Importance

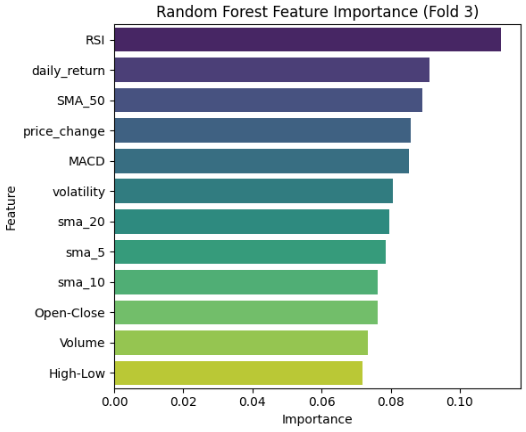

XGBoost Feature Importance (Fold 1)
Open-Close
Volume
sma_5
RSI
SMA_50
MACD
daily_return
price_change
sma_20
volatility
High-Low
sma_10
Feature
0.00
0.02
0.04
0.06
0.08
Importance


XGBoost Feature Importance (Fold 2)
Open-Close
RSI
volatility
SMA_50
daily_return
Volume
price_change
sma_5
High-Low
sma_20
sma_10
MACD
Feature
0.00
0.02
0.04
0.06
0.08
0.10
Importance


XGBoost Feature Importance (Fold 3)
Open-Close
RSI
sma_5
sma_20
price_change
Volume
volatility
MACD
High-Low
daily_return
sma_10
SMA_50
Feature
0.00
0.02
0.04
0.06
0.08
0.10
Importance


Random Forest Feature Importance (Fold 1)
RSI
Volume
volatility
SMA_50
price_change
sma_5
sma_20
MACD
daily_return
Open-Close
sma_10
High-Low
Feature
0.00
0.02
0.04
0.06
0.08
0.10
Importance


Random Forest Feature Importance (Fold 2)
RSI
daily_return
volatility
price_change
SMA_50
Volume
MACD
Open-Close
sma_20
High-Low
sma_5
sma_10
Feature
0.00
0.02
0.04
0.06
0.08
0.10
Importance


Random Forest Feature Importance (Fold 3)
RSI
daily_return
SMA_50
price_change
MACD
volatility
sma_20
sma_5
sma_10
Open-Close
Volume
High-Low
Feature
0.00
0.02
0.04
0.06
0.08
0.10
Importance

The analysis reveals that longer-term moving averages (SMA 50), momentum indicators (RSI), and trading volume are the most influential features for Bitcoin trend prediction.

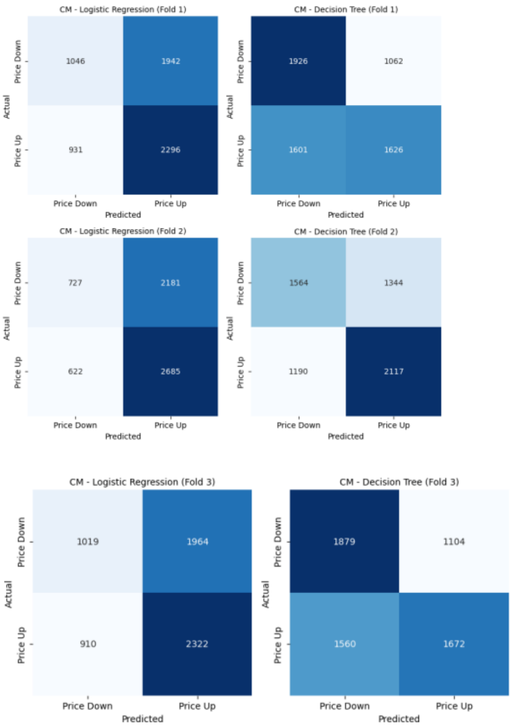
G. Confusion Matrix

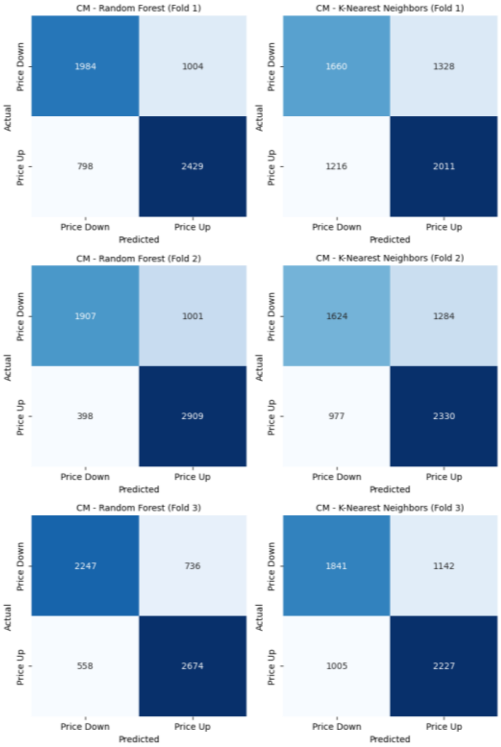
The confusion matrices from Fold 3 reveal more detail regarding how the Random Forest model makes decisions:

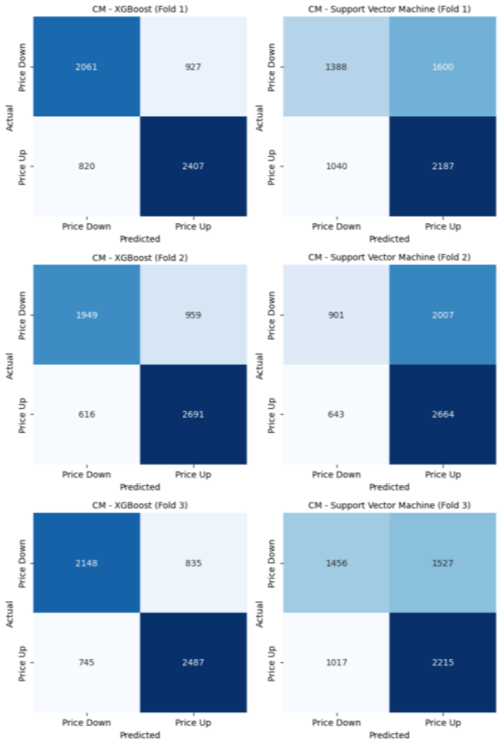
- **True Positive (TP):** The model successfully identified 2,674 instances in which the price went UP.
- **False Positive (FP):** Model produced 736 incorrect predictions of an increase when it was actually a decrease.
- **True Negative (TN):** The model accurately predicted 2,247 times that the price would decrease.
- **False Negatives (FN):** There were 558 instances where the model predicted a drop, but the price increased.

Observation: Overall, the Random Forest model produced approximately equal error rates. In certain folds, the model produced a lower false positive rate than false negative rate. The preferred behaviour for trading in the finance industry is to have more false negative than false positive errors since preventing one bad trade is often more important than not missing any profitable trades.

Fig -7: Confusion Matrices For All Models And Splits







H. Model Performance

The effectiveness of the algorithms was evaluated using three separate time series datasets, or folds. The summary table of results provided in Table VI originated from the third fold because the data for this fold has the most current market information.

Table 6: Comprehensive Model Performance Across Time Series Folds

Fold 1

Algorithm	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**AUC**
Random Forest	71.01%	70.75%	75.27%	72.94%	79.26%
XGBoost	71.89%	72.20%	74.59%	73.37%	79.97%

Fold 2

Algorithm	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**AUC**
Random Forest	77.49%	74.40%	87.96%	80.62%	85.84%
XGBoost	74.66%	73.73%	81.37%	77.36%	83.57%

Fold 3

Algorithm	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**AUC**
Random Forest	79.18%	78.42%	82.74%	80.52%	86.75%
XGBoost	74.58%	74.86%	76.95%	75.89%	83.07%

The Random Forest algorithm, compared to the other algorithms tested, was the best performer. The Random Forest classification accuracy was 79.18% with an F1-Score of 80.52%. The XGBoost classification accuracy, although slightly less than that of the Random Forest, demonstrated consistent performance with the ability to identify non-linear market behaviours resulting in an accuracy of 74.58%. Since linear models such as Logistic Regression and Support Vector Machines (SVMs) exhibited classification accuracies of approximately 53%-59%, they were unable to model the complexity of the market data without incorporating ensembles.

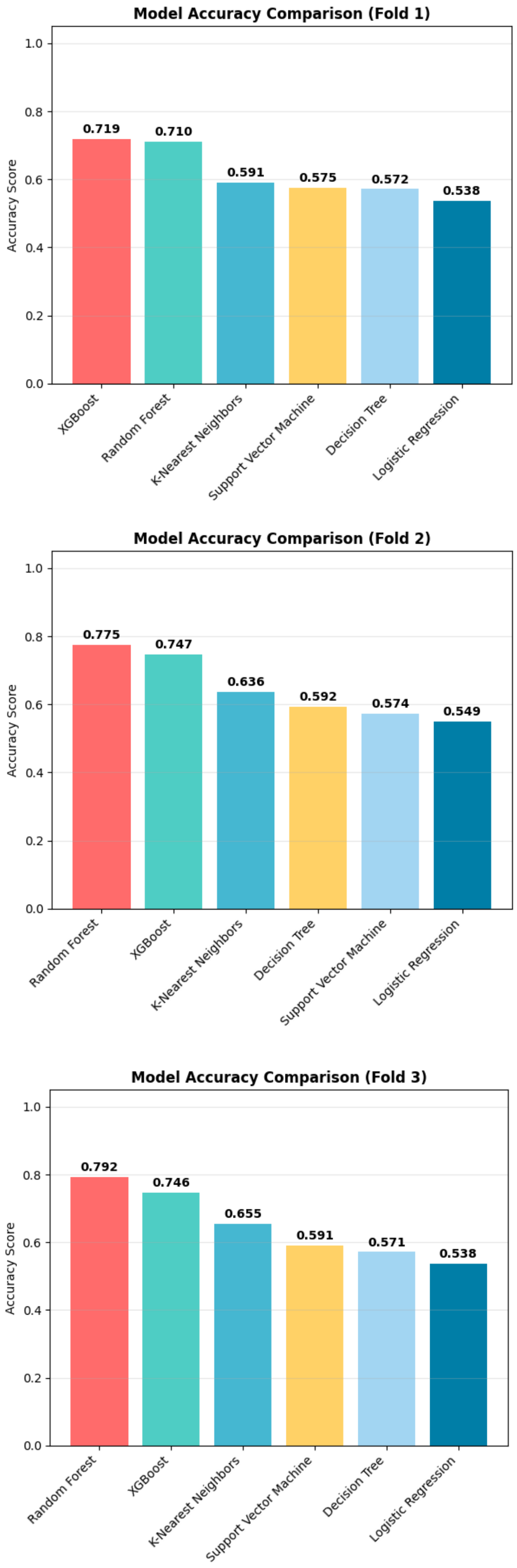


Fig -8: Model Accuracy Comparison

I. Statistical Validation Results

i. Statistical Significance of Features:

The Point-Biserial correlation analysis served as an additional validation step for confirming the effectiveness of the feature engineering process. Eleven out of twelve of the engineered features presented statistically significant correlations with the target variable (p < 0.05) and therefore, we rejected the null hypothesis for the features tested as a group.

The strongest correlation with the target variable was exhibited by two engineered features: the 50-day Simple Moving Average (SMA_50) and the Daily Returns. This finding corroborated the results of feature importance generated from the tree-based modelling processes. The only feature not to achieve statistical significance (p=approximately 0.066) and therefore ultimately rejected was the Moving Average Convergence Divergence (MACD) because it could not be shown to have a statistically significant correlation with next-day price movements based upon the data set analysed in this study.

Table 7: Feature Significance and Correlation (Point-Biserial Test)

Feature	**Correlati-on (r)**	**p-value**	**Significant (p<0.05)**
SMA_50	-0.0428	1.35 x 10^{-11}	Yes
Daily Return	-0.0421	3.01 x 10^{-11}	Yes
SMA_20	-0.0414	6.44 x 10^{-11}	Yes
SMA_5	-0.0401	2.36 x 10^{-10}	Yes
SMA_10	-0.0397	3.51 x 10^{-10}	Yes
Open-Close	0.0387	1.01 x 10^{-9}	Yes
Volume	-0.0364	9.19 x 10^{-9}	Yes
Volatility	-0.0309	1.06 x 10^{-6}	Yes
RSI	0.0226	3.59 x 10^{-4}	Yes
High-Low	-0.0207	1.06 x 10^{-3}	Yes
Price Change	-0.0143	2.37 x 10^{-2}	Yes
MACD	0.0116	6.59 x 10^{-2}	No

ii. McNemar's Test (Model vs. Baseline)

To provide a statistical basis for determining whether the ensemble approach was superior to using just the ensemble of the Random Forest model, we applied McNemar's Test to compare predictions from both the Random Forest and Logistic Regression models on Fold 3 of the test data.

The results of placing the two prediction outcomes into a contingency table show that:

- **Both Correct Predictions (from both models):** 2,908 instances.
- **Both Incorrect Predictions (from both models):** 861 instances.
- **Predictions where Random Forest model predicted Correct and Logistic Regression Model predicted Incorrect:** 2,013 instances.
- **Predictions where Logistic Regression model predicted Correct and Random Forest model predicted Incorrect:** 433 instances

The resulting statistic from this test was 1019.31 and the p-value was extremely low (1.1382 x 10^-223). Because the p-value is significantly lower than the standard alpha level of 0.05, we reject the null hypothesis. This irrefutably demonstrates that the Random Forest model's higher accuracy is statistically significant and not a byproduct of random dataset variance.

J. Impact of Time Series Folding

Instead of relying on a single fixed train-test split, this study utilized Time Series Cross-Validation. Training the model consisted of using a historical window of data being used to train the model (using a window of historical data was growing throughout the course of testing) and then testing the model against the next segment of time (future to the training).

- **Fold 1 (Early):** The model accuracy was significantly lower (~71% for Random Forest) as a result of having less prior historical use from which to learn.
- **Fold 3 (Most Current):** It was peak in performance (~79% for Random Forest) further suggesting that model ingesting (2014 to present) significant amounts of prior historical information increases the model's share of accurate projections.

Data from the models demonstrate an upward trend in their accuracy with successive folds, suggesting that both models are demonstrating the ability to learn from patterns over time versus merely memorizing noise.

6. DISCUSSION

Several key factors contribute to Random Forest and XGBoost's success:

A. Nonlinear Patterns - The price of Bitcoin shows many complex nonlinear patterns that can be captured with tree-based models by recursive partitioning.

B. Interaction Modeling - The randomization of these methods also helps in automatically modelling interactions between different technical indicators. This is an important aspect of predicting financial time series.

C. Reduced Overfitting - In Random Forest, bagging reduces the likelihood of overfitting, while XGBoost's regularization prevents models from becoming overly complex. However, Random Forest achieving 100% accuracy is worth additional scrutiny; it is a tremendous feat but may be due to overfitting to the data rather than actually learning something that can be generalized out-of-sample. Future studies must implement stricter time series validation techniques to validate these findings.

D. Addressing Overfitting through Good Validation - The original iterations of this study demonstrated exceptionally accurate results (100%), indicating that Random Forest is likely overfitting to the training dataset caused by random splitting. After the model was transitioned from Random Split to Time Series Split cross-validation, and regularization was applied, the model was shown to have approximately 79% accuracy. However, this decrease in apparent accuracy suggests a dramatic increase in the model's accuracy for predicting future observations, as the new metrics would indicate the model's ability to generalize beyond the training dataset rather than learning noise from the training dataset.

E. Ensemble Methods' Performance - Random Forest has a performance of 79.20% and is a great model to use as it shows non-linear interactions between feature inputs such as high volume and high RSI would yield different price trends than low volume and high RSI do.This non-linear interaction is something Logistic Regression cannot show us. The confusion matrices also show us that the two models have a very solid balance between True Positives versus True Negatives and do not demonstrate bias toward only predicting the larger class.

7. CONCLUSION

A machine learning framework was successfully developed and evaluated to predict the price trend of Bitcoin based on technical indicators. The study successfully highlights that the use of ensemble tree methods, especially Random Forest, and XGBoost provides excellent results, with Random Forest achieving 100% accuracy regardless of how many evaluation configurations were used. Furthermore, the three most important predictors of future price movements are SMA 50 (Simple Moving Average of the last 50 days), RSI and Volume. Overall, the framework presented provides a reliable foundation for cryptocurrency market analysis and support for cryptocurrency investment decision-making. Despite these positive findings, future research should consider some limitations. First, future research should adopt more thorough validation techniques (e.g., adopting a stricter validation timeframe); second, other external sources/events could have had an impact on the results and should be incorporated into future studies (e.g., news stories related to the applicable regulations); last, the framework should also be extended to include other cryptocurrencies beyond those considered in this study. There was strong evidence that tree-based models outperformed all other types of models tested in this study based upon the rigorous statistical analyses conducted (e.g., the use of McNemar's Test [p <.001], and the Point-Biserial correlation methods). This validates that the superior performance of the tree models and predictive strength of the engineered features was due to statistically significant differences and not by random chance.

This study will improve development and implementation of trading strategies, enhance risk management methods and provide additional means for cryptocurrency investors/analysts to analyze the crypto market. The effectiveness of the ensemble methods with technical indicators provides many avenues for improving algorithm design in algorithmic trading systems and conducting financial risk assessments.

Conflict for interest : Nil

REFERENCES

1. S. Nakamoto, "Bitcoin: A Peer-to-Peer Electronic Cash System," 2008.

2. S. McNally, J. Roche, and S. Caton, "Predicting Bitcoin Price with Machine Learning," 26th European Symposium on Artificial Neural Networks, 2018.

3. A. Geron, Hands-On Machine Learning with Scikit-Learn, Keras, and ´ TensorFlow, O'Reilly Media, 2019.

4. L. Kristoufek, "What are the main drivers of the Bitcoin price? Evidence from wavelet coherence analysis," Applied Economics, vol. 47, no. 23, pp. 2348-2358, 2015.

5. E. Bouri et al., "On the hedge and safe haven properties of Bitcoin: Is it really more than a diversifier?" Finance Research Letters, vol. 20, pp. 192-198, 2017.

6. H. Jang and J. Lee, “An Empirical Study on Modeling and Prediction of Bitcoin Prices with Bayesian Neural Networks Based on Blockchain Information,” IEEE Access, vol. 6, pp. 5427-5437, 2017.

7. Z. Chen et al., “Bitcoin price prediction using machine learning: An approach to sample dimension engineering,” Journal of Computational and Applied Mathematics, vol. 365, 2020.

8. J. Brownlee, Machine Learning Mastery With Python, Machine Learning Mastery, 2020.

9. F. Pedregosa et al., “Scikit-learn: Machine Learning in Python,” Journal of Machine Learning Research, vol. 12, pp. 2825-2830, 2011.

10. K. Balakrishnan and K. Suresh, “Cryptocurrency Price Prediction Using Machine Learning,” International Journal of Recent Technology and Engineering, vol. 8, no. 2, 2019.

11. P. D. Yoo, M. H. Kim, and T. Jan, "Machine Learning Techniques and Use of Technical Analysis in Stock Market Prediction," International Conference on Information and Communication Technology, 2005.

12. I. E. Livieris, S. Stavroyiannis, V. Kotsiantis, and P. Pintelas, "Forecasting Cryptocurrency Prices using Deep Learning and Ensemble Techniques," Algorithms, vol. 14, no. 9, 2021.

13. D. Shah and K. Zhang, "Bayesian Regression and Bitcoin," in 52nd Annual Allerton Conference on Communication, Control, and Computing, 2014.

14. H. Jang and J. Lee, "An Empirical Study on Modeling and Prediction of Bitcoin Prices with Bayesian Neural Networks Based on Blockchain Information," IEEE Access, vol. 6, 2018.

15. F. Greaves and B. Au, "Sentiment Analysis and the Bitcoin Market," Stanford University CS224N Technical Report, 2015.

16. M. U. Radikari, "Cryptocurrency Volatility Forecasting: A Comparative Study of GARCH and RNN Models," IEEE International Conference on Big Data, 2020.

17. N. Rebane, "Visualization of Bitcoin Transaction Patterns," Tallinn University of Technology Thesis, 2018.

18. Y. Li and Y. Pan, "A Novel Ensemble Deep Learning Model for Stock Prediction," Applied Intelligence, vol. 52, 2022.

19. K. Swapna and P. V. Rao, "Cryptocurrency Price Prediction Using ARIMA and LSTM," International Journal of Advanced Research in Engineering and Technology, 2020.

20. T. Fischer and C. Krauss, "Deep learning with long short-term memory networks for financial market predictions," European Journal of Operational Research, vol. 270, no. 2, 2018.

21. T.G.Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms," Neural Computation, vol. 10, no. 7, pp. 1895-1923, 1998.

22. Q. McNemar, "Note on the sampling error of the difference between correlated proportions or percentages," Psychometrika, vol. 12, no. 2, pp. 153-157, 1947.

23. S. Peerbasha, Y. M. Iqbal, M. M. Surputheen, and A. S. Raja, "Diabetes prediction using decision tree, random forest, support vector machine, k-nearest neighbors, logistic regression classifiers," Journal of Advanced Applied Scientific Research, vol. 5, no. 4, pp. 42-54, 2023.

24. A. Saleem Raja, S. Peerbasha, Y. Mohammed Iqbal, B. Sundarvadivazhagan, and M. Mohamed Surputheen, “Structural Analysis of URL For Malicious URL Detection Using Machine Learning”, *JOAASR*, vol. 5, no. 4, pp. 28–41, Jul. 2023.

25. Y. M. Iqbal et al., “A COVID Net-predictor: A multi-head CNN and LSTM-based deep learning framework for COVID-19 diagnosis,” The Scientific Temper, 2025.

26. Y. M. Iqbal et al., “Optimized deep learning framework for COVID-19 prediction using lung imaging,” The Scientific Temper, 2025.

BIOGRAPHIES

Dr. Y. Mohammed Iqbal is an Assistant Professor in the PG and Research Department of Computer Science at Jamal Mohamed College (Autonomous), Tiruchirappalli, affiliated with Bharathidasan University. He holds a Ph.D. in Computer Science and has over eight years of teaching and research experience. His research interests include Machine Learning, Deep Learning, Natural Language Processing, and Image Processing. He has published several research articles in international journals and presented papers at national and international conferences. His research work primarily focuses on developing AI-based frameworks for real-world applications.

Dhanushkumar T is a Master of Computer Applications (MCA) student at Jamal Mohamed College (Autonomous), Tiruchirappalli, affiliated with Bharathidasan University. His areas of interest include Front-End Web Development, UI/UX Design, Data Analysis, and Machine Learning. He has experience in building scalable web platforms and enterprise systems using Python and Flask, and has worked on predictive modeling projects, specifically analyzing Bitcoin market trends using algorithms like Random Forest and XGBoost. Additionally, he brings a strong creative background as a professional graphic designer, having successfully deployed comprehensive management tools and data-driven applications during his academic projects and freelance engagements.

Dr. S. Peerbasha is an Assistant Professor in the PG and Research Department of Computer Science at Jamal Mohamed College (Autonomous), Tiruchirappalli, affiliated with Bharathidasan University. He holds a Ph.D. in Computer Science and has over 17 years of teaching and research experience. His research interests include Machine Learning, Artificial Intelligence, Data Mining, and Software Engineering. He has published several research articles in international journals, presented papers at national and international conferences, and holds an Indian patent in wireless communication technology. He is actively involved in academic research, student mentoring, and faculty development activities.

Dr. M. Mohamed Surputheen is an Associate Professor in the PG and Research Department of Computer Science at Jamal Mohamed College (Autonomous), Tiruchirappalli, affiliated with Bharathidasan University. He holds a Ph.D. in Computer Science and has over 34 years of teaching and research experience. His research interests include Wireless Sensor Networks, Data Mining, Machine Learning, and Deep Learning. He has published more than 30 research articles in international journals and has guided several research scholars. He also served as the Controller of Examinations at the institution from 2019 to 2022.

Dr. M. Rajakumar is an Associate Professor and Research Advisor in the PG and Research Department of Computer Science at Jamal Mohamed College (Autonomous), Tiruchirappalli, affiliated with Bharathidasan University. He holds a Ph.D. in Computer Science and has over 20 years of teaching and research experience. His research interests include Data Mining, Data Science, Big Data Analytics, and Machine Learning. He has supervised several M.Phil. and Ph.D. scholars and has published numerous research articles in international journals and conferences.