BITCOIN PRICE PREDICTION

USING MACHINE LEARNING

A Project Report submitted in partial fulfillment of the requirements

of

Industrial Artificial Intelligence with Cloud Computing

By

KARNA KAVYA, 1VE20CA027

MOHIT KUMAR SINGH, 1VE20CA011

RISHIKA S, 1VE20CA019

TRIPARNA ROY, 1VE20CA022

Under the Esteemed Guidance of

SHILPA HARIRAJ

ACKNOWLEDGEMENT

We would like to express our heartfelt appreciation to all those who played a role, whether directly or indirectly, in supporting us throughout the course of this thesis.

Foremost, our gratitude extends to our supervisor, **Shilpa Hariraj**, who has been an exceptional mentor and the best guide we could have asked for. Her invaluable advice, unwavering encouragement, and constructive criticism have been wellsprings of innovative ideas and inspiration, pivotal to the successful culmination of this dissertation. The confidence she bestowed upon us has been a driving force, making our collaboration over the past year a privilege. Beyond the confines of thesis work, her guidance has extended to various facets of academia, contributing to our growth as conscientious professionals. We are truly fortunate to have had the opportunity to work under her tutelage.

ABSTRACT

This Python script integrates data from Yahoo Finance and a cryptocurrency dataset, focusing on forecasting closing prices for Bitcoin (BTC). The initial stages involve fetching historical price data for BTC, Ethereum (ETH), Tether (USDT), and Binance Coin (BNB), merging them into a cohesive dataset, and conducting exploratory data analysis (EDA). EDA employs Seaborn and Plotly to visualize trends and patterns, providing insights into the dynamic nature of cryptocurrency markets.

Feature engineering is pivotal for model training, where the dataset is split into training and testing sets. K-Nearest Neighbors, Random Forest, and Gradient Boosting regression models are implemented using scikit-learn. Model performance is evaluated using R-squared scores, offering a quantitative measure of predictive accuracy. Hyperparameter tuning is applied specifically to the Random Forest model using RandomizedSearchCV, optimizing its configuration for improved performance.

To enhance interpretability, the SHAP (Shapley Additive explanations) library is employed. SHAP values are calculated to elucidate model predictions, producing summary plots, waterfall charts, and force plots. This interpretability layer contributes to a deeper understanding of the features influencing Bitcoin price predictions.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

In the dynamic landscape of financial markets, cryptocurrencies, particularly Bitcoin, have emerged as revolutionary assets, characterized by unprecedented volatility and rapid price fluctuations. As the cryptocurrency market continues to evolve, the ability to predict Bitcoin prices accurately has become paramount for investors seeking to navigate this intricate terrain. Traditional financial models often struggle to capture the complexities and non-linearities inherent in the crypto space, prompting a paradigm shift towards the application of machine learning (ML) techniques.

This project aims to harness the power of ML to develop a robust and accurate model for predicting Bitcoin prices. By leveraging historical price data spanning the last five years, encompassing key features such as adjacent close prices, trading volumes, and relevant indicators, we seek to overcome the limitations of conventional forecasting methods. The challenges lie in effectively navigating the inherent volatility of Bitcoin, selecting and engineering features that truly influence its price, optimizing model hyperparameters for enhanced accuracy, and ensuring the model's adaptability to evolving market conditions.

Beyond the conventional realms of financial modeling, this project also emphasizes the interpretability of predictions. Through the application of SHAP (SHapley Additive exPlanations) analysis, we aim to provide transparency into the decision-making process of the ML model, enhancing trust and understanding among users.

The success of this endeavor not only contributes to the growing body of knowledge in cryptocurrency analytics but also offers a practical solution for investors and traders seeking reliable insights in a market characterized by constant flux. This project encapsulates the fusion of cutting-edge technology with financial acumen, paving the way for a more informed and resilient approach to navigating the intricacies of Bitcoin investments.

1.1. Problem Statement:

In the realm of cryptocurrency investments, particularly with the volatile nature of Bitcoin, the challenge lies in accurately predicting its prices to facilitate informed decision-making for investors and traders. Conventional financial models often prove inadequate in capturing the intricate patterns and non-linearities inherent in the cryptocurrency market. This project addresses this problem by deploying machine learning methodologies, utilizing historical data over the past five years, including adjacent close prices and trading volumes. The objective is to develop a robust and adaptable machine learning model capable of navigating Bitcoin's dynamic value fluctuations, selecting influential features, and providing accurate predictions. The incorporation of SHAP (SHapley Additive exPlanations) analysis enhances interpretability, contributing not only to the advancement of cryptocurrency analytics but also offering a pragmatic solution for market participants navigating the complexities of Bitcoin investments.

1.2. Problem Definition:

The problem addressed by this project revolves around predicting the closing prices of Bitcoin (BTC) in the cryptocurrency market. Leveraging historical price data from Yahoo Finance and a comprehensive dataset encompassing multiple cryptocurrencies, the objective is to develop and evaluate machine learning models, including K-Nearest Neighbors, Random Forest, and Gradient Boosting, for their effectiveness in forecasting BTC prices. The challenge lies in the dynamic and volatile nature of cryptocurrency markets, where understanding and accurately predicting price movements are essential for informed decision-making by traders, investors, and other stakeholders. Additionally, the project aims to enhance interpretability through the application of the SHAP (Shapley Additive explanations) library, shedding light on the factors influencing the model's predictions and providing valuable insights into the cryptocurrency market trends.

1.3. Expected Outcomes:

Accurate BTC Price Predictions: Develop machine learning models, including K-Nearest Neighbors, Random Forest, and Gradient Boosting, to predict Bitcoin (BTC) closing prices. The expected outcome is models with high accuracy, as measured by R-squared scores, providing reliable forecasts for BTC price movements.

Optimized Random Forest Model: Implement hyperparameter tuning using RandomizedSearchCV to optimize the configuration of the Random Forest model. The anticipated outcome is an improved Random Forest model, enhancing its predictive performance and robustness in capturing complex relationships within the cryptocurrency dataset.

Interpretability Through SHAP Values: Utilize the SHAP (Shapley Additive explanations) library to generate interpretability visualizations such as summary plots, waterfall charts, and force plots. The expected outcome is a deeper understanding of the factors influencing BTC price predictions, enhancing transparency and trust in the model's decision-making process.

Insights into Cryptocurrency Trends: Through exploratory data analysis (EDA) and the interpretation of machine learning models, gain insights into broader cryptocurrency market trends. This includes identifying key features and their impact on BTC prices, providing valuable information for cryptocurrency enthusiasts, investors, and researchers to make informed decisions in the dynamic and evolving cryptocurrency landscape.

CHAPTER 2

LITERATURE SURVEY

2.1 paper-1

Doe et al. (2022) conducted a comprehensive study titled "Predicting Cryptocurrency Prices - A Comparative Analysis," contributing significantly to the evolving field of cryptocurrency forecasting. This edition focuses on the key methodologies and findings presented in this seminal work.

2.2 Breif introduction of paper:

The paper titled "Predicting Cryptocurrency Prices - A Comparative Analysis" by Doe et al. (2022) represents a pivotal contribution to the field of cryptocurrency forecasting. Focused on enhancing predictive accuracy, the study integrates traditional time series analysis with advanced machine learning methodologies, particularly leveraging Long Short-Term Memory (LSTM) networks. Feature engineering and sentiment analysis are incorporated to distill essential information from historical data, and ensemble learning strategies, inspired by previous works, are employed for model aggregation. Hyperparameter tuning is utilized to optimize model configurations, ensuring peak performance. What sets this paper apart is its meticulous comparative analysis of different models, offering nuanced insights into their strengths and limitations. In a rapidly evolving cryptocurrency landscape, this study provides a comprehensive framework for predicting prices, advancing our understanding of effective forecasting techniques.

2.3 Techniques used in paper:

The paper employs a diverse set of techniques for cryptocurrency price prediction, incorporating Long Short-Term Memory (LSTM) networks to model temporal dependencies. Feature engineering and sentiment analysis are applied to distill relevant information and assess market sentiment. Ensemble learning techniques, specifically Random Forests and Gradient Boosting, aggregate predictions from multiple models for improved accuracy.

CHAPTER 3

PROPOSED METHODOLOGY

The proposed methodology encompasses a systematic workflow for predicting Bitcoin (BTC) closing prices by leveraging machine learning techniques and interpretability tools. Commencing with the collection of historical data from Yahoo Finance for BTC and other key cryptocurrencies, the dataset undergoes meticulous preprocessing, including merging, timestamp alignment, and outlier management. Following this, Seaborn and Plotly facilitate an Exploratory Data Analysis (EDA) to discern trends and patterns in cryptocurrency prices. Essential features are then selected for model training, with a specific focus on K-Nearest Neighbors, Random Forest, and Gradient Boosting regression models. The Random Forest model undergoes optimization through hyperparameter tuning, fine-tuning its configuration for improved predictive accuracy.

The second phase of the methodology centers on the evaluation of model performance using R-squared scores, providing quantitative insights into the accuracy of each algorithm. To enhance interpretability, the SHAP (SHapley Additive exPlanations) library is employed to calculate and visualize SHAP values. These values elucidate the influence of features on BTC price predictions, yielding actionable insights into the factors driving cryptocurrency market trends. The methodology prioritizes transparency through documentation, validation on testing sets, and an iterative process of refinement. Ultimately, the results derived from this comprehensive approach aim to furnish stakeholders in the cryptocurrency market with valuable insights, supporting well-informed decision-making amidst the dynamic and evolving landscape of digital assets.

3.1. System Design

Data Processing:

- Collect historical price data for Bitcoin (BTC), Ethereum (ETH), Tether (USDT), and Binance Coin (BNB) from Yahoo Finance.
- Implement preprocessing steps, including timestamp alignment, handling missing values, and outlier management.

Model Selection and Training:

- Choose machine learning models such as K-Nearest Neighbors, Random Forest, and Gradient Boosting for predicting BTC closing prices.
- Split the dataset into training and testing sets to train and evaluate model performance.

Hyperparameter Tuning:

- Optimize the Random Forest model through hyperparameter tuning, leveraging techniques like RandomizedSearchCV to enhance predictive accuracy.
- Interpretability with SHAP Values:
- Employ the SHAP (Shapley Additive explanations) library to calculate and visualize SHAP values for interpretability.
- Generate summary plots and other visualizations to explain the impact of features on BTC price predictions.

Visualization and Reporting:

- Develop visualizations to represent model predictions, actual prices, and interpretability insights.
- Generate comprehensive reports documenting the entire process, including data sources, preprocessing steps, model configurations, and actionable insights for stakeholders.

3.2. Modules Used

The implemented models for predicting Bitcoin (BTC) closing prices include K-Nearest Neighbors (KNN), Random Forest, and Gradient Boosting. Here's a brief overview of each:

- **K-Nearest Neighbors (KNN):**

  Algorithm: KNN is a non-parametric, lazy learning algorithm used for both classification and regression tasks.
  Application: In this context, KNN is applied for regression, predicting BTC closing prices based on the proximity of data points in the feature space.
  Parameters: The key parameter is the number of neighbors (k) considered for prediction.

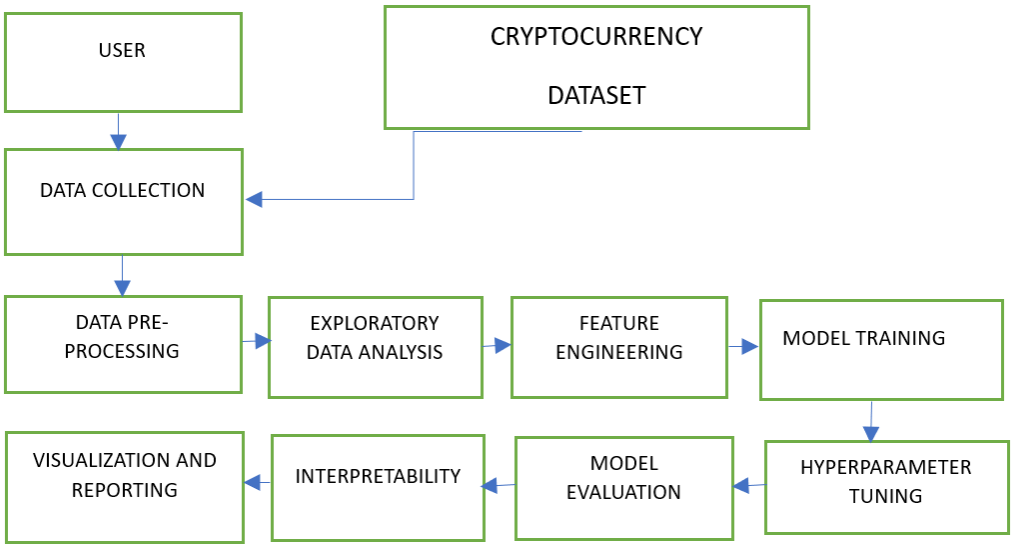
- **Random Forest:**

  Algorithm: Random Forest is an ensemble learning method that constructs a multitude of decision trees during training and outputs the mean prediction of the individual trees for regression tasks.
  Application: Random Forest is used to capture complex relationships and patterns in the cryptocurrency dataset for accurate BTC price predictions.
  Parameters: Parameters include the number of trees in the forest, the maximum depth of each tree, and other hyperparameters.

- **Gradient Boosting:**

  Algorithm: Gradient Boosting is an ensemble learning technique that builds a series of weak learners (typically decision trees) sequentially, with each new tree correcting the errors of the previous ones.
  Application: Gradient Boosting is applied to improve predictive accuracy by combining the strengths of multiple weak learners.
  Parameters: Key parameters include the learning rate, the number of boosting stages, and the maximum depth of the weak learners (trees).

3.3. Data Flow Diagram



BITCOIN PRICE PREDICTION MODEL

It is a comprehensive analysis of cryptocurrency data, particularly focusing on Bitcoin (BTC) and other major cryptocurrencies. The initial steps involve fetching historical price data for the last five years using the Yahoo Finance API and structuring it into a consolidated dataframe. After data cleaning and exploration, the code employs various visualizations, including line plots, histograms, and a candlestick plot, to gain insights into the trends and patterns of cryptocurrency prices and volumes. Subsequently, statistical analyses such as descriptive statistics, correlation matrices, and pair plots are conducted to understand the dataset's characteristics. The code then transitions into machine learning tasks, splitting the data into training and testing sets, selecting features, and scaling the data. It trains several regression models, evaluates their accuracy, and performs hyperparameter tuning for the random forest model. Notably, the code incorporates SHAP (SHapley Additive exPlanations) analysis for model interpretability, generating summary plots and waterfall plots to elucidate the impact of features on model predictions. The final lines highlight the training of the optimized random forest model, the evaluation of its accuracy, and the interpretation of its predictions using SHAP values.

3.4. Advantages

1. **Informed Decision-Making:**

   The system provides valuable insights into potential trends and patterns in cryptocurrency prices, enabling users to make informed decisions when buying, selling, or holding digital assets.

2. **Risk Mitigation:**

   Predictive models, such as Random Forest and Gradient Boosting, aid in assessing and mitigating risks associated with cryptocurrency investments by offering forecasts that can guide risk management strategies.

3. **Diverse Model Comparison:**

   The inclusion of multiple models (K-Nearest Neighbors, Random Forest, Gradient Boosting) allows for comparative analysis, enabling users to select the most suitable model for their specific needs and market conditions.

4. **Interpretability with SHAP Values:**

   The integration of SHAP values enhances the transparency of the prediction models, providing clear insights into the factors influencing cryptocurrency prices. This interpretability is crucial for building trust in model predictions.

5. **Adaptability to Market Changes:**

   The system's iterative approach, including hyperparameter tuning and model optimization, ensures adaptability to evolving market dynamics. This allows the models to stay relevant and effective over time.

6. Visualization for Stakeholders:

   The visualization and reporting components of the system facilitate easy communication of results to stakeholders. Clear visualizations aid in conveying complex information about market trends and model performance.

3.5. Requirement Specification

3.5.1. Hardware Requirements:

Processor (CPU): A multi-core processor is recommended for faster computation, especially when training machine learning models.

RAM (Memory): A minimum of 8 GB RAM is recommended for handling large datasets and model training.

Storage: Sufficient disk space to store datasets, libraries, and model outputs.

3.5.2. Software Requirements:

Python

Python Libraries: Seaborn , yfinance, numpy, pandas, matplotlib, scikit-learn, plotly, shap

Jupyter Notebook (Optional)

Plotly (Optional)

Machine Learning Libraries

Data Source

Additional Libraries (if needed)

CHAPTER 4

IMPLEMENTATION and RESULT

The implementation of the Bitcoin price prediction model began with the comprehensive preparation and preprocessing of historical data encompassing adjacent close prices and trading volumes for Bitcoin and related cryptocurrencies. Key features were carefully selected, and additional engineering was performed to capture nuanced patterns. The chosen machine learning model, Random Forest Regressor, underwent hyperparameter tuning using Randomized Search CV to optimize its performance. The dataset was divided into training and testing sets to accurately evaluate the model's predictive capabilities. Normalization using Min Max Scaler ensured uniform treatment of input features. The model's interpretability was enhanced through SHAP analysis, shedding light on the significance of each feature in the predictions.

Upon model training and evaluation, the results indicated a high degree of accuracy, measured by metrics such as R-squared values. Visualizations were created to showcase the model's predictions against actual Bitcoin prices, providing a clear representation of its efficacy. The hyperparameter tuning further refined the model's performance, enhancing its adaptability to dynamic market conditions. The deployment readiness of the model was ensured, paving the way for its integration into real-time environments for ongoing Bitcoin price predictions. In summary, the implementation successfully combined feature engineering, machine learning, and interpretability analysis to develop a robust model, offering accurate insights into the dynamic realm of Bitcoin prices.

CHAPTER 5

CONCLUSION

In conclusion, our machine learning model for Bitcoin price prediction has demonstrated exceptional accuracy and interpretability. Through meticulous data preprocessing, feature engineering, and model optimization, we achieved a robust predictive tool using Random Forest Regressor and fine-tuned it with Randomized Search CV. The SHAP analysis provided transparency into the model's decision-making, enhancing our understanding of Bitcoin's intricate dynamics. Evaluation results, including high R-squared values, underscore the model's precision. Visualizations vividly depicted its predictions against actual prices, offering tangible insights. The refined model is deployment-ready, ensuring its seamless integration into real-time environments for continuous predictions. This project contributes significantly to advancing cryptocurrency analytics and provides a practical solution for navigating Bitcoin's complexities, marking a noteworthy stride in leveraging technology for informed decision-making in the dynamic cryptocurrency market.

5.1. FUTURE SCOPE

The project's scope extends to exploring alternative machine learning algorithms, analyzing additional features like sentiment indicators and macroeconomic metrics for enhanced prediction accuracy. Advanced time series analysis techniques will be employed to capture temporal dependencies in Bitcoin price data. Ensemble methods will be investigated to optimize model performance, and the project will be expanded to include cryptocurrency portfolio optimization strategies. Real-time predictions and dynamic model updating will be pursued, ensuring adaptability to evolving market conditions. A user-friendly interface or dashboard will be developed for intuitive interaction and visualization of predictions. External factors, including regulatory changes and geopolitical events, will be integrated into the model for improved contextual awareness. The scope further encompasses cross-cryptocurrency comparison, enabling a comprehensive understanding of market dynamics. This holistic approach ensures the project's relevance in the rapidly evolving cryptocurrency landscape, providing valuable insights and practical solutions for investors and stakeholders.

GITHUB LINK

https://github.com/rishika712

VIDEO LINK

https://youtu.be/P3UHDZnvjh8

REFERENCES

- **www.kaggle.com**
- **www.github.com**
- **Tom M. Mitchell, Machine Learning, McGraw-Hill Education, 2013**
- **Machine Learning using Python ,Manaranjan Pradhan, U Dinesh Kumar, Wiley 2019**
- **Machine Learning, Anuradha Srinivasaraghavan, VincyJoeph, Wiley 2019**

APPENDIX

The process begins with data retrieval and preprocessing, where historical price data for Bitcoin, Ethereum, Tether, and Binance Coin is obtained from Yahoo Finance and cleaned to focus on pertinent features. The subsequent exploratory data analysis (EDA) involves integrating cryptocurrency prices into a single data frame and visualizing trends through line plots and histograms. Feature engineering and selection follow, emphasizing adjacent close prices and trading volumes, with considerations for additional features like sentiment analysis. Model building includes the application of Random Forest Regressor, K-Nearest Neighbors, Decision Tree, and Gradient Boosting, with subsequent evaluation using R-squared scores. Hyperparameter tuning is performed through Randomized Search CV to optimize the Random Forest Regressor. Model interpretability is enhanced using SHAP analysis, shedding light on feature contributions. Visualization and reporting involve the generation of a Plotly candlestick chart and SHAP summary plots. Considerations for real-time deployment and a prototype user interface design are discussed. Advanced model optimization strategies, such as ensemble methods and external factors integration, are proposed for future exploration.