Bitcoin Price Prediction

A report on
Big Data Analytics Lab Project
[CSE-3263]

Submitted By
SAKSHAM ARORA – Reg. No. 210962202
SHARON PLAMKUDIYIL REJI – Reg. No. 210962087

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
MANIPAL INSTITUTE OF TECHNOLOGY,
MANIPAL ACADEMY OF HIGHER EDUCATION
APRIL 2024

Bitcoin Price Prediction

Saksham Arora
Department Of Computer Science and Engineering
Manipal Institute of Technology
Karnataka, India
saksham.study19@gmail.com

Sharon Plamkudiyil Reji
Department Of Computer Science and Engineering
Manipal Institute of Technology
Karnataka, India
sharonpreji@gmail.com

Abstract—In this project we aim to predict the closing price of Bitcoin on a particular day using PySpark and machine learning algorithms like Linear Regression, Decision Trees and Random Forest. Based on the predicted price a person can decide whether to buy Bitcoin or not. The performance of the models is evaluated using metrics like RMSE. This has real life applications in the field of financial analysis and can help investors and stakeholders in making better decisions.

Keywords— Bitcoin, PySpark, Machine Learning, Linear Regression, Decision Trees, Random Forest , RMSE, financial analysis

I. INTRODUCTION

Cryptocurrencies have become a disruptive force in the constantly changing financial landscape, capturing the interest of investors all around the world. These digital assets have become a main topic of discussion in the financial community since they provide characteristics that go beyond the limits of traditional banking. As a leader in this space, Bitcoin represents a fully virtual currency that has the ability to transform financial transactions. It functions as an online cash alternative. Notwithstanding their attraction, there is a cloud of doubt regarding the broad acceptance of cryptocurrencies, mostly because of their infamous volatility and unpredictability. The idea of decentralized currencies appeals to investors, but they are also concerned about price volatility. This creates a paradox where the very characteristics of cryptocurrencies that make them appealing also pose significant challenges. The quintessential decentralized digital currency, Bitcoin, runs on a public key infrastructure where users use public keys to conduct transactions. The terrain is not without its dangers, though. Only a small percentage of retailers have accepted Bitcoin, and other countries have explicitly prohibited its use. The bitcoin landscape is complicated, as seen by this contrast between promise and risk. We focus on the complex issue of price prediction among the many difficulties the bitcoin ecosystem faces. The intrinsic unpredictability and uncertainty ingrained in cryptocurrencies have made precise forecasts challenging, even with the widespread availability of forecasting tools

II. LITERATURE REVIEW

In [1], Aravindan Jg and Sankara Rama Krishnan V's paper "Parent Coin based Cryptocurrency Price Prediction using Regression Techniques" investigates the use of regression techniques in predicting the price of a cryptocurrency based on its parent coin. The research looks into the relationship between the value of a particular cryptocurrency and the performance of its parent coin. Regression analysis is used as a predictive tool to improve the accuracy of cryptocurrency price predictions.

In [2], The paper "Automated Bitcoin Trading via Machine Learning Algorithms" by Isaac Madan, Shaurya Saluja, and Aojia Zhao from Stanford University's Department of Computer Science investigates the use of machine learning algorithms in automating Bitcoin trading. The authors delve into the design and implementation of algorithms for analyzing market data and making trading decisions in the volatile Bitcoin market. The emphasis is on using machine learning techniques to improve trading

strategies and potentially achieve more effective and informed cryptocurrency trading outcomes.
In [3], The paper "Prediction of Bitcoin Prices with Machine Learning Methods Using Time Series Data" by Seçkin Karasu, Aytaç Altan, Zehra Saraç, and Rfat Hacolu from the Department of Electrical and Electronics Engineering at Bülent Ecevit University in Zonguldak, Turkey, has focused on using machine learning methods to forecast Bitcoin prices. The authors have looked into using time series data to train machine learning models to forecast future Bitcoin price movements. The research includes the development and evaluation of various machine learning techniques for their effectiveness in predicting Bitcoin prices.

In [4], G. L. Joshila et al.'s paper, presented at the 2021 ICOEI, introduces a Bitcoin price prediction model based on . The model, which ensures accurate predictions on specific dates, is intended for regular Bitcoin traders. Its simplicity allows for simple deployment in a variety of environments. In the conclusion, future work on developing an hourly prediction model using ensemble techniques is proposed, indicating potential improvements for finer grained forecasting.

In [5], The paper outlines a comprehensive framework for predictive analytics in stock market trading, integrating Moving Average (MA) method and Long Short Term Memory (LSTM) model. Through extensive data collection, preprocessing, and correlation analysis, it achieves a prediction accuracy of 95.82% with LSTM and identifies a 54% correlation between Monarch Staffing (MSTF) and Alphabet (GOOG) stocks using MA. The study underscores the significance of technical analysis and neural networks in stock trading algorithms, providing valuable insights for students, researchers, and industry practitioners.

In [6], The paper "Bitcoin Price Prediction Using Machine Learning" by S. Velankar, S. Valecha, and S. Maji, presented at the 2018 International Conference on Advanced Communication Technology, investigates the use of machine learning to forecast Bitcoin prices. In their study, the authors used Bayesian Regression and GLM/Random Forest methods. The emphasis is on determining the efficacy of these machine learning techniques in predicting Bitcoin price movements.

In[7], P. V. Rane and S. N. Dhage's paper "Systematic Erudition of Bitcoin Price Prediction Using Machine Learning Techniques' ' investigates machine learning approaches for predicting Bitcoin prices. The authors emphasize the Bitcoin system's uniqueness and its impact on volatility. The study offers a survey of techniques, with the NARX Model being the most accurate. While some existing systems achieve 60-70% accuracy, the paper recommends more research into advanced methods for a more comprehensive understanding and precise forecasting of Bitcoin prices. Overall, the study's goal is to help investors navigate the volatile cryptocurrency market.

In[8], Fernandes et al.'s paper, which was presented at the 2021 ICAC3 Conference, delves into Bitcoin price prediction. Recognizing the difficulties posed by market fluctuations, the study employs historical Bitcoin transaction data, employing LSTM, GRU, and RNN models. Sentiment analysis was abandoned due to data mapping issues. The system provides near-accurate results by relying on time series observations at 30-minute intervals. With its superior sentimental analysis accuracy, LSTM is effective in predicting Bitcoin prices, which is critical in the highly volatile cryptocurrency market. The authors also use Gradio to create a user-friendly interface and incorporate web scraping for relevant news updates, increasing the model's comprehensiveness.

In[9], M. Ali and S. Shatabda's paper, which they presented at the 2020 ICAICT Conference, focuses on a data selection methodology that is used to train a linear regression model that predicts Bitcoin prices. The research, carried out in Dhaka, Bangladesh, yields a noteworthy 96.97 percent accuracy rate for the linear regression model. The paper's methodology advances our knowledge of efficient data selection techniques that can be used to raise the precision of Bitcoin price prediction models.

In[10], The paper presented at the 2021 ICECA Conference by S. E. Freeda, T. C. E. Selvan, and I. G. Hemanandhini focuses on predicting Bitcoin prices using a deep learning model. The study, which was conducted in Coimbatore, India, investigates the use of deep learning techniques for accurate Bitcoin price forecasting.

III. RESEARCH GAPS AND OBJECTIVES

Research Gaps:

Data Volatility: The project recognizes that there is an inherent difficulty with cryptocurrency data volatility, since fluctuations in the price of Bitcoin can occur suddenly and without warning. This volatility needs to be carefully taken into account when developing models since it could affect how accurate the predictions are.

Abrupt Data Changes: The constraint stems from the peculiarities of cryptocurrency markets, wherein abrupt and unforeseen occurrences have the potential to substantially influence valuations. Because of this, it may be difficult to appropriately record and predict such sudden changes using historical data.

Deep Learning Exclusion: Neither deep learning nor neural networks are used in this project; instead, it makes use of conventional machine learning techniques. Although the project acknowledges the potential superiority of these sophisticated techniques, it may not fully utilize their potential for sophisticated pattern recognition and prediction.

Limited Historical Data: The quantity and caliber of historical data may limit how accurate predictions can be made. Inadequate historical data may make it more difficult for the model to correctly identify long-term trends and patterns.

Research Objectives :

Improving Financial Decision-Making: The main goal is to give financial analysts and investors a useful tool to help them decide whether to buy or sell Bitcoin. The project intends to advance the field of financial studies and enable better-informed investment strategies by utilizing machine learning algorithms.

Pyspark and Machine Learning: The goal of the project is to evaluate how well different machine learning algorithms predict Bitcoin prices using PySpark. This project covers Linear Regression, Decision Tree and Random Forest algorithms.

Understanding Long-Term Trends: The project seeks to identify and comprehend long-term trends in Bitcoin prices through the analysis of historical data. With this knowledge, investors can create strategies that take into consideration the past performance and behavior of cryptocurrencies.

Practical Applicability: The project's goal is to develop a model that is applicable in real-world situations and both accurate and practical. This involves taking into account factors like usability, environment adaptability, and accessibility for regular people who trade bitcoins.

Future Model Improvements: The study aims to establish the foundation for upcoming developments, such as the investigation of neural networks and deep learning models. This goal is motivated by the realization that improvements in these methods could further improve the precision of Bitcoin price forecasts.

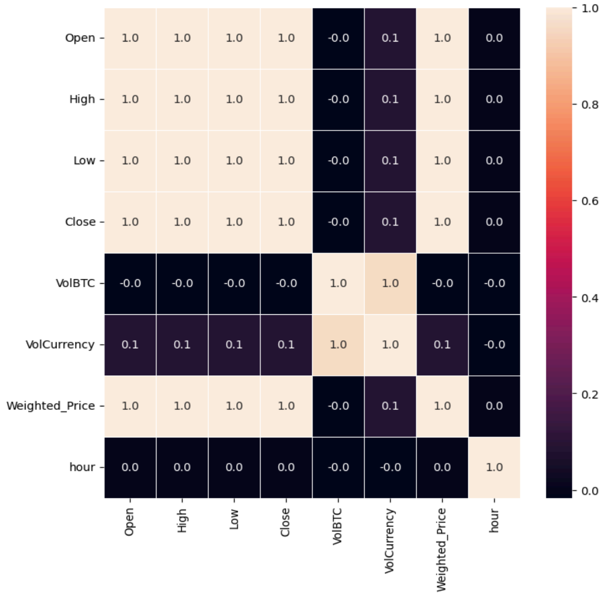
IV. METHODOLOGY

4.1 Data Collection

4.1.1 Data Source: Historical Bitcoin price information was gathered for this project from dependable sources including financial databases and cryptocurrency exchanges.
For the data, we particularly used
https://www.kaggle.com/code/sandeepanmukherjee/pyspark-bitcoin-trend-analysis
which provides daily bitcoin prices data.
The dataset captures variations in the bitcoin price at different intervals of the day for the course of 2 years.

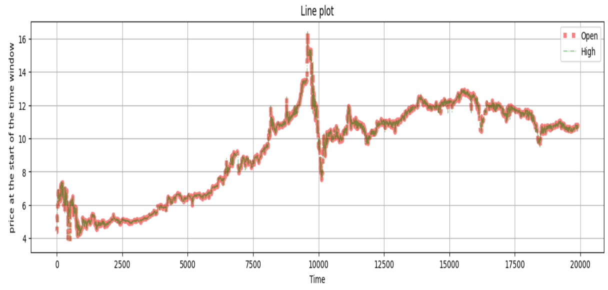
4.1.2 Data Preprocessing: Preprocessing data is done to make sure the dataset is consistent and of high quality. This included dealing with NULL values, eliminating outliers, renaming the required columns, preprocessing the dataset including manipulating the required columns like extracting data, time and hour from date_time column and fixing any problems with the quality of the data that would have impacted how well the machine learning models performed.

4.1.3 Data Visualization: Data visualization is done to visualize and analyze the given data using correlation matrix.

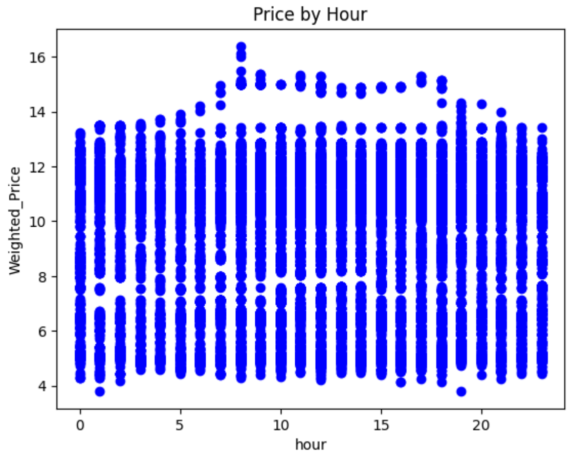


The correlation matrix computes correlation coefficients among numerical features in a dataset, unveiling the magnitude and orientation of linear associations.

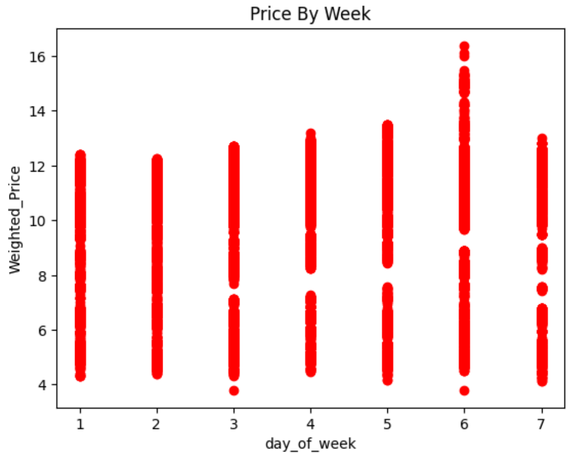
There is a strong negative correlation between High and Open. This means that when High is high, Open tends to be low.



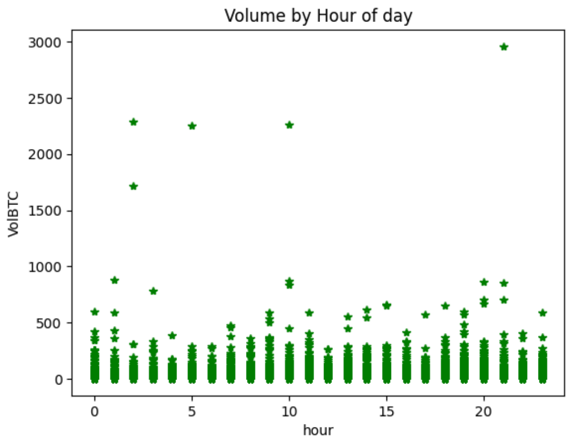
On analyzing the line plot, we observe not much difference in the open and high value which means the peak price during the day is close to the open price during the day.



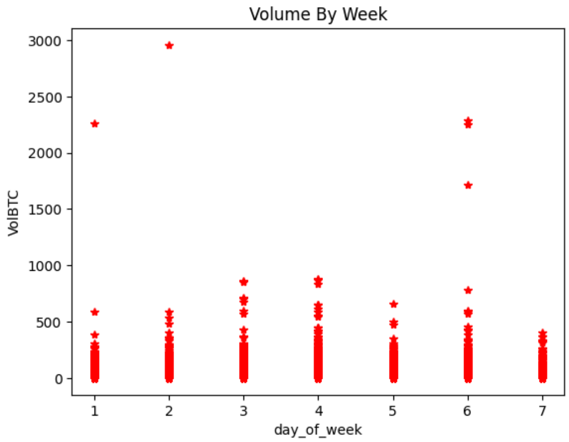
Scatter plot is done to visualize the relationship between hour of the day and the Closing price(Weighted_price) of Bitcoin.The prices tend to be slightly higher in the later hours of the day. The two variables have a weak positive correlation between each other.



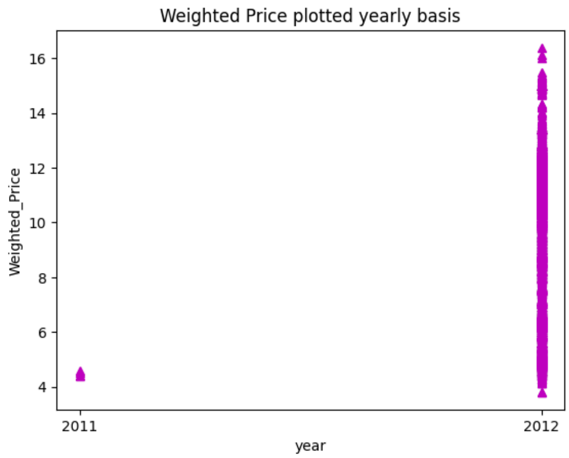
Scatter plot is done to visualize the relationship between the day of week and the Closing of the bitcoin.The prices are mostly consistent throughout the week with a little rise on the 6th day of the week.



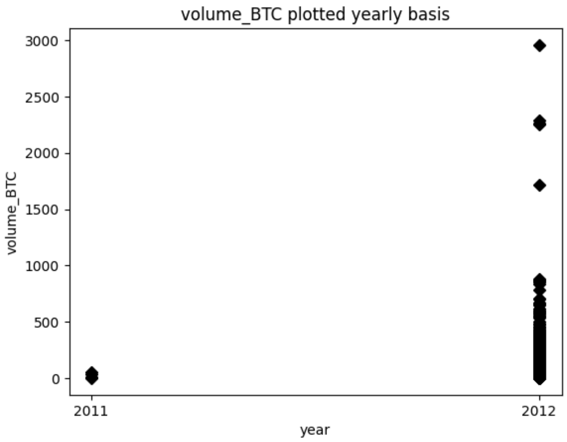
Scatter plot is done to visualize the relationship between the hour of the day and the volume of Bitcoin per hour.The data is consistent with few outliers.



Scatter plot is done to visualize the relationship between the day of week and the volume of bitcoin per week.The data is consistent with a few outliers.



Scatter plot is done to visualize the relationship between the year and the Weighted price of the bitcoin.The price of the Bitcoin was very low in the year 2011 but started rising in 2012 which means Bitcoin gained popularity in 2012.



Scatter plot is done to visualize the relationship between the year and the volume of bitcoin per year. The volume in the year 2011 is lower compared to 2012.

4.2 Feature Selection and Scaling

4.2.1 Identification of Features: Selecting relevant features to forecast the price of Bitcoin. Possible attributes consist of: Historical trading volume and prices (open, close, high, and low)

4.2.2 Feature Vector Assembler: Feature vector helps us to encode and convert the features into numbers in order to perform machine learning tasks.

The transform() method takes the input DataFrame and generates a new DataFrame assembled_df by assembling the specified input columns into the "features" column using the VectorAssembler.
4.2.3 Feature Scaling: It is done in order to squish the values between 0 and 1. It is needed in machine learning to ensure that features are on similar scales, which helps algorithms perform optimally and prevents bias in the model.
The fit() method computes summary statistics of the feature vectors in the DataFrame to determine how to scale them. The transform() method scales the feature vectors in the DataFrame using the computed min-max values.

4.3 Baseline Model: Linear Regression

4.3.1 Model Implementation: Created a linear regression model with the chosen characteristics. The model functioned as a reference point for contrasting with more complex machine learning models.

4.3.2 Training and Validation: Divide the dataset into sets for validation and training. Utilizing a different validation set, test the linear regression model's performance after training it on the training set.

4.3.3 Evaluation Metrics: Metrics like , Root Mean Squared Error (RMSE), was used to assess the model's performance. Its measures offered a starting point for evaluating the performance of more sophisticated models.

4.4 Decision Tree

4.4.1 Model Implementation: Developed a decision tree model to forecast the price of bitcoin. modified the model's hyperparameters, such as the minimum sample split and tree depth, to maximize performance.

4.4.2 Training and Validation: Utilizing a different validation set, assess the decision tree model's performance after training it on the training set.

4.4.3 Evaluation Metrics: Metrics like , Root Mean Squared Error, was used to assess the model's performance. Utilizing the established evaluation metrics, evaluate the decision tree model's performance.

4.5 Random Forest

4.5.1 Model Implementation: Combine many decision trees to create a random forest model. To maximize the performance of the model, adjust hyperparameters like the number of trees and the maximum number of features per tree.

4.5.2 Training and Validation: Utilizing a different validation set, assess the random forest model's performance after training it on the training set.

4.5.3 Evaluation Metrics: Used the defined metrics to assess the random forest model's overall performance. Compared the outcomes to the baseline decision tree, and linear regression models.

5. Model Evaluation: The purpose of this report is to give the users a thorough understanding of the capabilities and possible uses of the models.
In addition to addressing potential issues and limitations, this thorough methodology guarantees a methodical approach to investigating the predictive powers of Decision Trees, Random Forest, and Linear Regression models in the context of stock price prediction.
The report's following sections will provide in-depth analysis, insights, and conclusions from applying these models to actual financial data.

V. DISCUSSION AND ANALYSIS OF RESULTS

The outcomes of our models for predicting the price of Bitcoin using linear regression, decision trees, and random forests are shown in this section.

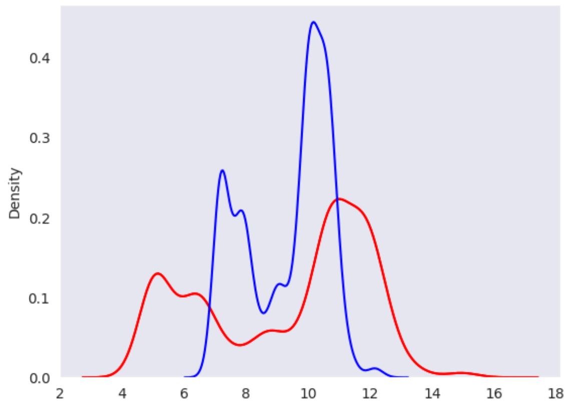
6.1 Linear Regression Model

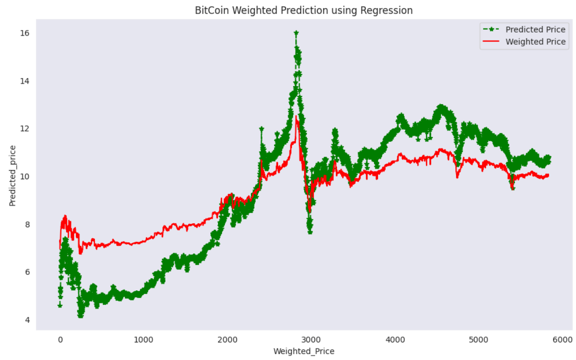
6.1.1 Model Performance

The baseline against which we can compare is the linear regression model. The evaluation metrics are listed below:
RMSE: 1.33583

6.1.2 Analysis

The Root Mean Squared Error (RMSE) of around 1.33498 indicates the average discrepancy between observed and predicted Bitcoin prices within the test dataset. A lower RMSE value suggests better model accuracy.
In the context of Bitcoin prediction, this RMSE implies predictions of reasonable accuracy.
Further evaluation against baseline benchmarks and potential model refinements may enhance prediction performance.





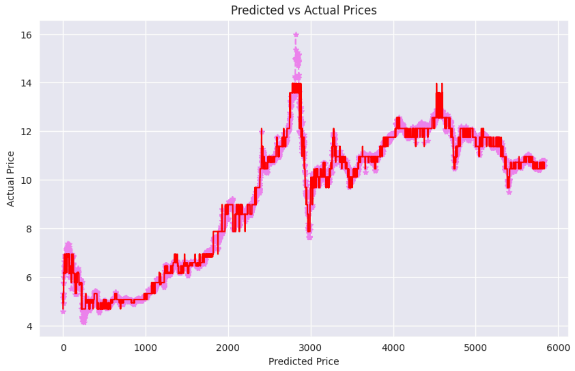
The graphs depict that Regression proved to be a good technique in predicting the Bitcoin prices. It predicted the prices accurately for most cases and is able to identify the general trend although there are a few cases in which the predictions are not very close to the actual prices. This happens because the Bitcoin market is a very volatile market and price fluctuation is unpredictable. Therefore, further we implement better algorithms like Decision Tree and Random Forest.

6.2 Decision Tree Model

6.2.1 Model Performance

Decision Tree Model was trained and evaluated using the following metrics:
RMSE: 0.174082

6.2.2 Analysis



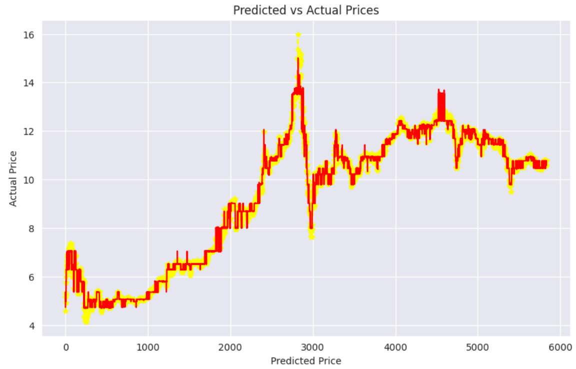
Decision Tree Regressor detected the trend accurately and was able to generate better results than Regression with lesser error. The predicted values are almost accurate as seen in the graph of Predicted price v/s Actual price.

6.3 Random Forest Model

6.3.1 Model Performance

Random Forest Model was trained and evaluated using the following metrics:
RMSE: 0.160666

6.3.2 Analysis



Random Forest is an ensemble learning method that builds multiple decision trees and aggregates their predictions to improve the results compared to one single decision tree. Random forest stood out to be the best algorithm to predict Bitcoin prices. It outperformed the Regression and Decision Tree models and was able to predict the Bitcoin prices accurately.

VI. CONCLUSIONS AND FUTURE WORK

In summary, the goal of our research was to create and assess machine learning models using the random forest, decision tree and linear regression algorithms for predicting Bitcoin prices.
A thorough examination of the data allows us to make a number of important deductions and discoveries.

8.1 Recap of Results

Model of Linear Regression: It functioned as our foundational model, offering a simple linear connection between input features and Bitcoin prices.

Although it provided a fair starting point, its simplicity might have made it less capable of capturing the intricate non-linear patterns that are present in the fluctuations of cryptocurrency prices.

Decision Tree Model: Although decision trees are renowned for being easily interpreted, it's possible that the standalone decision tree had trouble either overfitting or underfitting the training set, which would have limited its ability to generalize to new data.

The decision tree model's interpretability is valuable, but the intricacy of the dynamics affecting Bitcoin prices may have had an effect on its performance.

Random Forest Model: The Random Forest model demonstrates strong potential for Bitcoin price prediction, evidenced by its low RMSE of 0.160666. While effective in capturing complex patterns and relationships in Bitcoin data, vigilance is required to avoid overfitting.

Nevertheless, its adaptability and ensemble approach position it as a valuable tool for forecasting Bitcoin prices, subject to ongoing refinement and optimization for real-world application.

8.3 Future Work

The application of machine learning using PySpark in cryptocurrency price prediction is significantly impacted by the study's findings.
In order to further improve predictive performance, future research directions might examine more sophisticated machine learning techniques, add new features, and take ensemble methods into consideration. Additionally, in the constantly shifting landscape of cryptocurrency prices, models' efficacy will depend on constant monitoring and adaptation to changing market conditions.

VII. ACKNOWLEDGEMENT

We would like to take this opportunity to thank our professors – Professor, Dr. Prakash Kalingarao Aithal and Professor Ms. Suma D for their invaluable guidance and mentorship throughout the duration of the project.Their endless support and expertise about the subject helped us in our growth in the subject and the project’s success.

VIII. REFERENCES

[1] J. Aravindan and R. K. V. Sankara, "Parent Coin based Cryptocurrency Price Prediction using Regression Techniques," 2022 IEEE Region 10 Symposium (TENSYMP), Mumbai, India, 2022, pp. 1-6, doi: 10.1109/TENSYMP54529.2022.9864452.

[2] Madan, S. Saluja, and A. Zhao, Dept. Comput. Sci., Stanford Univ., Stanford, CA, USA, Tech. Rep., 2015, “Automated Bitcoin Trading via Machine Learning Algorithms”

[3] S. Karasu, A. Altan, Z. Saraç and R. Hacioğlu, "Prediction of Bitcoin prices with machine learning methods using time series data," 2018 26th Signal Processing and Communications Applications Conference (SIU), Izmir, Turkey, 2018, pp. 1-4, doi: 10.1109/SIU.2018.8404760.

[4] G. L. Joshila, A. P, D. U. Nandini and G. Kalaiarasi, "Price Prediction of Bitcoin," 2021 5th International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India, 2021, pp. 113-116, doi: 10.1109/ICOEI51242.2021.9452976.

[5] K. N. Myint and Y. Y. Hlaing, "Predictive Analytics System for Stock Data: methodology, data pre-processing and case studies," 2023 IEEE Conference on Computer Applications (ICCA), Yangon, Myanmar, 2023, pp. 77-82, doi: 10.1109/ICCA51723.2023.10182047.

[6] S. Velankar, S. Valecha and S. Maji, "Bitcoin price prediction using machine learning," 2018 20th International Conference on Advanced Communication Technology (ICACT), Chuncheon, Korea (South), 2018, pp. 144-147, doi: 10.23919/ICACT.2018.8323676.

[7] P. V. Rane and S. N. Dhage, "Systematic Erudition of Bitcoin Price Prediction using Machine Learning Techniques," 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS), Coimbatore, India, 2019, pp. 594-598, doi: 10.1109/ICACCS.2019.8728424.

[8] M. Fernandes, S. Khanna, L. Monteiro, A. Thomas and G. Tripathi, "Bitcoin Price Prediction," 2021 International Conference on Advances in Computing, Communication, and Control (ICAC3), Mumbai, India, 2021, pp. 1-4, doi: 10.1109/ICAC353642.2021.9697202.

[9] M. Ali and S. Shatabda, "A Data Selection Methodology to Train Linear Regression Model to Predict Bitcoin Price," 2020 2nd International Conference on Advanced Information and Communication Technology (ICAICT), Dhaka, Bangladesh, 2020, pp. 330-335, doi: 10.1109/ICAICT51780.2020.9333525.

[10]S. E. Freeda, T. C. E. Selvan and I. G. Hemanandhini, "Prediction of Bitcoin Price using Deep Learning Model," 2021 5th International Conference on Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, India, 2021, pp. 1702-1706, doi: 10.1109/ICECA52323.2021.9676048.