Erasmus University Rotterdam

Erasmus School of Economics

Master Thesis Data Science and Marketing Analytics

Predicting Solana Price Trends Using Various Market Indicators and Blockchain Activity

Nikilesh Jagan (506809)

Supervisor:	Alfons Andreas
Second assessor:	Name of your second assessor
Date final version:	30th July 2024

The content of this thesis is the sole responsibility of the author and does not reflect the view of the supervisor, second assessor, Erasmus School of Economics or Erasmus University.

Contents

Abstract

This study aimed to predict Solana's price trends by integrating traditional stock market indicators, inflation metrics, and blockchain-specific features using the XGBoost model. A comprehensive analysis was conducted, examining cumulative SHAP values and individual SHAP values for the lags of each feature to provide a nuanced understanding of their impacts. The findings reveal that traditional stock market indicators and blockchain-specific features exhibit a mixed effect on Solana's price trends depending on the specific lag considered. However, the inflation metric consistently shows a clear downward pressure on Solana's price trends.

Introduction

2.1. Overview

The cryptocurrency market has experienced explosive growth in recent years, becoming a significant component of the global financial system. Cryptocurrencies, which are decentralized digital assets based on blockchain technology, have revolutionized financial transactions by offering increased security, transparency, and efficiency. Among the myriad of digital assets, Solana has attracted substantial interest from both retail and institutional investors due to its high throughput (table 1) and low transaction costs (Yakovenko, 2018). Although Ethereum leads the altcoin market, Solana is growing rapidly, enjoying $9.6 million in inflows.[1] In addition, Shopify's recent partnership with Solana to explore new use cases in commerce underscores Solana's potential and its growing adoption across industries.[2]

Table 1: Throughput Comparison of Blockchain Platforms

Blockchain Platform	**Transactions per Second (TPS)**	**Source**
Solana	65,000	(Yakovenko, 2018)
Bitcoin	7	(Nakamoto, 2008)
Ethereum	30	(Buterin, 2013)

The existing body of literature on cryptocurrency price prediction has predominantly centered on Bitcoin and Ethereum, with assets like Solana receiving considerably less attention. Furthermore, while previous research has utilized various independent variables, such as sentiment analysis from social media, internal blockchain activity metrics, and historical cryptocurrency prices (Gurrib, 2022; Olivier Kraaijeveld, 2020), there has been a lack of exploration into integrating traditional financial market indicators and macroeconomic variables with cryptocurrency-specific data. This study seeks to address this gap by incorporating stock market indices, such as the S&P 500, macroeconomic indicators like the CPILFESL[3] and cryptocurrency prices and volumes traded from Bitcoin and Ethereum to predict Solana price trends. This multi-faceted approach will provide an understanding of the links between various financial indicators.

[1] Ethereum Overtakes Solana With Most Altcoin Inflows Year-to-Date As Positive Sentiment Continues

[2] Solana Pay Integrates with Shopify as New Payment Option to Transform Commerce

[3] Consumer Price Index for All Urban Consumers: All Items Less Food and Energy in U.S. City Average

Solana operates in a highly volatile market environment, where prices can fluctuate rapidly due to several factors. This inherent volatility in the cryptocurrency market presents significant challenges for price prediction, necessitating the development of robust predictive models capable of handling such complexity. Traditional financial prediction models, such as Holt-Winters exponential smoothing, which rely on linear assumptions and require data to be divisible into trend, seasonal, and noise components, have proven less effective in capturing the non-linear dependencies characteristic of cryptocurrency markets (McNally, 2018). In contrast, machine learning techniques offer superior performance by effectively modeling these non-linear relationships. This study employs an XGBoost model for its predictive analysis. Additionally, Shapley values are used to determine the effect of all the features on Solana price trends.

The structure of this study will begin with detailing the research question and hypotheses that are being addressed. Following this, a comprehensive literature review will be presented, covering existing studies on cryptocurrency price prediction, the XGBoost model and Shapley values. The subsequent section will focus on data, describing the various data sources and presenting descriptive statistics. The methodology section will then cover all data pre-processing steps, model development, evaluation, result interpretation, and the limitations of the methodology. Finally, the study will finalise with the results section answering all the hypotheses, followed by a conclusion.

2.2. Research Question & Hypotheses

The central research question that this study aims to address is:

What is the impact of the U.S. stock market, U.S. inflation, and blockchain market metrics on Solana's price trends?

The research question can be investigated through the following hypotheses:

H1: An increase in the S&P 500 index leads to an upward price trend for Solana

The S&P 500 index serves as a widely recognized barometer of the overall health of the stock market and the broader economy (Hashemi et al., 2017; Mustapa & Ismail, 2019). When the S&P 500 experiences an upward trend, it generally reflects positive investor sentiment and favorable economic conditions. This optimistic outlook can extend beyond traditional equities, influencing investor behavior in other asset classes, including cryptocurrencies. Consequently, increased investment flows into the cryptocurrency market can elevate the demand and valuation of Solana, driving up its price. This phenomenon highlights the interconnectedness of financial markets and underscores how positive developments in the stock market can bolster confidence and investment in the cryptocurrency sector.

H2: An increase in transaction volumes for Solana, Bitcoin and Ethereum leads to an upward price trend for Solana

When metrics such as transaction volume increase, they reflect a growing adoption and usage of the Solana, Bitcoin, and Ethereum blockchains. As more users and applications find value in these blockchain capabilities, the demand for the respective cryptocurrencies rises. This increased activity can enhance the perceived value and utility of Solana, driving up investor interest and further elevating its market price. This relationship underscores the importance of network activity as a predictive indicator of market performance for these cryptocurrencies, highlighting how higher transaction volumes can signal positive market trends and investor confidence.

H3: Upward price movements of Bitcoin, Ethereum and Solana result in upward trends in Solana's market price

Bitcoin and Ethereum, as the two largest cryptocurrencies by market capitalization, often serve as bellwethers for the broader cryptocurrency market. When Bitcoin and Ethereum experience positive price movements, they tend to generate favorable market sentiment and increased investor confidence across the entire cryptocurrency ecosystem. This positive sentiment can lead investors to explore and invest in other cryptocurrencies, including Solana. This interconnected market behavior underscores how major cryptocurrencies can influence and propel the market trends of other digital assets. Furthermore, rising Solana prices are likely to indicate an upward trend for Solana. It is a common and effective practice in price prediction studies to use historical data to forecast future trends.

H4: An increase in the CPI of the U.S. results in a decrease in Solana's price trend

The CPI can be used as a measure of inflation in the U.S economy (Bryan & Cecchetti, 1993; Zellner et al., 1980). When the CPI rises, it signals increasing inflation, which often prompts the Federal Reserve to implement higher interest rates and tighter monetary policies to curb inflationary pressures. These measures can negatively impact investor sentiment, as higher interest rates typically make borrowing more expensive and saving more attractive, thereby reducing the liquidity available for investment in riskier assets like cryptocurrencies. As a result, investors might shift their capital towards more stable, lower-risk investments, leading to decreased demand for cryptocurrencies such as Solana. This hypothesis highlights the sensitivity of cryptocurrency markets to macroeconomic indicators and their potential inverse relationship.

Literature Review

3.1. Cryptocurrency Price Prediction

Prediction efforts for established financial markets, such as the stock market, are well-documented. Similarly, recent research has increasingly focused on predicting cryptocurrency prices, particularly Bitcoin and Ethereum, by leveraging various machine learning techniques and performing sentiment analysis. Below, we delve into several such studies, showcasing the diverse approaches and findings within this field.

McNally, 2018 critiques traditional time series models like Holt-Winters exponential smoothing, which rely on linear assumptions and require data divisible into trend, seasonal, and noise components. Due to the high volatility of the cryptocurrency market, these methods prove less effective. In his study, McNally focused on predicting Bitcoin price trends, using Bitcoin's closing price on Coindesk as the dependent variable, with opening price, daily high, daily low, mining difficulty, and hash rate as independent variables. He found that the LSTM model achieved an accuracy of 52%.

Wu et al., 2022 explores the challenges associated with predicting the highly volatile cryptocurrency market, noting that it is more difficult to forecast compared to traditional financial products like stocks due to the susceptibility of cryptocurrency prices to various economic, political, and other factors. This paper leverages the XGBoost algorithm to predict the short-term returns of 14 different cryptocurrency markets. The authors conducted experiments using data from the KAGGLE competition platform and enhanced the dataset through feature engineering. The findings demonstrate that the XGBoost model significantly outperformed traditional machine learning algorithms, showing 12.5%, 16.6%, and 43.3% higher prediction performance than Gradient Boosting, SVM, and Linear Regression models, respectively. This highlights the effectiveness of advanced machine learning techniques, particularly XGBoost, in improving the accuracy of cryptocurrency market predictions.

Srivastava et al., 2023 present an advanced model for predicting cryptocurrency prices, addressing the high volatility inherent in digital currencies like Bitcoin, Dogecoin, and Ethereum. The study integrates a regression algorithm and Particle Swarm Optimization (PSO) with XGBoost algorithm to enhance prediction accuracy. The approach uses time series data consisting of daily cryptocurrency prices. Comparative assessments indicate that the proposed model outperforms traditional methods, showing lower values for Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Mean Squared Error (MSE). The results underscore the XGBoost model's superior efficiency and accuracy in predicting cryptocurrency closing prices.

In their study, Z. Chen et al., 2020 examined the prediction of Bitcoin prices using various machine learning techniques, highlighting the importance of sample dimension engineering. They differentiated Bitcoin price prediction into daily and high-frequency (5-minute interval) categories. The study utilized high-dimension features such as property, network, trading, market attention, and gold spot price for daily predictions, achieving an accuracy of 66% with statistical methods like Logistic Regression and Linear Discriminant Analysis. Notably, machine learning models such as Random Forest, XGBoost, Quadratic Discriminant Analysis, Support Vector Machine, and Long Short-Term Memory outperformed statistical methods for 5-minute interval predictions, achieving an accuracy of 67.2%.

Kim H., 2021 delve into the relationship between inherent Ethereum Blockchain information and Ethereum prices. Moreover, they examine how Blockchain information from other publicly available cryptocurrencies is associated with Ethereum prices. The key findings indicate that macroeconomic factors, Ethereum-specific Blockchain data, and the Blockchain information of other cryptocurrencies play critical roles in predicting Ethereum prices. This research underscores the importance of considering a broad array of factors, including inter-cryptocurrency Blockchain dynamics, for accurate cryptocurrency price predictions.

Gurrib, 2022 investigates Bitcoin price prediction using sentiment analysis. The authors trained a Latent Dirichlet Allocation (LDA)-based classifier that utilized current BTC price data and news headlines to predict the next day's BTC price movement. Their results were compared with a Support Vector Machine (SVM) model and a random guessing approach. The SVM model outperformed the LDA classifier in predicting next-day BTC price trends. All models more accurately forecasted increases rather than decreases in BTC prices. Incorporating news sentiment data significantly improved the forecast accuracy, yielding a 0.585 accuracy score on the test data, outperforming random guessing. The LDA (with asset-specific features) and SVM (with both news sentiment and asset-specific features) models ranked highest within their classifier categories, indicating that both BTC news sentiment and asset-specific features are key factors in predicting next-day price direction.

Wołk, 2020 believed that Bitcoin, being one of the largest cryptocurrencies in terms of market capitalization, often sees its price influenced by public sentiment rather than institutional regulations. Wołk, 2020 explores this phenomenon by proposing that sentiment analysis can be effectively utilized as a computational tool to predict cryptocurrency prices. The study leverages data from Twitter and Google Trends to forecast short-term prices for major cryptocurrencies, demonstrating that psychological and behavioral attitudes of individuals have a substantial impact on the speculative nature of cryptocurrency prices.

Abraham et al., 2018 explored the predictive power of social media and web search data on the price changes of Bitcoin and Ethereum, the two largest cryptocurrencies by market capitalization. Their study utilized Twitter data and Google Trends to develop a model capable of predicting the direction of cryptocurrency price changes. Contrary to previous research that focused on sentiment analysis, their findings revealed that tweet volume, rather than sentiment, was a more reliable predictor of price direction. This insight is attributed to the inherently positive bias in cryptocurrency-related tweets. By incorporating tweet volumes and Google Trends data into a linear regression model with lagged variables, they achieved accurate predictions of future price movements. This approach underscores the importance of overall interest metrics, such as search volume and tweet frequency, in forecasting cryptocurrency prices.

3.2. XGBoost and Shapley Additive Values

Tree boosting is a highly effective and widely used machine learning method, extensively applied in diverse fields, including spam detection, advertising systems, fraud detection, and anomaly event detection in experimental physics. It is particularly valued for its ability to capture complex data dependencies and for its scalability in learning from large datasets. Among the various tree boosting methods, gradient tree boosting stands out for its application in many machine learning challenges, often achieving state-of-the-art results in classification benchmarks and ranking problems.

T. Chen and Guestrin, 2016 introduced and invented XGBoost, a scalable end-to-end tree boosting system that has gained widespread recognition in the machine learning community. The system has been highly effective, winning numerous machine learning competitions and consistently delivering top-tier results. XGBoost incorporates several innovative features, including a sparsity-aware algorithm for handling sparse data and a theoretically justified weighted quantile sketch for approximate tree learning. These innovations enable XGBoost to scale efficiently, handling billions of examples while using fewer resources than other systems. The system's scalability is further enhanced by its ability to exploit parallel and distributed computing, as well as out-of-core computation, allowing data scientists to process hundreds of millions of examples on a desktop. These capabilities make XGBoost an ideal choice for handling large datasets and complex machine learning tasks.

The empirical analysis conducted by Bentéjac et al., 2021 provides an in-depth evaluation of XGBoost, particularly focusing on its performance in terms of training speed, accuracy, and parameter tuning compared to gradient boosting and random forest. The study highlights that while gradient boosting showed the highest accuracy across various classification tasks, the differences between XGBoost and gradient boosting were not statistically significant in terms of average ranks. The research emphasizes the importance of meticulous parameter tuning for achieving optimal results with XGBoost, a necessity not as critical for random forests, which performed well with default settings.

One of the key findings is that parameter tuning for XGBoost, particularly the subsampling rate and the number of features selected at each split, significantly improves its performance. The study found that fixing the subsampling rate to 0.75 and the number of features to the square root of the total features reduced the parameter grid search size and enhanced the average performance of XGBoost. Additionally, the tuning phase accounted for the majority of the computational effort in training gradient boosting or XGBoost models.

The Shapley value, a concept from cooperative game theory, has gained significant traction in machine learning over recent years, demonstrating its utility across a variety of applications. In their comprehensive review, Rozemberczki et al., 2022 discuss fundamental concepts of cooperative game theory and elucidate the axiomatic properties of the Shapley value, such as fairness, symmetry, and efficiency. The Shapley value has been employed in diverse areas within machine learning, including feature selection, explainability, multi-agent reinforcement learning, ensemble pruning, and data valuation. This approach offers a theoretically motivated solution to measuring importance and attributing gains, central problems in many machine learning tasks. For instance, it provides a rigorous method to evaluate the contribution of individual features, data points, or models within an ensemble, facilitating more transparent and interpretable machine learning models.

3.3. Gaps in the Current Literature

Despite advancements in cryptocurrency price prediction using historical price, volume data, and sentiment analysis from social media and news sources, critical gaps remain. One significant gap is the insufficient incorporation of traditional financial market indicators and macroeconomic variables. Limited research investigates the relationship between cryptocurrency prices and broader market indices like the S&P 500 and macroeconomic indicators such as the CPILFESL in the U.S. Addressing these gaps could enhance the accuracy of predictive models.

Another significant gap is the focus on improving predictive models rather than understanding the features impacting cryptocurrency prices. Many studies aim to enhance predictive accuracy using multiple machine learning models. However, there is a lack of studies identifying and analyzing the importance of different features. This study addresses this gap by using Shapley Additive Explanations (SHAP) values to determine each feature's impact on cryptocurrency price prediction.

There is also a notable lack of academic focus on Solana compared to Bitcoin and Ethereum, despite Solana being a significant player in the blockchain space. This study aims to fill this gap by focusing on Solana, analyzing the factors influencing its price trends, and expanding academic research to include this important cryptocurrency. By addressing these gaps, this research will contribute to a more holistic understanding of cryptocurrency markets and enhance the predictive capabilities of financial models.

Data

4.1. Data Sources

4.1.1. COMPUSTAT

COMPUSTAT, part of S&P Global Market Intelligence, offers standardized North American and global financial statement and market data for over 80,000 active and inactive publicly traded companies, a resource that financial professionals have trusted for more than 50 years. The COMPUSTAT database used in this study has been accessed through Wharton Research Data Services (WRDS). WRDS is renowned for providing reliable access to a comprehensive range of business data, including financial, economic, and marketing information.

A component of COMPUSTAT is Compustat Daily Updates - Index Daily Prices, which provides detailed information on index prices. For this study, the closing prices of the S&P 500 index from April 2020 to June 2024 were sourced from here. Data source can be accessed here - S&P 500 Data (WRDS). The closing price, represented by the mnemonic PRCCD, reflects the last trade price with volume for the day for the security. This means that the closing price is the price at which the last transaction occurred on that trading day, ensuring that there was actual trading activity at that price.

4.1.2. FRED (Federal Reserve Economic Data)

Federal Reserve Economic Data (FRED) is an extensive online database created and maintained by the Research Department at the Federal Reserve Bank of St. Louis since the early 1990s. FRED provides access to a vast array of economic data time series, encompassing hundreds of thousands of data points from a multitude of national, international, public, and private sources.

In this study, FRED was utilized to obtain the Consumer Price Index for All Urban Consumers: All Items Less Food & Energy (CPILFESL) data from April 2020 to June 2024. Data source can be accessed here - CPILFESL Data (FRED). The CPILFESL is an aggregate measure of the prices paid by urban consumers for a typical basket of goods and services, excluding the highly volatile categories of food and energy. Commonly referred to as the "Core CPI," this index is widely utilized by economists and policymakers to assess underlying inflation trends without the noise caused by the frequent price fluctuations in food and energy.

4.1.3. Yahoo Finance

Yahoo Finance is a comprehensive financial news and data platform operated by Yahoo. It provides real-time and historical data on stock prices, indices, commodities, and cryptocurrencies, making it a valuable resource for investors, researchers, and financial analysts.

For this analysis, closing prices and trading volumes for Solana, Bitcoin, and Ethereum from April 2020 to June 2024 were collected. Data source can be accessed here - Cryptocurrency Data (Yahoo Finance).

4.2. Descriptive Statistics

Below tables showcase the descriptive statistics for all variables in this study. This includes metrics such as the minimum, 1st quartile, median, mean, 3rd quartile, maximum value, and standard deviation. These are crucial for understanding the distribution, central tendency, and variability of the dataset. Additionally, a correlation matrix will be presented to illustrate the relationships between the variables.[4]

Table 2: Descriptive Statistics for Volumes (in Hundreds of Millions)

	Solana	Bitcoin	Ethereum
Min.	0.007	53.31	20.82
1st Qu.	1.98	196.29	88.75
Median	6.93	282.24	141.56
Mean	13.11	318.66	161.48
3rd Qu.	19.24	389.32	202.58
Max.	170.69	3509.68	844.83
Std.Dev.	16.74	188.77	103.19

Solana's trading volume ranges from 700,000 to 17.07 billion, with a mean of 1.31 billion, indicating significant variability in trading activity. This high variability suggests that Solana experiences periods of intense trading activity, likely driven by market events or investor sentiment. Bitcoin's trading volume exhibits an even broader range, from 5.33 billion to 350.97 billion, with a mean of 31.87 billion. The extensive range and high mean volume highlight Bitcoin's status as the most actively traded cryptocurrency, often serving as a market leader and influencer. Ethereum's trading volume ranges from 2.08 billion to 84.48 billion, with a mean of

[4] The data pre-processing steps conducted prior to calculating the descriptive statistics are detailed in the Methodology section (5.2). Note that the presented descriptive statistics were calculated before applying Z-scale normalization.

16.15 billion, reflecting somewhat lower trading activity compared to Bitcoin but still substantial. This indicates that Ethereum maintains a strong presence in the market, likely due to its widespread use in decentralized applications and smart contracts. To visually observe the flow of data, time series plots of the volumes are included in the Appendix (see figures 8 to 10).

Table 3: Descriptive Statistics for Closing Prices (USD) and CPILFESL

	Solana	Bitcoin	Ethereum	SP&P 500	CPILFESL
Min.	0.52	6642.11	153.29	2736.56	265.46
1st Qu.	14.12	20200.05	1286.48	3841.71	275.16
Median	30.05	30139.05	1831.12	4167.59	292.50
Mean	54.64	33299.88	1957.74	4143.93	291.80
3rd Qu.	90.87	44140.57	2727.20	4479.62	308.02
Max.	258.93	73083.50	4812.09	5321.41	318.35
Std.Dev.	59.22	16486.20	1086.44	538.62	17.13

Solana's closing prices exhibit significant variability, ranging from a minimum of $0.52 to a maximum of $258.93, with a mean of $54.64, reflecting the high volatility and dynamic nature of the cryptocurrency market. Bitcoin and Ethereum also show substantial ranges and volatility, with Bitcoin's closing prices spanning from $6642.11 to $73083.50, and Ethereum's from $153.29 to $4812.09, with means of $33299.88 and $1957.74, respectively. In contrast, the S&P 500 index's closing prices range from $2736.56 to $5321.41, with a mean of $4143.93, highlighting the relative stability of the traditional stock market. The CPILFESL for the U.S. shows values ranging from 265.46 to 318.35, with a mean of 291.80, indicating moderate inflation over the study period. To visually observe the flow of data, time series plots of the closing prices and CPILFESL are included in the Appendix (see figures 11 to 15).

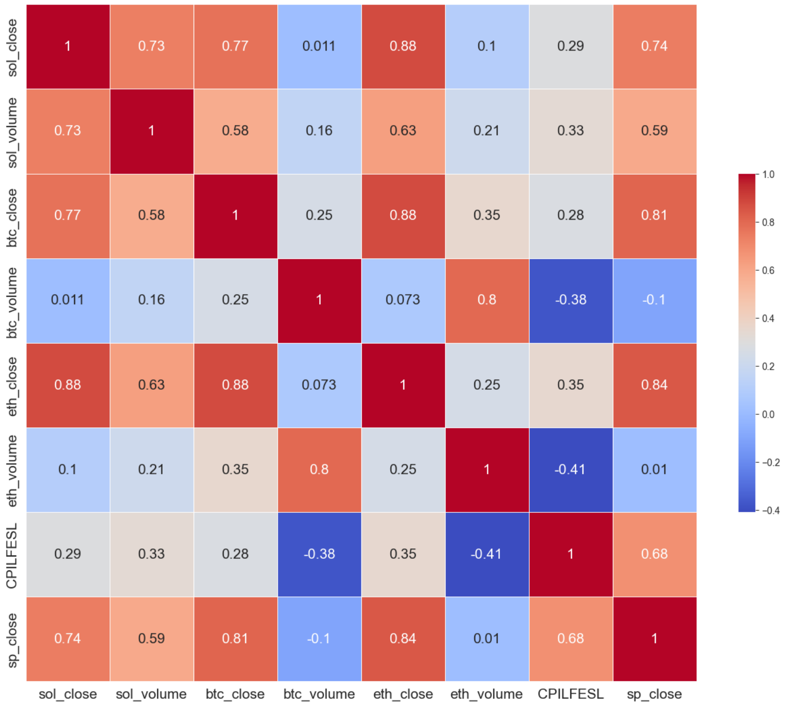


Figure 1: Correlation Matrix

Notably, all cryptocurrency closing prices—Solana, Bitcoin, and Ethereum—exhibit high positive correlations with each other. For instance, Solana's closing price (sol_close) shows a correlation of 0.77 with Bitcoin's closing price (btc_close) and 0.88 with Ethereum's closing price (eth_close). This strong positive correlation indicates that the price movements of these cryptocurrencies are closely aligned, suggesting that they often move in tandem, likely due to shared market factors and investor sentiment affecting the broader cryptocurrency market. Furthermore, Solana's trading volume (sol_volume) has a significant positive correlation with its closing price (0.73), implying that higher trading volumes are often associated with higher prices for Solana. Bitcoin and Ethereum also show similar patterns, though with varying correlation strengths.

The S&P 500 closing price (sp_close) also shows moderate to strong positive correlations with the cryptocurrency closing prices, particularly with Bitcoin (0.81) and Ethereum (0.84). This suggests that there may be some degree of co-movement between the stock market and cryptocurrency markets, although the correlation is not as strong as within the cryptocurrencies themselves.

Interestingly, CPILFESL shows a moderate positive correlation with the S&P 500 closing price (0.68) and weaker correlations with cryptocurrency closing prices, such as Solana (0.29) and Ethereum (0.35). This indicates that while inflation measures might influence traditional stock markets significantly, their impact on cryptocurrency markets is less direct.

The strong positive correlations observed among the cryptocurrency closing prices can potentially lead to issues of multicollinearity. Multicollinearity can adversely affect the performance and interpretability of traditional linear models, as it can inflate the variance of coefficient estimates and make it difficult to determine the individual effect of each predictor. However, the XGBoost model used in this study, a gradient boosting framework, is robust to such correlations. The framework of this model will be explored in detail in the Methodology (5).

Methodology

5.1. Introduction

This study adopts an explanatory approach, aiming to develop an XGBoost model to elucidate Solana price trends. The objective is to create a robust predictive model that not only forecasts price movements but also identifies the key variables driving these trends using Shapley Additive explanations (SHAP) values. Ultimately, this interpretation method will address the research question and hypotheses detailed in the introduction (2.2).

The resources used for this study include Python 3.12.4 and several libraries such as Scikit-learn (Sklearn), XGBoost, Optuna, and Shap. The development and testing of models were conducted using Jupyter Lab, providing an interactive environment for data analysis and model building. The algorithms were executed on a Windows desktop equipped with an AMD Ryzen 5 3600 6-Core Processor, 16 GB of RAM, and a 2060 GPU. This computational setup ensured efficient processing and model training, allowing for extensive hyperparameter tuning and model evaluation.

This section will first cover all the data pre-processing methods, ensuring the dataset is suitable for analysis. Following this, the design and implementation of the XGBoost model will be discussed, highlighting its robustness against multicollinearity due to its ability to handle correlated features effectively. The model evaluation will be detailed next, explaining the metrics used to assess model performance and the rationale behind selecting these metrics. The use of SHAP values for interpreting the model's results will then be explored, illustrating how these methods provide insights into the key variables influencing Solana price trends. Finally, the limitations of the employed methodology will be addressed, discussing potential biases and areas for future improvement.

5.2. Data Pre-Processing

Data preprocessing is a crucial step in ensuring the datasets are suitable for analysis and modeling. In this study, various preprocessing techniques were applied to handle differences in data frequency, missing values, and synchronization of time series.

5.2.1. Handling Missing Values

The datasets for the S&P 500, Solana, Bitcoin, and Ethereum all consist of daily data, whereas the CPILFESL data is available on a monthly basis. To harmonize these datasets, the monthly CPILFESL values were linearly interpolated to generate daily values. This method assumes that CPILFESL exhibits smooth changes between monthly observations, allowing for a more accurate alignment of different temporal frequencies. Linear interpolation is a common practice in economic research for aligning datasets with varying frequencies and ensuring that all data points are synchronized for robust analysis (Lepot et al., 2017).

The S&P 500 dataset contained missing values corresponding to weekends and public holidays when the stock market was closed. In contrast, the cryptocurrency market operates 24/7, resulting in a continuous dataset without such gaps. To address the missing values in the S&P 500 data, forward filling was implemented. This technique involves carrying the last observed value forward to fill the missing entries, ensuring continuity in the dataset. Given that the missing days in the S&P 500 dataset are weekends and holidays when the market is closed and no price fluctuations occur, forward filling is particularly appropriate. Additionally, Kamalov and Sulieman, 2021 found that forward and backward fill methods are well-suited for time series with large positive correlations, further validating the use of forward filling for this study (see figure 1).

5.2.2. Feature Engineering

Solana Price Trend

To determine Solana's price trends from its closing prices, a binary feature was created using the Moving Average Convergence Divergence (MACD) method, which is a widely recognized momentum indicator in technical analysis.

The first step involved calculating the MACD line, which is the difference between the 6-day Exponential Moving Average (EMA) and the 13-day EMA of Solana's closing price. Subsequently, the Signal Line was computed as the 9-day EMA of the MACD line. The formula used to calculate the EMA is as follows:

$$EMA_t = \alpha \cdot P_t + (1 - \alpha) \cdot EMA_{t-1} \tag{5.1}$$

where:

- EMA_t is the EMA at time t
- P_t is the price at time t
- α is the smoothing factor, calculated as $\frac{2}{N+1}$, where N is the number of periods
- EMA_{t-1} is the EMA at time $t-1$

The trend classes were defined based on the relationship between the MACD line and the Signal Line. Specifically, when the MACD line was above the Signal Line, it indicated an upward movement, and the trend was marked with a value of 1. Conversely, when the MACD line was below the Signal Line, it indicated a downward or stable movement, and the trend was marked with a value of 0.

The resulting class distribution (figure 2) shows negligible class imbalance, which brings several benefits. Firstly, improved model performance is achieved, as models trained on balanced datasets are less likely to be biased towards the majority class, leading to more accurate and fair predictions. Secondly, the training process is simplified since there is no need for special techniques to address imbalance, such as oversampling, undersampling, or adjusting class weights. Thirdly, performance metrics like accuracy, precision, recall, and F1-score provide a more reliable evaluation of the model's performance when classes are balanced. Lastly, a model trained on balanced data is more likely to generalize well to new, unseen data, ensuring robust performance in real-world applications.

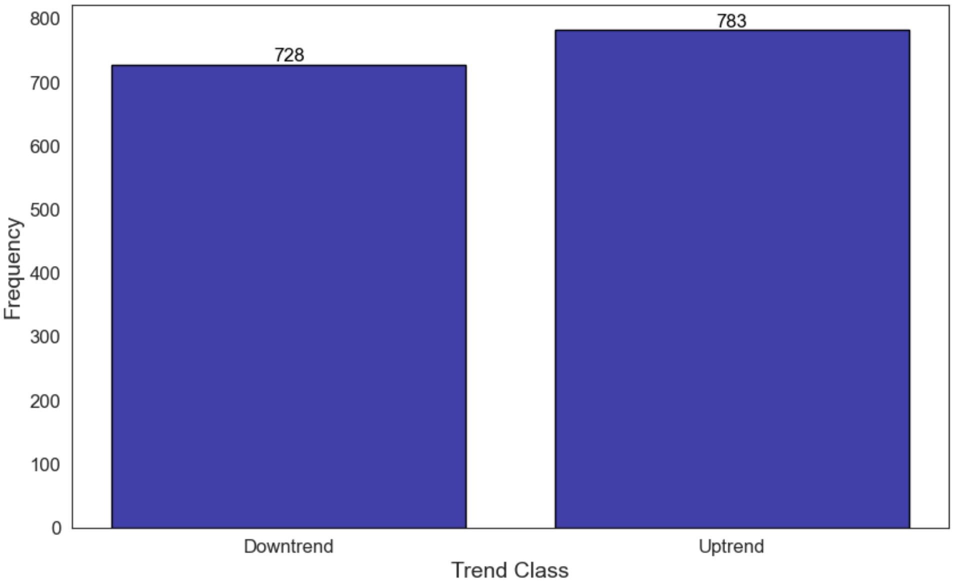


Figure 2: Solana Trend Distribution

Lagged Features

Creating explicit lagged features can enhance model performance; however, it also increases the complexity of feature interpretability. Different lags may have varying impacts on Solana trends, complicating the generalization of the impact of the original feature.

Another consideration is autocorrelation, the inherent correlation of time series features with their own past values. Autocorrelation can pose challenges in time series analysis. Thus, XGBoost was chosen for this study due to its ability to effectively handle autocorrelated features and its resilience against multicollinearity.

Considering these factors, the final decision was made to implement 10 lags for all features. Each lag in this study represents a day, so the 10th lag corresponds to data from 10 days ago. This approach ensures that the temporal nature of the data is accounted for, leveraging XGBoost's strengths in handling highly correlated variables while enhancing performance through the inclusion of more lags. However, interpreting results will necessitate a more nuanced analysis of the impacts of all the lags on Solana price trends to adequately address the hypotheses.

5.2.3. Z-Scale Normalization

Z-scale normalization, also known as standardization, is a technique used to transform data into a standard normal distribution with a mean of zero and a standard deviation of one. This method is particularly useful in machine learning and statistical analysis as it ensures that each feature contributes equally to the model, preventing features with larger scales from dominating the learning process. The formula for z-scale normalization is:

$$z = \frac{x - \mu}{\sigma} \tag{5.2}$$

where z is the standardized value, x is the original value, μ is the mean of the dataset, and σ is the standard deviation of the dataset. This technique is essential for improving the performance and convergence of gradient-based optimization algorithms like XGBoost. (Abdi, 2022).

Z-scale normalization was applied to the entire feature set except for the Solana trend variable, which is a binary indicator and does not require normalization. This step was crucial due to the significant differences in the scales of the feature set, with volumes traded ranging in the billions and other features like prices, S&P 500 index, and CPILFESL ranging in the hundreds and thousands. Normalizing these features ensures that they contribute equally to the predictive model, avoiding bias towards features with larger scales.

5.2.4. Data Split

The dataset was split into a training set and a test set in an 80-20 ratio, maintaining the temporal order of the data. This approach ensures that the last 20 percent of the data, representing the most recent time period, is used as the test set.

5.3. Model Development

5.3.1. XGBoost

XGBoost, short for eXtreme Gradient Boosting, is a powerful and efficient implementation of gradient boosting algorithms. It has gained popularity for its scalability, speed, and performance in machine learning tasks (T. Chen & Guestrin, 2016). XGBoost enhances traditional gradient boosting methods by incorporating several algorithmic optimizations and system design improvements.

XGBoost is designed to handle a variety of predictive modeling tasks, including regression, classification, and ranking, by building an ensemble of decision trees sequentially, where each tree corrects the errors of its predecessors. The primary goal of XGBoost is to minimize the loss function by iteratively adding new trees that predict the residuals of the previous trees. The overall prediction is given by:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i) \tag{5.3}$$

where $\hat{y}_i$ is the predicted value for instance i, K is the number of trees, and f_k represents the k-th decision tree.

The objective function to be minimized in XGBoost consists of two parts: the loss function and the regularization term. The objective function is given by:

$$\text{Obj} = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{5.4}$$

where l is the loss function that measures the difference between the actual and predicted values, and Ω is the regularization term that penalizes the complexity of the model to avoid overfitting.

The regularization term $\Omega(f_k)$ is defined as:

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{5.5}$$

where γ is a parameter that controls the complexity of the tree (number of leaves T), λ is a regularization parameter, and w_j are the leaf weights.

XGBoost was selected for this study due to its numerous advantageous properties, making it an ideal choice for predicting Solana price trends. Its efficiency, scalability, and high performance are well-documented, allowing it to handle large datasets and complex models with ease. XGBoost has been successfully employed in previous studies on cryptocurrency price prediction, as detailed in the literature review (3), validating its effectiveness in this domain. Additionally, XGBoost's ability to handle autocorrelated values and multicollinearity is particularly beneficial for this study, given the presence of such characteristics in the feature set. This capability ensures that the model can manage complex interactions within the data without compromising accuracy or stability.

5.3.2. Hyperparameter Tuning

Hyperparameter tuning is a critical step in optimizing the performance of the XGBoost model. In this study, hyperparameters were tuned using Bayesian optimization and cross-validation with the goal of maximizing accuracy. The final model was selected based on the best performance metrics obtained from the training set. The hyperparameters used in this study are detailed below and shown in table 4.

Tree-specific hyperparameters control the construction and complexity of the decision trees:

- **Maximum Depth:** This parameter defines the maximum depth of a tree. Deeper trees have the capability to capture more intricate patterns within the data, but they also run the risk of overfitting.
- **Minimum Leaf Weight:** This parameter establishes the minimum sum of instance weight (hessian) required in a child node. It controls the complexity of the decision tree by preventing the formation of overly small leaves.
- **Subsample Ratio:** This parameter specifies the percentage of rows utilized for the construction of each tree. Reducing this value can help prevent overfitting by training on a smaller subset of the data.
- **Feature Subsample Ratio:** This parameter determines the percentage of columns used for each tree construction. Lowering this value can prevent overfitting by training on a subset of the features.

Learning task-specific hyperparameters govern the overall behavior of the model and the learning process:

- **Learning Rate:** Also known as the step size shrinkage, this parameter is used in updates to mitigate overfitting. Smaller values make the model more robust by taking smaller steps during training.
- **Gamma:** This parameter sets the minimum loss reduction necessary to make an additional partition on a leaf node of the tree. Higher values enhance the regularization.
- **Lambda:** This is the L2 regularization term on weights. Increasing this value strengthens the regularization.
- **Alpha:** This is the L1 regularization term on weights. Higher values enhance the regularization.

Table 4: Hyperparameters Tuned in XGBoost Model

Hyperparameter	**Range**	**Optimal Value**
Maximum Depth	[3, 10]	7.00
Min Leaf Weight	[1, 10]	6.00
Subsample Ratio	[0.5, 1.0]	0.98
Feature Subsample Ratio	[0.5, 1]	0.69
Learning Rate	[0.01, 0.3]	0.01
Gamma (γ)	[0, 5]	1.32
Lambda (λ)	[0.1, 10]	0.26
Alpha (α)	[0, 10]	2.72

5.4. Model Evaluation

The tuned XGBoost model will be fitted on the training set and used to make predictions on the test set. The performance of the model will be evaluated using several key metrics such as the confusion matrix, accuracy, precision, recall, F1-score and specificity.

The confusion matrix provides a detailed breakdown of the model's performance, showing the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). This allows for a clear visualization of how well the model distinguishes between classes.

Accuracy represents the proportion of correctly predicted instances (both true positives and true negatives) out of the total instances. It is defined as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision measures the accuracy of positive predictions, defined as the ratio of true positive predictions to the total positive predictions:

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall indicates the model's ability to correctly identify positive instances, calculated as the ratio of true positive predictions to the actual positives:

$$\text{Recall} = \frac{TP}{TP + FN}$$

The F1-score, the harmonic mean of precision and recall, provides a balance between the two. It is particularly useful when the class distribution is relatively balanced, as it considers both false positives and false negatives:

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

Specificity measures the proportion of actual negatives that are correctly identified, providing insight into the model's ability to detect negative instances:

$$\text{Specificity} = \frac{TN}{TN + FP}$$

5.5. Model Interpretation

The interpretation of the XGBoost model is a crucial aspect of this study, enabling us to understand the influence of various features on Solana's price trends. To achieve this, Shapley Additive Explanations (SHAP) values are employed, as explained below.

SHAP values are a unified measure of feature importance based on cooperative game theory. They provide a way to explain the output of any machine learning model by attributing the contribution of each feature to the prediction. The core idea of SHAP values is to fairly distribute the prediction among the features, ensuring that the sum of the feature contributions equals the actual prediction minus the mean prediction. This is achieved by considering all possible combinations of features and calculating the average contribution of a feature across these combinations (Rozemberczki et al., 2022).

Mathematically, the SHAP value for a feature i in a given prediction is defined as:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} [f(S \cup \{i\}) - f(S)] \tag{5.6}$$

where:

- ϕ_i is the SHAP value for feature i,
- N is the set of all features,
- S is any subset of N that does not include feature i,
- $f(S)$ is the prediction for the subset of features S,
- $|S|$ is the number of features in subset S,
- $|N|$ is the total number of features.

SHAP values are particularly useful for interpreting machine learning models because they provide a clear and consistent method for understanding the impact of each feature on the model's predictions. By decomposing the prediction into contributions from each feature, SHAP values allow us to see how individual features influence the model's output, which is essential for gaining insights into the underlying data and the model's behavior. In this study, SHAP values will be calculated for all lags of all features to obtain a nuanced understanding of their impact on Solana price trends.

Additionally, to capture the aggregated effects of all lags per feature, a cumulative SHAP value plot will also be derived. This cumulative SHAP value is calculated by summing the SHAP values of all lags for each feature, providing a holistic view of the feature's overall impact on the model's predictions. The formula for the aggregated SHAP value for a group of features G is:

$$\phi_G = \sum_{i \in G} \phi_i \tag{5.7}$$

where each ϕ_i represents the SHAP value for a specific lag of a feature.

This approach ensures that we capture the temporal dynamics and their combined influence on the prediction. In this formulation, we do not take the absolute values of the individual SHAP values, thus maintaining the correct attribution of feature interactions and their directions. This method allows us to understand not only the impact of individual lags but also the aggregated influence of a feature over time, providing deeper insights into the factors driving Solana price trends.

5.6. Limitations in Methodology

Despite the comprehensive approach adopted in this study, several limitations must be acknowledged to provide a balanced perspective on the findings and the methodologies employed.

Firstly, the dataset includes a range of features such as cryptocurrency prices, trading volumes, the S&P 500 index, and the CPILFESL. While these features provide a comprehensive view of market factors, they may not cover all relevant variables that influence Solana price trends. Factors such as regulatory changes, technological advancements, and macroeconomic policies were not included, potentially limiting the model's explanatory power.

Secondly, the creation of 10 lagged features, while intended to capture temporal dependencies, introduces the challenge of multicollinearity. Although XGBoost is relatively robust to multicollinearity, the presence of highly correlated lagged features can still impact the interpretability of the model outputs, particularly when using interpretation methods such as SHAP values.

Thirdly, the study uses a train-test split of 80-20 to maintain the temporal order of the data, with the last 20 percent being the most recent data. While this approach preserves the sequential nature of time series data, it may not account for potential structural breaks or regime shifts within the dataset. Such events could affect the stability and predictability of the model.

Lastly, the generalizability of the findings might be limited by the specific timeframe and market conditions considered in this study. The data spans from April 2020 to June 2024, a period characterized by significant economic and market events that may not be representative of other time periods. As such, the model's applicability to different market conditions or future scenarios remains to be validated.

In summary, while the methodologies employed in this study are robust and well-supported by existing research, these limitations highlight the need for careful interpretation of the results and suggest areas for future research to address these constraints.

Results

6.1. Model Assessment

Table 5: Model Evaluation Metrics

Metric	Value
Accuracy	0.73
Precision	0.75
Recall	0.75
F1 Score	0.75
Specificity	0.71

The evaluation metrics for the XGBoost model indicate solid performance in predicting Solana price trends. An accuracy of 0.73 means the model correctly classified 73 percent of all instances, which is a strong indication of its reliability. With precision, recall and F1 score at 0.75, the model shows a good balance between identifying true positives and minimizing false positives. In addition, the specificity of 0.71 suggests the model also performs well in correctly identifying negative instances. A confusion matrix for visual representation is provided in the Appendix (figure 16).

6.2. Hypotheses

The cumulative SHAP values plot offers a comprehensive insight into the overall impact of each feature on the model's predictions for Solana's price trends. Using insights from the cumulative SHAP values, as well as the individual SHAP values of each feature's lags, we can address the following hypotheses with detailed analysis.

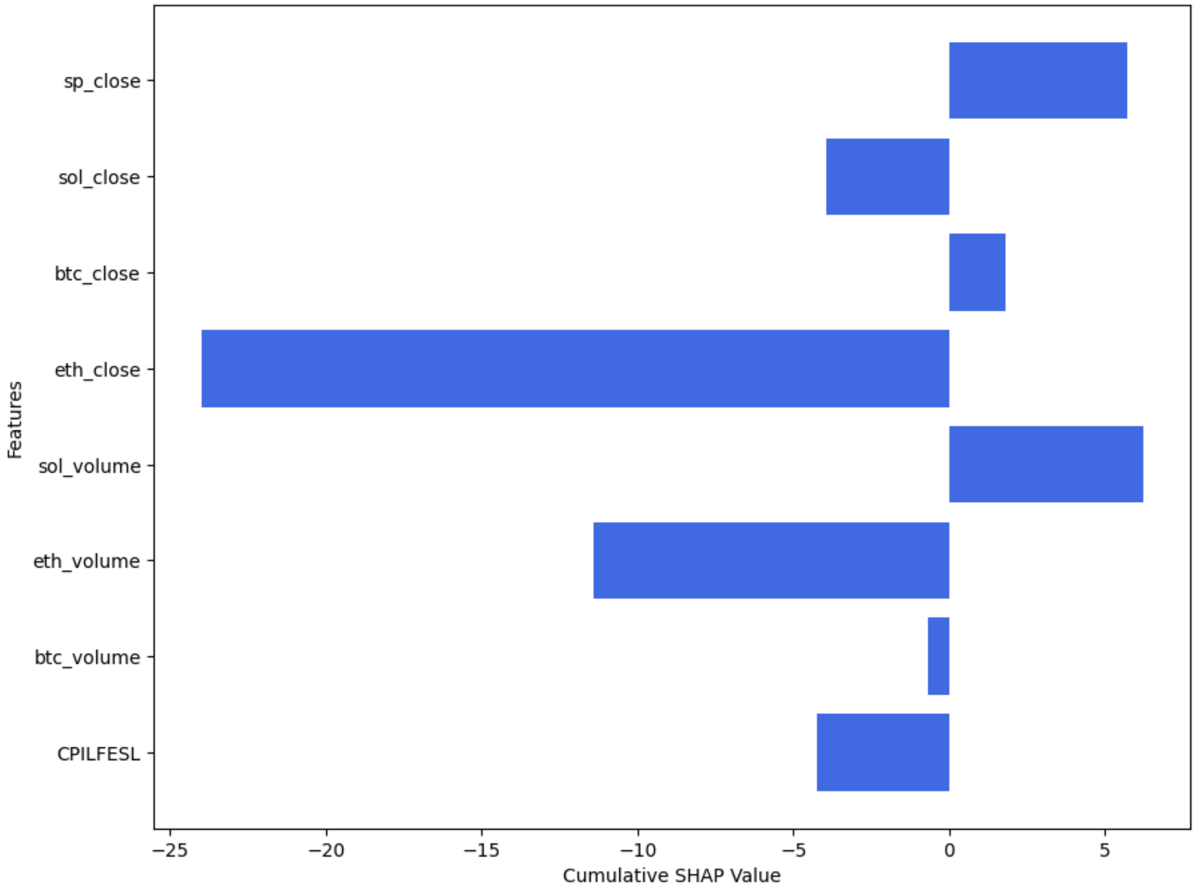


Figure 3: Cumulative SHAP Values

6.2.1. H1: An increase in the S&P 500 index results in an upward price for Solana

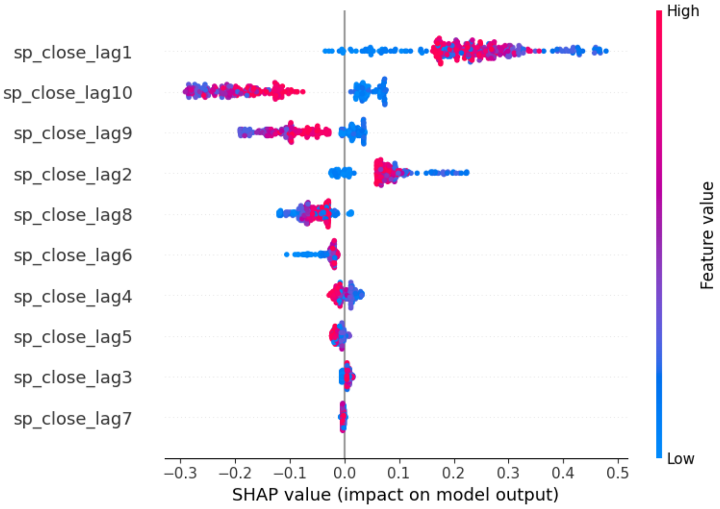


Figure 4: S&P 500 SHAP Values

The S&P 500 index (sp_close) feature has a high positive cumulative SHAP value, suggesting that an increase is likely to result in an upward trend for Solana's price. This highlights the broader stock market's health has a significant influence on Solana's price trends.

In the SHAP values plot for individual lags of the sp_close feature, we observe that higher values (represented in red) of lag 1 and lag 2 are associated with positive SHAP values, indicating a strong positive impact on the prediction of an upward trend. This implies that recent increases in the S&P 500 index are strong indicators of a positive trend in Solana's price.

In contrast, medium-term lags such as lag 3 to lag 7 show mixed impacts. These lags exhibit both positive and negative SHAP values, suggesting that these past values have a less consistent effect on Solana's price trends. Additionally, the SHAP values for these medium-term lags are very low and negligible compared to the other terms, indicating that their overall impact on the model's predictions is minimal. This may indicate that while immediate changes in the S&P 500 index have a clear and strong impact, the influence of medium-term values is more variable, context-dependent, and generally less significant.

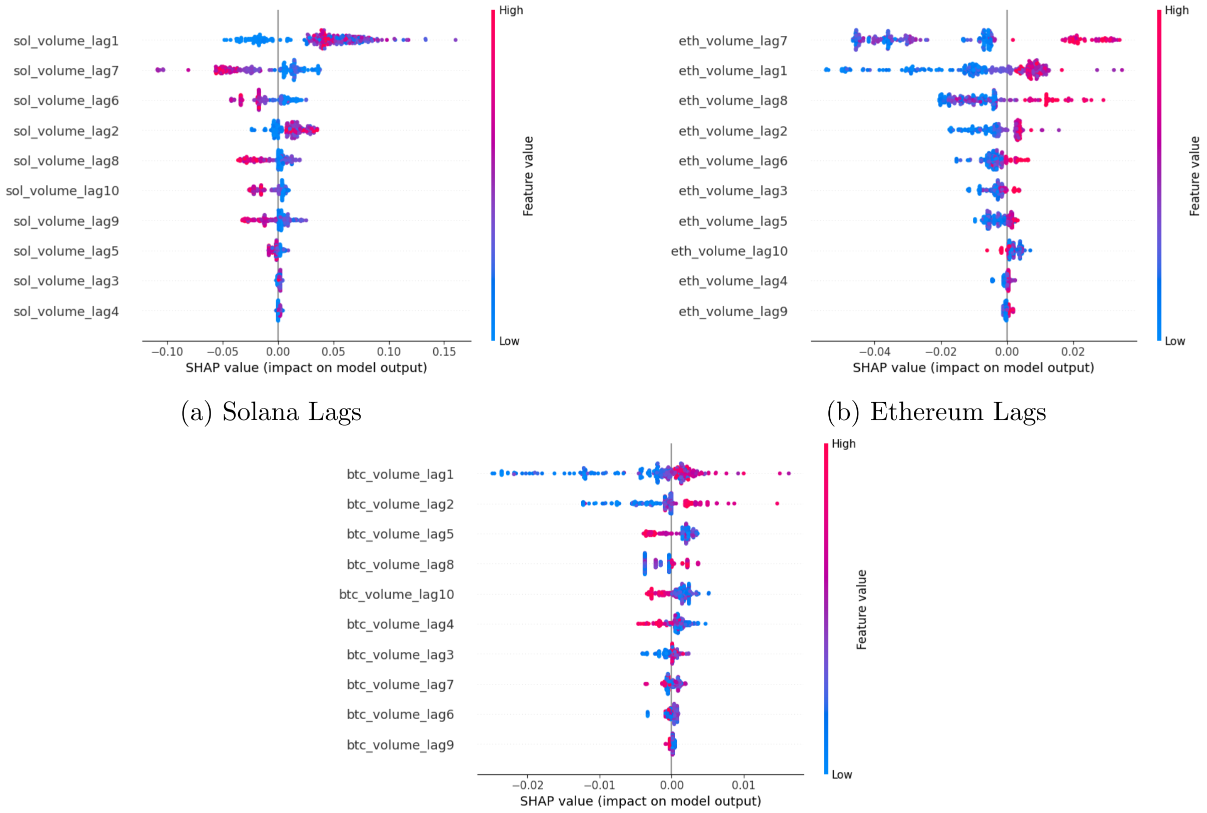
The later lags, such as lag 8 to lag 10, show significant negative impacts with higher values. This suggests that higher values of the S&P 500 index in the more distant past are associated with a downward trend in Solana's price. This might be interpreted as a correction or market adjustment effect where past high values eventually lead to a decrease in Solana's price.

In summary, the overall effect of the S&P 500 index, particularly its most recent lags, supports the hypothesis that it influences Solana's price trends positively. However, the S&P 500 index in the long term seems to have a negative influence, indicating that the effect on Solana's price trends is multifaceted and varies across different time horizons.

6.2.2. H2: An increase in transaction volumes for Solana, Bitcoin and Ethereum leads to an upward price trend for Solana

The volume feature of Solana (sol_volume) exhibits a strong positive cumulative SHAP value, suggesting that an increase in volume is associated with an upward price trend. This indicates that heightened network activity correlates with increased demand for Solana, thereby driving its price trends upward.

Higher values of Solana's volume lag 1 consistently correspond to high positive SHAP values, whereas lower values are associated with negative SHAP values. This indicates that an increase in transaction volume at lag 1 strongly suggests an upward price trend for Solana. Lag 2 follows a similar pattern, though with comparatively lower SHAP values, emphasizing the significant positive impact of recent transaction volumes on Solana's price movements. Lags 3 to 5 present mixed SHAP values for both high and low feature values, and their overall low SHAP values suggest that mid-term transaction volumes have a minimal impact on the model's predictions. Interestingly, lags 6 to 10 of Solana's volume display an inverse effect compared to lags 1 and



(a) Solana Lags

(b) Ethereum Lags

(c) Bitcoin Lags

Figure 5: Solana, Ethereum & Bitcoin Volumes SHAP Values

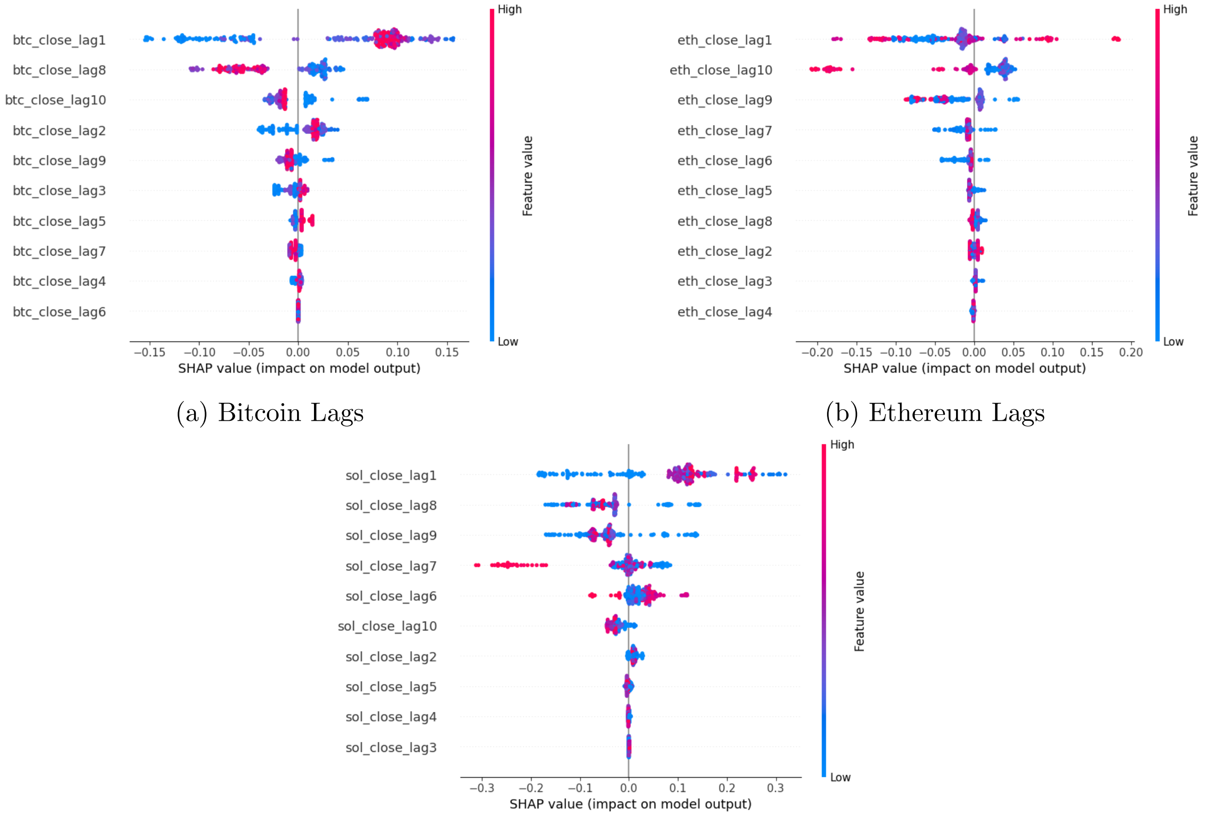
2, with higher transaction volumes indicating a downward price trend for Solana. Notably, lags 6 and 7 have relatively high negative SHAP values, illustrating that in the long term, increased transaction volumes are associated with a downward price trend for Solana.

Ethereum's volume (eth_volume) feature has a high negative cumulative SHAP value, second only to its closing price, indicating that higher trading volumes for Ethereum may be associated with a downward trend in Solana's price. Similarly, Bitcoin's volume (btc_volume) feature also shows a negative cumulative SHAP value, though with a lesser impact compared to Ethereum. When users engage more with Ethereum or Bitcoin, there may be a corresponding decrease in Solana's user base, leading to downward pressure on Solana's price. This dynamic highlights how interdependencies and competition within the cryptocurrency market can affect individual blockchain networks.

The SHAP plot for Ethereum's volume lags reveals that higher values across almost all lags are associated with positive SHAP values, suggesting that increase corresponds to an upward trend in Solana's price. This observation contrasts with the cumulative SHAP values plot, which indicates a negative overall impact for Ethereum's volume. The most recent lags for Bitcoin's volume (btc_volume), lag 1 and lag 2, show a positive trend, suggesting that an increase in Bitcoin trading volume leads to an upward trend in Solana's price. However, the other lags display mixed or negative trends. Overall, the SHAP values for Bitcoin volume's lags are relatively low, indicating a minimal effect on predicting Solana's price trends.

In summary, the recent lags of Bitcoin, Ethereum, and Solana support the hypothesis that these factors positively influence Solana's price trends, with Solana's cumulative SHAP value further reinforcing this positive impact. However, in the long term, the influence appears to be negative, and the cumulative SHAP values for Bitcoin and Ethereum also indicate a negative impact. This suggests that the effect on Solana's price trends is varied between cryptocurrencies and changes across different time horizons.

6.2.3. H3: Upward price movements of Bitcoin, Ethereum and Solana result in upward trends in Solana's market price



(a) Bitcoin Lags

(b) Ethereum Lags

(c) Solana Lags

Figure 6: Bitcoin, Ethereum and Solana Prices SHAP Values

Bitcoin's closing price (btc_close) feature shows a small positive cumulative SHAP value, indicating that higher Bitcoin closing prices tend to predict an upward trend for Solana. Considering Bitcoin is the leading cryptocurrency on the broader market, it's price increase may signal investors in the health of the overall cryptocurrency market leading to more investments in solana. Ethereum's closing price (eth_close) feature exhibits the highest absolute cumulative SHAP value, suggesting it is the most influential factor in predicting Solana's price trends. The negative cumulative SHAP value for eth_close indicates that higher values are generally associated with a downward trend in Solana's price. This significant influence may be attributed to the interconnectedness and competitive dynamics within the cryptocurrency market. Surprisingly, Solana's closing price (sol_close) feature also shows a small negative cumulative SHAP value, indicating that higher historical closing prices of Solana itself are predictors of a downward trend in its future prices.

No consistent trend can be discerned from the individual lags of Bitcoin, Solana and Ethereum. They all show a mixed effect, with both lower and higher values of the closing price leading to upward and downward trend predictions. This underscores the complex and nuanced relationship between the historical prices of these cryptocurrencies and Solana's price trends.

In summary, this hypothesis is invalid as cumulative SHAP values indicate that Ethereum and Solana's historical prices lead to downward price trends for Solana. Although Bitcoin does show a positive impact, its SHAP value is very low. Furthermore, the individual lags present a mixed effect with no discernible trend.

6.2.4. H4: An increase in the CPI of the U.S. results in a decrease in Solana's price trend

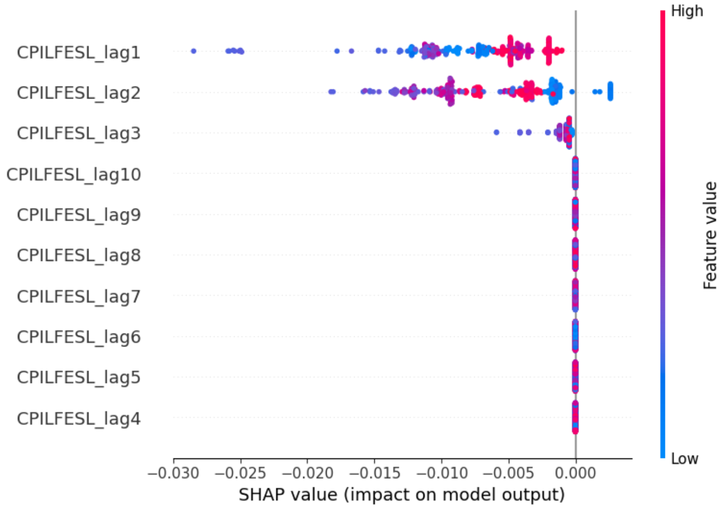


Figure 7: CPILFESL SHAP Values

The CPILFESL exhibits a negative cumulative SHAP value, indicating that increases in the CPILFESL of the U.S. are associated with a decrease in Solana's price. Additionally, the SHAP values for its individual lags are predominantly negative or zero, with very few positive values, reinforcing this negative correlation. Given these insights, there is substantial evidence to support the hypothesis that higher CPILFESL values lead to downward trends in Solana's price. This relationship may stem from investors' tendencies to shift towards more stable assets during inflationary periods, thereby reducing demand for more volatile investments like cryptocurrencies.

Conclusion

This study sought to answer the research question:

What is the impact of the U.S. stock market, U.S. inflation, and blockchain market metrics on Solana's price trends?

Using the S&P 500 index as a proxy for the U.S. stock market health, the analysis revealed a mixed effect on Solana's price trends. When considering individual lags, the impact was varied, but the cumulative effect of the S&P 500 index on Solana's trends was positive, suggesting that overall, movements in the U.S. stock market have a positive influence on Solana's price trends.

Regarding blockchain metrics, the study examined the closing prices and trading volumes of Bitcoin, Ethereum, and Solana. Ethereum's closing price had the strongest negative impact on Solana's price trends, followed by Solana's own historical prices. Conversely, Bitcoin's closing prices had a positive cumulative effect on Solana, though individual lags for all blockchains showed mixed results. In terms of trading volumes, recent lags of Bitcoin, Ethereum, and Solana positively influenced Solana's price trends, with Solana's cumulative SHAP value reinforcing this positive impact. However, the long-term influence of these volumes appeared negative, with cumulative SHAP values for Bitcoin and Ethereum also indicating a negative impact. This variability underscores the complex relationships within the cryptocurrency market.

U.S. inflation, proxied by the Consumer Price Index for All Urban Consumers: All Items Less Food and Energy (CPILFESL), had the clearest effect on Solana's price trends. Both cumulative and individual lag analyses showed that increases in inflation led to downward trends in Solana's prices, suggesting a strong inverse relationship.

In conclusion, this research contributes to the existing literature by incorporating a broader range of influential factors and leveraging advanced machine learning techniques for enhanced interpretability. The findings highlight the complex interplay between stock market indicators, macroeconomic indicators, and cryptocurrency-specific metrics. Future studies should continue to explore these relationships to refine predictive models and support better-informed investment decisions in the volatile cryptocurrency market.

Appendix

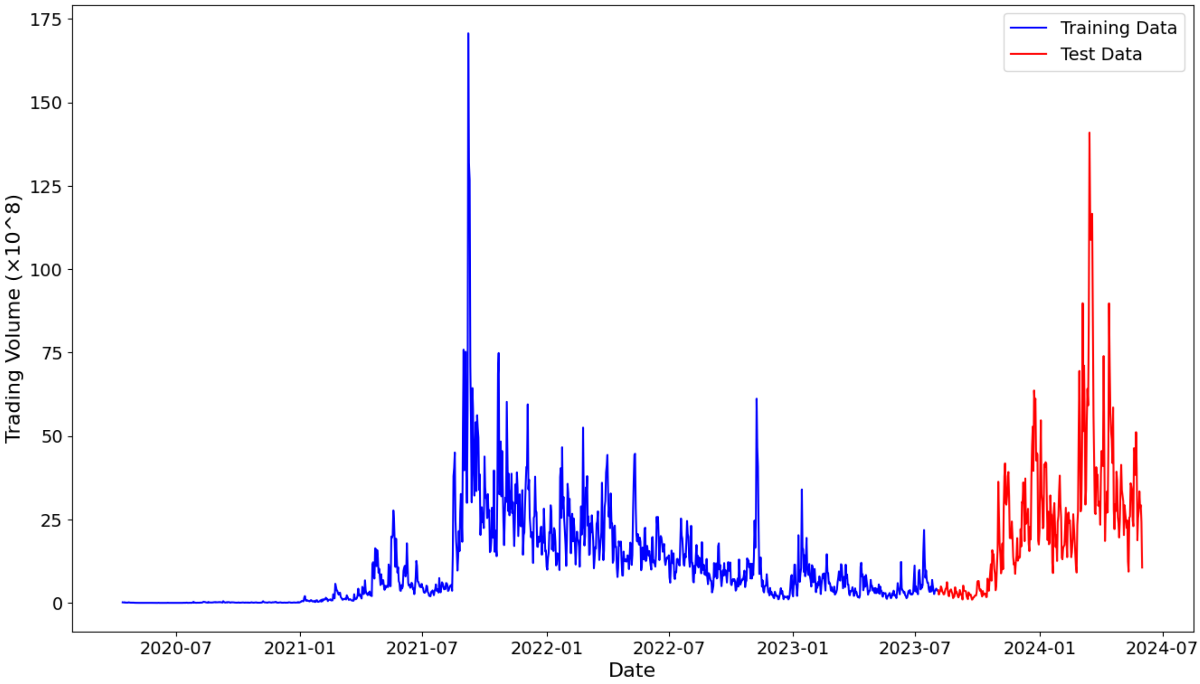


Figure 8: Solana Trading Volume Over Time

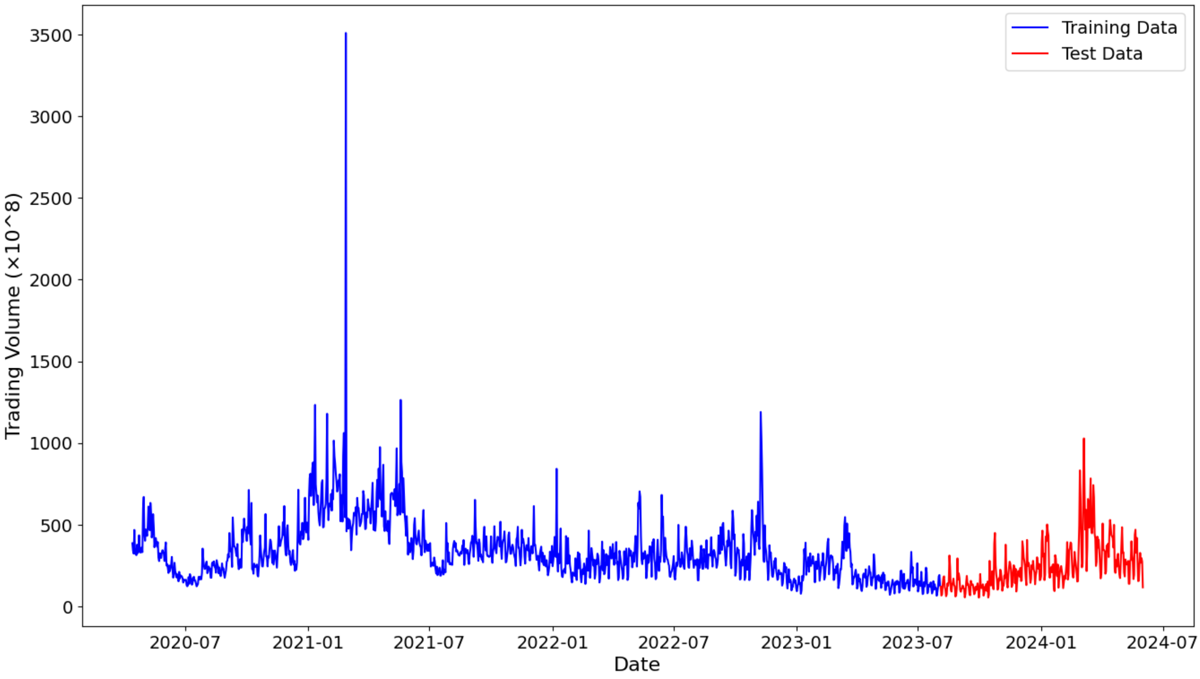


Figure 9: Bitcoin Trading Volume Over Time

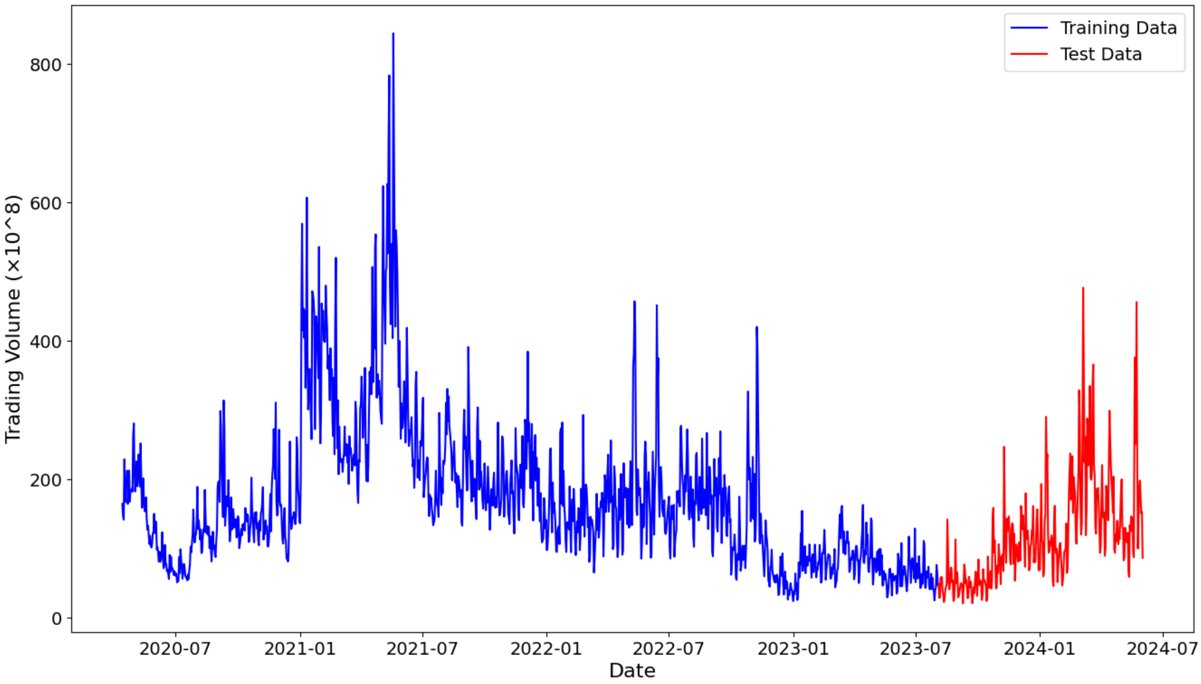


Figure 10: Ethereum Trading Volume Over Time

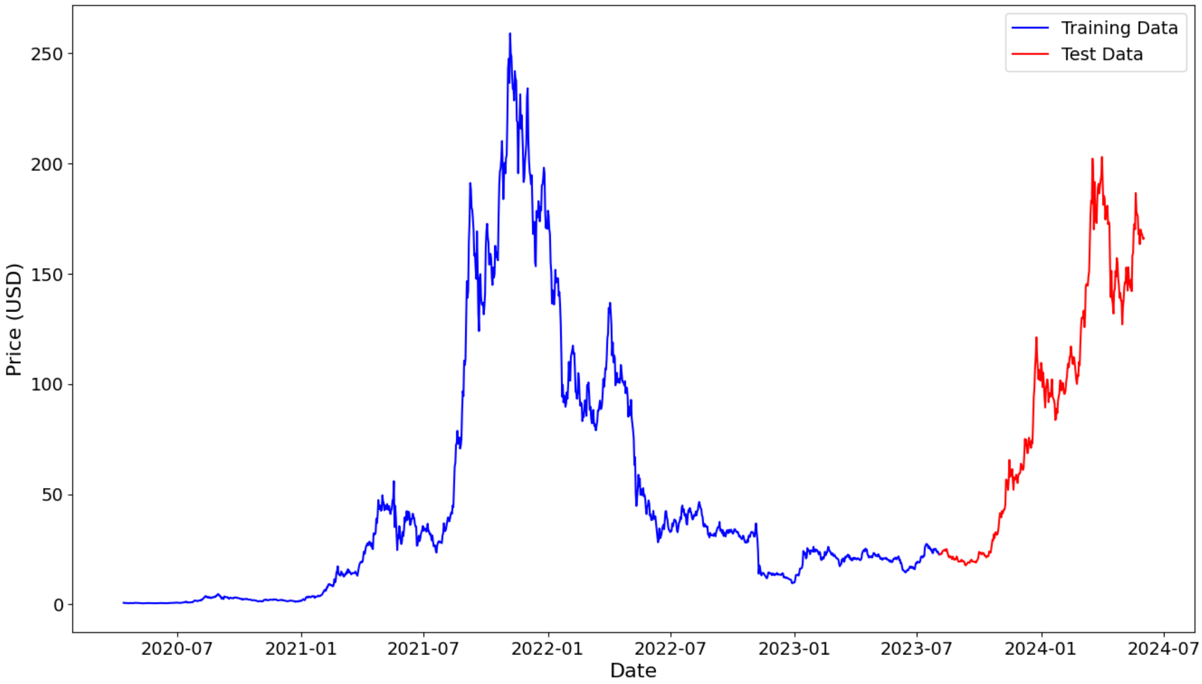


Figure 11: Solana Closing Prices Over Time

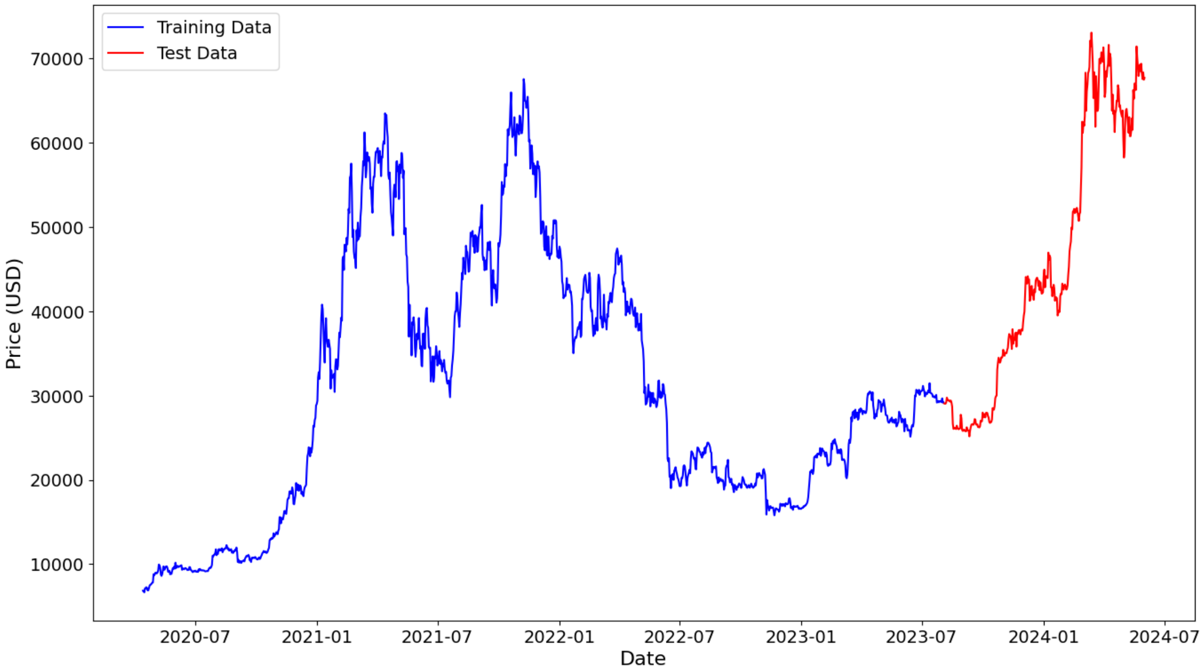


Figure 12: Bitcoin Closing Prices Over Time

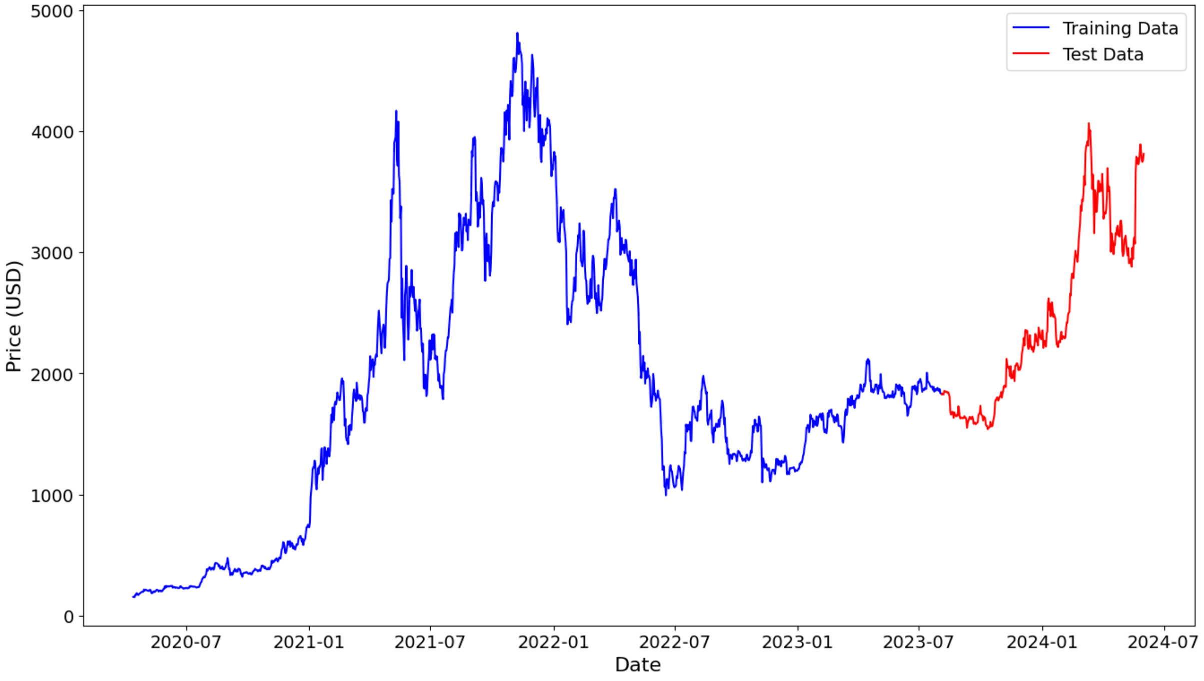


Figure 13: Ethereum Closing Prices Over Time

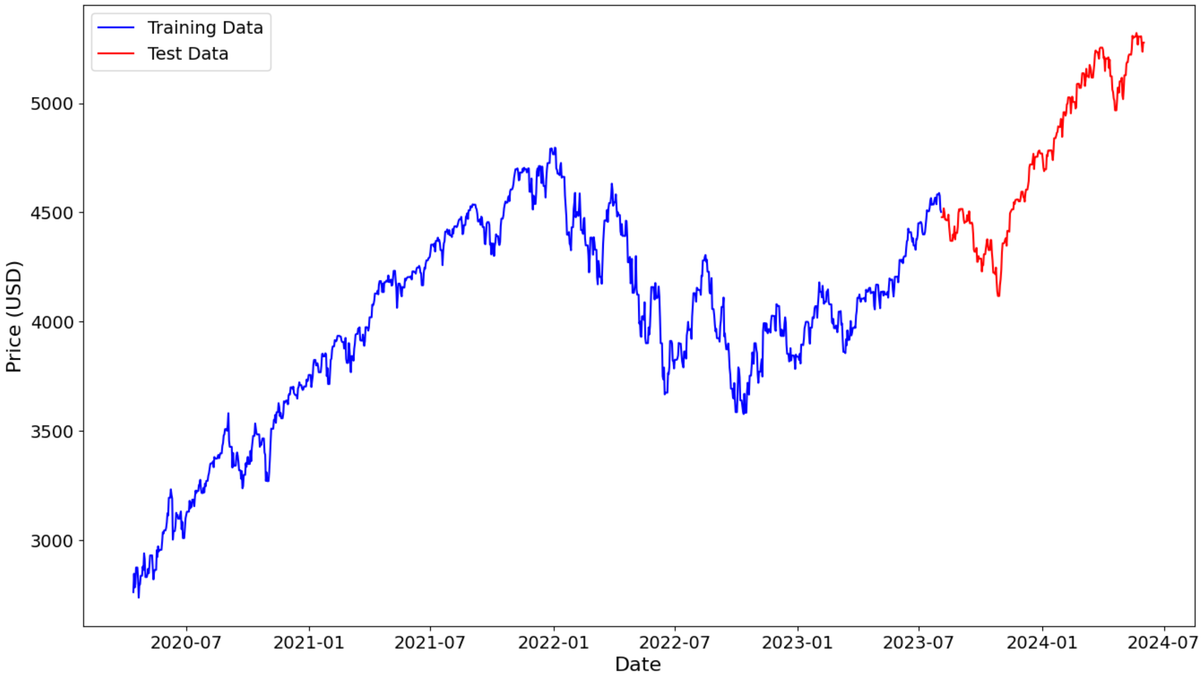


Figure 14: S&P 500 Closing Prices Over Time

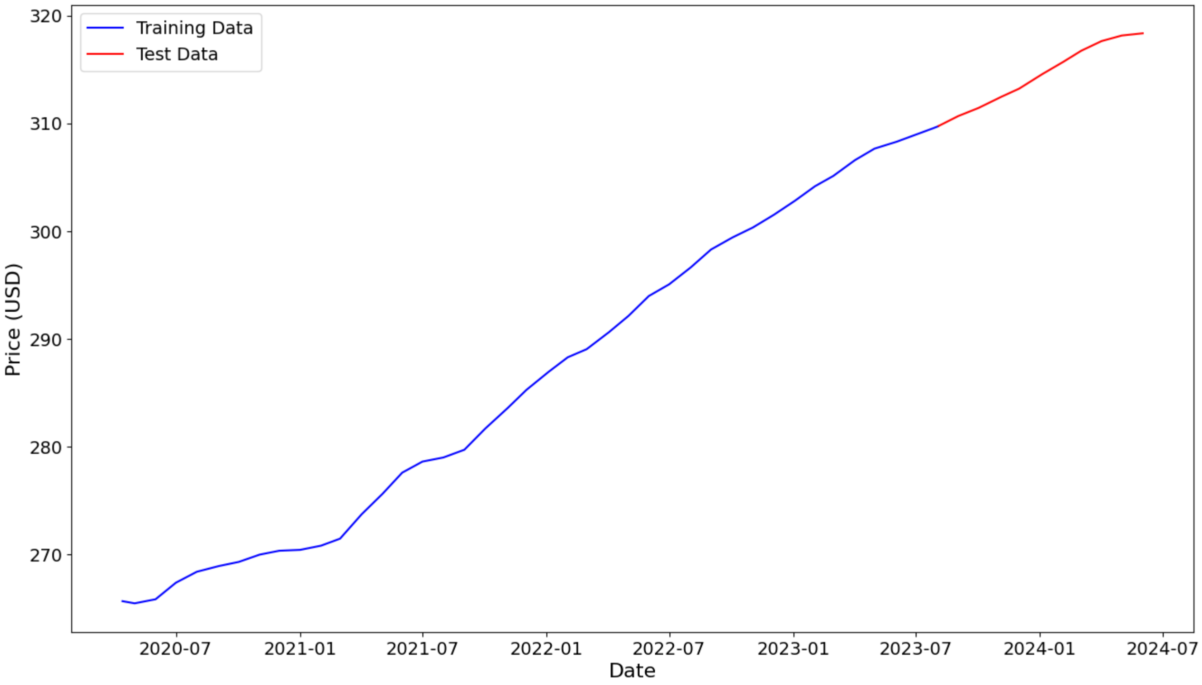


Figure 15: CPI US Over Time

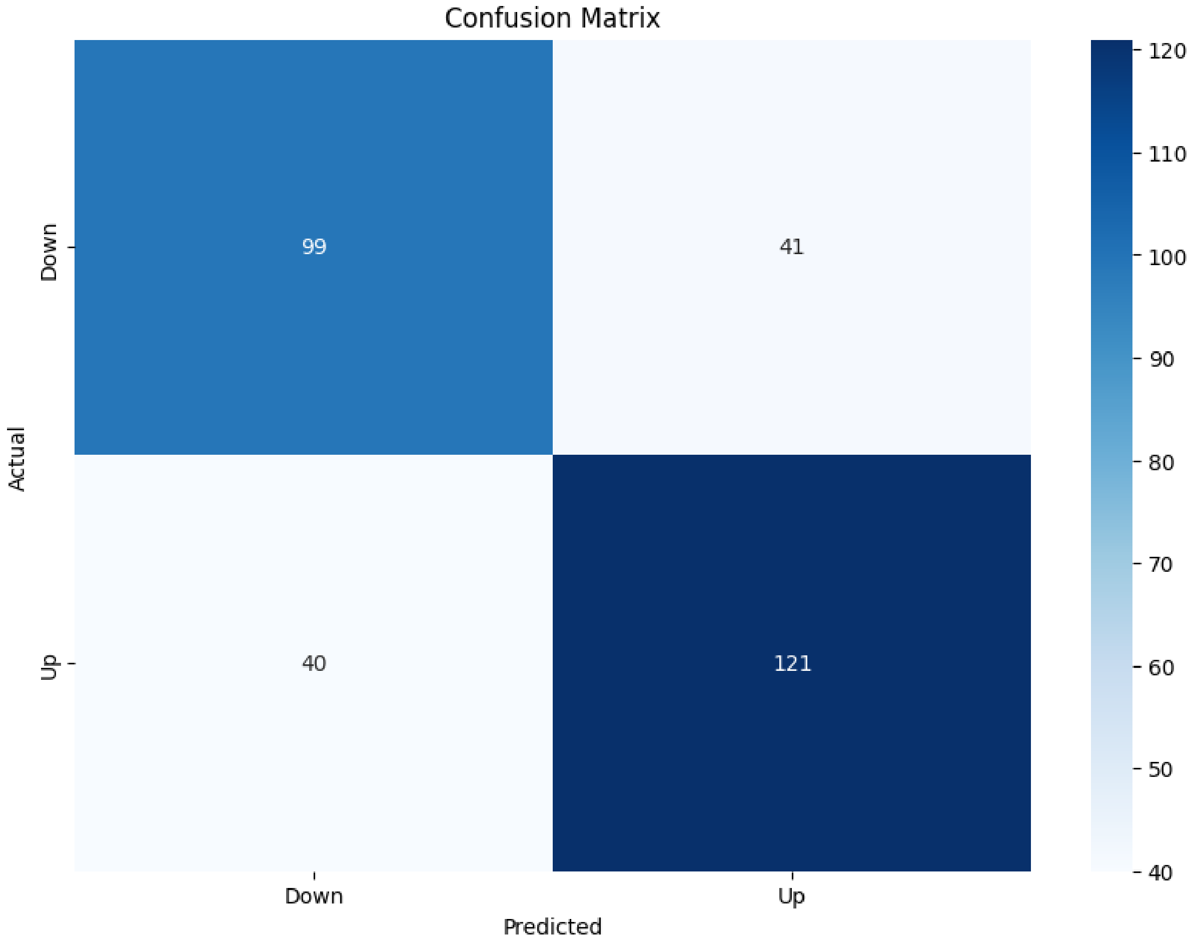


Figure 16: Confusion Matrix

Bibliography

Abdi, H. (2022). Normalizing data. In *Experiments of the mind*. Princeton University Press. https://doi.org/10.2307/j.ctv1n1bs5c.11

Abraham, J., Higdon, D., Nelson, J., & Ibarra, J. (2018). Cryptocurrency price prediction using tweet volumes and sentiment analysis. *SMU Data Science Review*, *1*(3), 1.

Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2021). A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review*, *54*, 1937–1967.

Bryan, M. F., & Cecchetti, S. G. (1993). The consumer price index as a measure of inflation. *Monetary Economics*. https://doi.org/10.3386/W4505

Buterin, V. (2013). Ethereum whitepaper.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 785–794.

Chen, Z., Li, C., & Sun, W. (2020). Bitcoin price prediction using machine learning: An approach to sample dimension engineering. *Journal of Computational and Applied Mathematics*, *365*, 112395. https://doi.org/10.1016/j.cam.2019.112395

Gurrib, I. (2022). Bitcoin price prediction using sentiment analysis. *Journal of Financial Technology and Innovation*, *2*(1), 45–60.

Hashemi, R., Ardakani, O., Bahrami, A., & Young, J. (2017). Extraction of the essential constituents of the s&p 500 index. *2017 International Conference on Computational Science and Computational Intelligence (CSCI)*, 351–356. https://doi.org/10.1109/CSCI.2017.59

Kamalov, F., & Sulieman, H. (2021). Time series signal recovery methods: Comparative study. *2021 International Symposium on Networks, Computers and Communications (ISNCC)*, 1–5. https://doi.org/10.1109/ISNCC52172.2021.9615669

Kim H., B. G.

bibinitperiod L. G. (2021). Predicting ethereum prices with machine learning based on blockchain information. *Expert Systems with Applications*, *184*, 115480. https://doi.org/10.1016/J.ESWA.2021.115480

Lepot, M., Aubin, J. B., & Clemens, F. H. (2017). Interpolation in time series: An introductive overview of existing methods, their performance criteria and uncertainty assessment. *Water*, *9*(10), 796.

McNally, S. (2018). Predicting the price of bitcoin using machine learning. *arXiv preprint arXiv:1803.04319*.

Mustapa, F., & Ismail, M. (2019). Modelling and forecasting s&p 500 stock prices using hybrid arima-garch model. *Journal of Physics: Conference Series*, *1366*. https://doi.org/10.1088/1742-6596/1366/1/012130

Nakamoto, S. (2008). Bitcoin: A peer-to-peer electronic cash system. https://bitcoin.org/bitcoin.pdf.

Olivier Kraaijeveld, J. D. S. (2020). The predictive power of public twitter sentiment for forecasting cryptocurrency prices. *Journal of International Financial Markets, Institutions and Money*, *65*, 101188. https://doi.org/10.1016/j.intfin.2020.101188

Rozemberczki, B., Watson, L., Bayer, P., Yang, H., Kiss, O., Nilsson, S., & Sarkar, R. (2022). The shapley value in machine learning. *ArXiv*, *abs/2202.05594*. https://doi.org/10.24963/ijcai.2022/778

Srivastava, V., Dwivedi, V., & Singh, A. (2023). Cryptocurrency price prediction using enhanced pso with extreme gradient boosting algorithm. *Cybernetics and Information Technologies*, *23*, 170–187. https://doi.org/10.2478/cait-2023-0020

Wołk, K. (2020). Advanced social media sentiment analysis for short-term cryptocurrency price prediction. *Expert Systems*, *37*(2), e12493. https://doi.org/10.1111/exsy.12493

Wu, J., Guo, X., Fang, M., & Zhang, J. (2022). Short term return prediction of cryptocurrency based on xgboost algorithm. *2022 International Conference on Big Data, Information and Computer Network (BDICN)*, 39–42. https://doi.org/10.1109/BDICN55575.2022.00015

Yakovenko, A. (2018). Solana: A new architecture for a high performance blockchain.

Zellner, J., Gallo, A., & Levey, B. (1980). Alternative measures of inflation. *Journal of Monetary Economics*, *10*. https://doi.org/10.22004/AG.ECON.281095