

Claremont Colleges

Scholarship @ Claremont

CMC Senior Theses

CMC Student Scholarship

2023

Application of Sentiment Analysis and Machine Learning Techniques to Predict Daily Cryptocurrency Price Returns

Edward Wu

Follow this and additional works at: https://scholarship.claremont.edu/cmc_theses



Part of the [Data Science Commons](#), [Econometrics Commons](#), [Finance Commons](#), [Longitudinal Data Analysis and Time Series Commons](#), and the [Statistical Models Commons](#)

Recommended Citation

Wu, Edward, "Application of Sentiment Analysis and Machine Learning Techniques to Predict Daily Cryptocurrency Price Returns" (2023). *CMC Senior Theses*. 3220.
https://scholarship.claremont.edu/cmc_theses/3220

This Open Access Senior Thesis is brought to you by Scholarship@Claremont. It has been accepted for inclusion in this collection by an authorized administrator. For more information, please contact scholarship@cuc.claremont.edu.

Claremont McKenna College

Application of Sentiment Analysis and Machine Learning
Techniques to Predict Daily Cryptocurrency Price Returns

Submitted to
Professor Michael Gelman

by
Edward Wu

for
Senior Thesis
Fall 2022
12/5/22

Acknowledgments

First off, I would like to express my gratitude to my thesis advisor, Professor Michael Gelman for his invaluable guidance and feedback during this semester-long journey. Without your support and patience, it would have been difficult for my thesis to come to fruition.

Additionally, I would like to thank Professor Mike Izbicki for providing me with access to historical data of all geolocated tweets dating back to 2017 as well as additional guidance with debugging my code.

Lastly, I would like to thank my family and friends who provided me with immense emotional support during this tough, grueling journey.

Abstract

This paper examines the effects of social media sentiment relating to Bitcoin on the daily price returns of Bitcoin and other popular cryptocurrencies by utilizing sentiment analysis and machine learning techniques to predict daily price returns. Many investors think that social media sentiment affects cryptocurrency prices. However, the results of this paper find that social media sentiment relating to Bitcoin does not add significant predictive value to forecasting daily price returns for each of the six cryptocurrencies used for analysis and that machine learning models that do not assume linearity between the current day price return and previous daily price returns combined with previous daily sentiment scores were more accurate than machine learning models that assume linearity.

Keywords: Cryptocurrency Price Prediction, Machine Learning, Sentiment Analysis, Twitter Sentiment

Table of Contents

I. Introduction	6
II. Literature Review	7
1. Sentiment Analysis	8
2. Machine Learning	10
III. Data and Methodologies	12
1. Data	12
2. Methodologies	15
IV. Empirical Results	20
1. Baseline Model Results	20
2. Experiment Model Results	23
V. Conclusion	26
VI. Bibliography	28
VII. Appendix	30
Appendix A: Additional Baseline Model Results	30
Appendix B: Additional Experiment Model Results	31
Appendix C: Baseline Model Results Figures	32
Appendix D: Experiment Model Results Figures	41

I. Introduction

Although there are various studies focused on predicting stock market price movements based on social media sentiment, research applying sentiment analysis and machine learning techniques to predict cryptocurrency price movements is comparatively sparse. This paper examines the common sentiment analysis and machine learning methods utilized in studies addressing the task of predicting stock prices and assesses whether these methods can be applied to forecasting cryptocurrency prices.

In comparison to the stock market, the cryptocurrency market has been historically more volatile and has a significantly smaller market capitalization. The cryptocurrency market appears to behave independently from other financial markets (Trabelsi 2018). Part of the reason behind the rise of the cryptocurrency market in the past few years is that it provides an alternative investment market for individuals with a higher risk tolerance. As a result, more and more investors are diversifying their investment portfolios to include popular cryptocurrencies such as Bitcoin, Ethereum, Tether, and Solana.

Within the past year, we have seen social media heavily affect the prices of both stocks and cryptocurrencies. For example, Elon Musk tweeted about Dogecoin being his favorite cryptocurrency sending the price of Dogecoin to jump 10% in the following days. Additionally, the Reddit community WallStreetBets pushed shares of Gamestop from \$20 to \$480 per share at its peak after a member identified and posted on Reddit that numerous hedge funds were trying to short the gaming merchandise retailer towards bankruptcy. Through these examples, we can observe that public sentiment and investor emotions may potentially have a large impact on stock and cryptocurrency prices. Furthermore, cryptocurrency prices will often move together in tandem meaning that when a particular cryptocurrency receives positive news and its price

surges, the other cryptocurrency prices will also increase. This paper uses various sentiment analysis and machine learning techniques to analyze whether social media sentiment relating to Bitcoin affects alternative cryptocurrency daily price returns. Hence, this thesis contributes to the fields of Quantitative Finance and Cryptocurrency Research by

1. Addressing whether social media and public opinion relating to Bitcoin affect Bitcoin and other popular cryptocurrency's prices
2. Incorporating sentiment analysis techniques such as VADER and machine learning models such as neural networks and regression analysis to forecast cryptocurrency prices
3. Bridging gaps in understanding the impact of social media and public opinion on cryptocurrency price movements

Section 2 of this paper reviews previous literature relating to the application of various sentiment analysis techniques and machine learning models to conduct cryptocurrency and stock price predictions. Section 3 details the collected data used for empirical analysis as well as the methodologies of specific sentiment analysis techniques and machine learning models that were used to generate prediction results. Section 4 shares the results from the machine learning models and addresses implications for the paper's area of interest. Section 5 summarizes the findings from the paper and proposes further areas of research.

II. Literature Review

Stock and cryptocurrency price prediction using sentiment analysis of social media and news articles has been a widely researched topic in the past few years. With main social media platforms such as Twitter and Reddit providing famed users such as Elon Musk and large digital communities such as WallStreetBets a space to openly share their opinions on cryptocurrencies

and stocks, we have seen prices of cryptocurrencies and stocks heavily fluctuate in response to their tweets and posts as highlighted in the examples mentioned in the introduction. Additionally, previous research such as R.J. Dolan's "Emotion, Cognition, and Behavior" (2002) and George Loewenstein's "Emotions in Economic Theory and Economic Behavior" (2000) have shown that decision-making involving economic and financial consequences is strongly affected by emotions. With concrete evidence and previous research highlighting that public sentiment can heavily affect cryptocurrency and stock prices, tools such as sentiment analysis can be more useful than purely economic fundamentals in predicting prices.

The literature review will be split into two components, sentiment analysis and machine learning prediction, and will conclude with the integration of both components jointly. The first part of the literature review will be an evaluation of common techniques used to perform sentiment analysis. The second part will focus on frequent machine learning techniques used to predict cryptocurrency and stock prices, and the last part will address previous research that has incorporated both sentiment analysis and machine learning techniques to predict cryptocurrency and stock prices.

1. Sentiment Analysis

Previous research addressing sentiment analysis of financial news and social media have often used two methods of analyzing sentiment: lexicon-based methods and transformers. VADER (Valence Aware Dictionary for Sentiment Reasoning), a lexicon-based method, is optimized for social media data and has been used by many researchers who analyze the sentiment of Twitter and other financial social media data. C.J. Hutto and Eric Gilbert (2014) developed VADER, which is a simple rule-based model for general sentiment analysis and compared its effectiveness to 11 typical state-of-the-practice benchmarks, including Affective

Norms for English Words (ANEW), Linguistic Inquiry and Word Count (LIWC), the General Inquirer, SentiWordNet, and machine learning-oriented techniques that rely on the Naive Bayes, Maximum Entropy, and Support Vector Machine algorithms. The study describes the development, validation, and performance evaluation of VADER. The researchers use a combination of qualitative and quantitative methods to construct a list of lexical features that are commonly used in social media. VADER utilizes a parsimonious rule-based model to assess the sentiment of tweets. Sohangir, Petty, and Wang (2018) have done a comparison between VADER and machine learning methods such as Naive Bayes and Support Vector Machine on financial social media. The results from their study showed that VADER was effective and efficient in computing the sentiment of financial social media and outperformed machine learning methods and other sentiment analysis techniques such as SentiWordNet, Textblob, Logistic Regression, and Linear Support Vector Machine.

Studies using transformers to output the sentiment of financial social media data have also shown great results. A transformer is a deep-learning model that uses the mechanism of self-attention to weigh the significance of each part of the input data. Instead of employing a pre-built dictionary that contains lexical features, transformers utilize a Sequence-to-Sequence learning algorithm to predict the next sequence. Sonkiya, Bajpai, and Bansal (2021) performed a version of BERT, a pre-trained transformer model by Google for Natural Language Processing (NLP), to compute the sentiment of news and headlines for the company Apple Inc. A particular result of their study showed that BERT outperformed other machine learning methods such as Support Vector Machine and Long short-term memory (LSTM) neural network.

There have been studies within the past year that have done a comparison between lexicon-based methods and BERT. Catelli, Pelosi, and Esposito (2022) have done a comparative

study of both methods on Italian opinionated documents for e-commerce and opinion websites. The results of their study showed that lexicon-based methods worked better for datasets that are small and the available computational resources limited while BERT was more accurate for larger datasets with the availability of machines with high computational power. Nonetheless, both methods are proven to be efficient and effective in performing sentiment analysis on financial news and social media sentiment.

2. Machine Learning

After performing sentiment analysis on social media data and computing the sentiment scores for each piece of text, we will use machine learning techniques to predict cryptocurrency and stock prices. As this paper is focused on predicting prices over time, time-series models will be employed for prediction. The LSTM neural network, a type of recurrent neural network, is a commonly used method to predict cryptocurrency and stock prices. Traditionally, recurrent neural networks have suffered from the problem of gradient disappearance, but LSTM neural networks have mitigated this problem by learning what to keep and forget from the previous cell states and adding relevant information to the intermediate cell state. The LSTM neural network utilizes a new structure containing a cell, an input gate, an output gate, and a forget gate. These gates decide what information should be stored and when to allow reading, writing, and forgetting. Recent studies have found that LSTM neural networks are effective in analyzing time series data due to its capability of storing past information and embedding sentiment scores in the set of selected features.

Ko and Chang (2021) utilize BERT to perform sentiment analysis on news relating to specific stocks in the Taiwan Stock Exchange and an LSTM neural network to predict the corresponding stock price. They found that applying LSTM neural networks to forecast stock

prices with stock historical transaction information and text sentiments improved average root mean square error (RMSE) by 12.05% in comparison to just using stock historical transaction information. Furthermore, Sidogi et al. (2021) analyze the influence of financial news headline sentiment on the predictability of stock prices using Long Term Short Term Memory (LSTM). The results from their study suggest that the use of news headline sentiment features improved the predictive performance of LSTM neural networks in intraday stock price prediction by 17.34% in terms of RMSE.

Vo et al. (2019) use sentiment analysis of news articles relating to cryptocurrency to predict price fluctuations for the second largest cryptocurrency in terms of market capitalization: Ethereum. After computing the sentiment scores of each piece of text, an LSTM neural network is employed to predict the price of Ethereum from July 30, 2017 to October 5, 2018. They found that their model was able to achieve a 2.16% lower mean absolute normalized error (MANE) when historical prices were combined with sentiment analysis features as opposed to only using historical prices. This highlights that public sentiment is an important factor in cryptocurrency price prediction. Additionally, Valencia et al. (2019) apply sentiment analysis of Twitter tweets relating to Bitcoin, Ethereum, Ripple, and Litecoin to predict the prices of these corresponding cryptocurrencies. They compare the utilization of multilayer perceptron neural networks, support vector machines, and random forests to predict the price movement of these cryptocurrencies. Their results show that neural networks outperform the other models in predicting cryptocurrency prices using machine learning and sentiment analysis of Twitter data. More specifically, the multilayer perceptron model had a 20% higher accuracy than the support vector machine and the random forest regression when predicting the movement of Bitcoin price for the

next day and a 5% higher accuracy than the 2 other models when predicting the movement of Ethereum price for the next day using tweets relating to Bitcoin, Ethereum, Ripple, or Litecoin.

Lastly, Ider (2022) studies the contribution of investor sentiment from news articles, Twitter, and Reddit to predict the daily prices of Bitcoin and Ethereum. Bitcoin and Ethereum are scraped for each respective currency and have their own text datasets. The study then compares the prediction results of twenty-three models, eleven of which are regressors and twelve are classifiers, for each of the selected cryptocurrencies. The results indicate that sentiment adds value to the return prediction models in most applications of these machine learning models and that test accuracy is higher for regressors with sentiment.

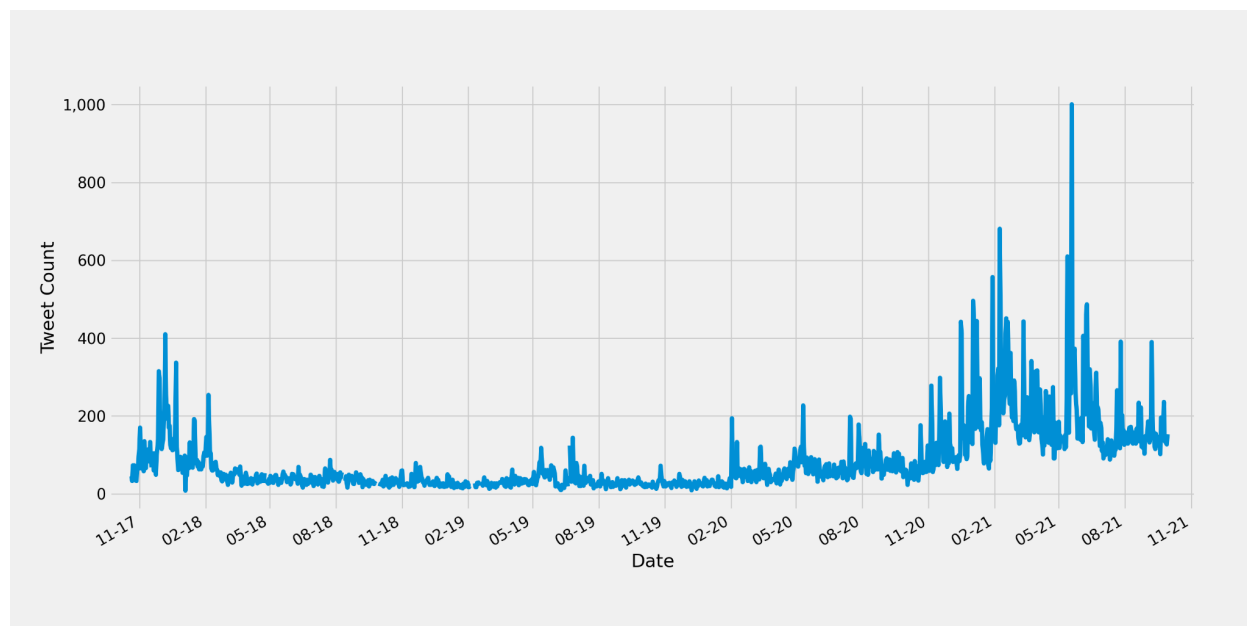
Although there have been various studies focused on examining whether sentiment relating to a particular cryptocurrency improves the price prediction results of the same cryptocurrency, research aimed towards addressing whether sentiment relating to mainstream cryptocurrencies improves the prediction results of alternative cryptocurrencies is comparatively scarce. This paper analyzes whether social media and public opinion on Bitcoin affect alternative cryptocurrency price movements. More specifically, this paper fills the gap in understanding if social media sentiment relating to Bitcoin adds predictive value to the prediction results of alternative cryptocurrencies' daily price returns.

III. Data and Methodologies

1. Data

This thesis mainly uses two datasets, a dataset containing Twitter data and a dataset containing cryptocurrency price metrics. For the Twitter data, tweets are collected in the time period spanning from October 21, 2017 to September 29, 2021. Since we perform our price

prediction and analysis on a daily frequency, the tweets are parsed by day using timestamp information. Twitter data was compiled through special access to the Twitter API Academic Research track with the help of Professor Mike Izbicki. Initially, collected and utilized by Izbicki, Papalexakis, and Tsotras (2019), the Twitter data contains all tweets with geolocation information going back to October 21, 2017. Due to the data consisting of 2.7 Terabytes of JSON files in over 100 languages, the dataset was filtered by specific hashtags and languages. The hashtags that were used to collect all tweets relating to Bitcoin were #Bitcoin and #BTC, case-insensitive. Furthermore, the dataset was restricted to only tweets in English in order to prepare for the sentiment analysis process. As illustrated in Figure 1, we can observe that there are fluctuations in the number of tweets gathered per day as the first 10,000 tweets were randomly sampled for each day and filtered for the relevant hashtags. As a result, a total of 116,344 tweets were parsed and relevant tables of metrics including a tweet's timestamp, text, hashtags and a user's name, follower count, and verification status were exported as comma-separated values (CSV) files. These CSV files were then imported to a remote server maintained by Professor Izbicki. The scope of this project is largely restricted due to three main factors: Twitter Dataset, API Limitation, and Limited Computing Power.

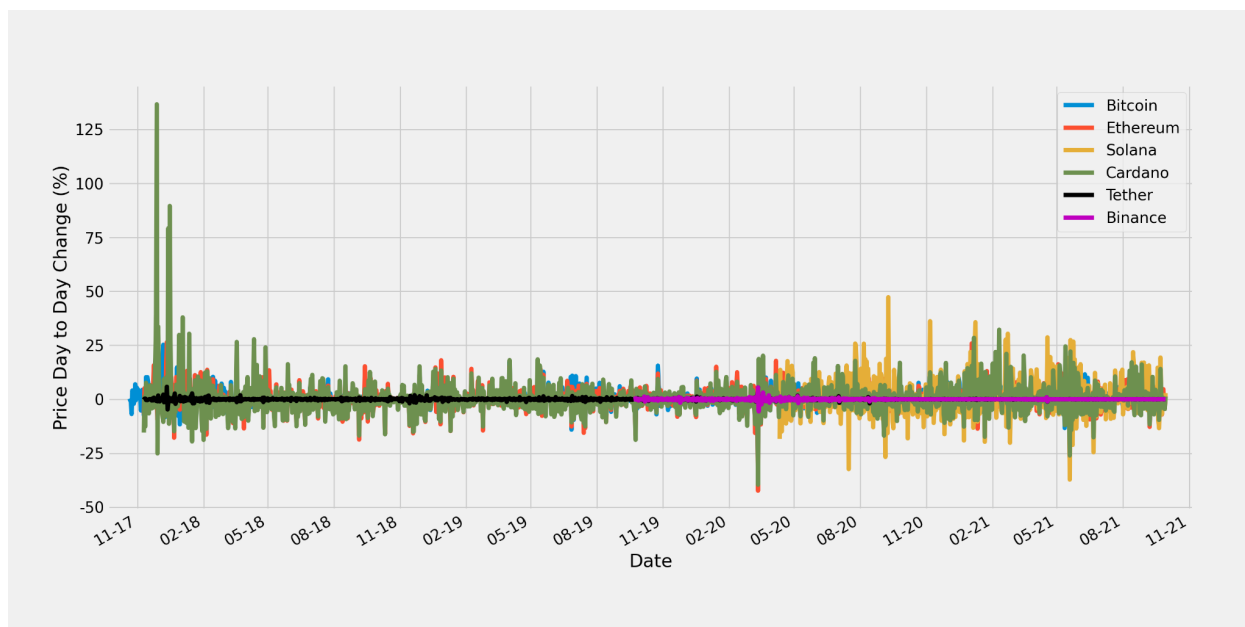
Figure 1: Total Tweet Counts Per Day

Source: Twitter API

For the cryptocurrency price dataset, market data was gathered for 6 different cryptocurrencies: Bitcoin (Ticker: BTC-USD), Ethereum (Ticker: ETH-USD), Tether (Ticker: USDT-USD), Cardano (Ticker: ADA-USD), Binance (Ticker: BUSD-USD), and Solana (Ticker: SOL-USD) for the same time period as the social media sentiment dataset. Even though longer time-series data for the majority of these cryptocurrency's prices were available, the timeframe of the data was reduced mainly due to Twitter API restrictions and call limitations. Solana and Binance were both launched and publicly traded after 2017, so the time period used for both cryptocurrencies' daily returns begins when the coin was first publicly available on the trading market. The cryptocurrency price data was obtained using Yahoo! Finance. Daily-adjusted time-series historical data from Yahoo! Finance includes the open, close, adjusted close, high, and low prices for a given day. Daily returns were calculated by subtracting the previous day's close price from the current day's close price and dividing this difference by the previous day's

close price. Figure 2 displays the daily returns for all six cryptocurrencies. From this figure, we can observe that Cardano and Solana have the largest fluctuations in daily returns while Tether and Binance have relatively stable returns and small changes in price.

Figure 2: Daily Returns of Cryptocurrency Tickers



Source: Yahoo! Finance

2. Methodologies

Training a deep-learning sentiment model and gathering sentiment for more than 100,000 tweets can be computationally expensive. VADER, a lexicon and rule-based sentiment analysis tool, was used to compute the sentiment scores of the tweets. In general, lexicon-based methods work better for projects where the available computational resources are limited. VADER utilizes a pre-defined dictionary of emotions rather than training a model using labeled data to assess the sentiment of a piece of text. The sentiment of each word is computed within a sentence and a comprehensive sentiment score is calculated for each sentence leading to an overall polarity score evaluated for a text. After running the vaderSentiment library through each tweet,

positivity scores, negativity scores, and neutrality scores were calculated for each tweet. Additionally, polarity scores, a weighted average of the positive, negative, and neutral scores normalized to -1 and 1, were also computed. If the polarity score of a tweet is greater than 0, the sentiment of the tweet would be labeled as positive. If the polarity score of a tweet is equal to 0, the sentiment of the tweet would be labeled as neutral, and if the polarity score of a tweet is less than 0, the sentiment of the tweet would be labeled as negative. As shown in Table 1, most tweets were classified as either neutral or positive with the vast majority of the positive tweets being mildly positive and having a polarity score close to 0. Table 2 shows an example of a positive tweet, a neutral tweet, and a negative tweet. After computing the individual polarity scores for each tweet, a weighted sentiment score was calculated for each day. Specifically, a proportion of the number of followers a user has to the total number of followers for every user on a particular day is multiplied by the corresponding sentiment score for the user tweet, which is then aggregated to derive a weighted sentiment score of a day. The equation is as follows:

$$\text{Weighted Sentiment Score}_{it} = \sum_{i=1}^N \frac{f_{it} * s_{it}}{f_t}$$

where i specifies an individual tweet, t specifies a particular day, N specifies the number of tweets in a day, f specifies the number of followers, and s specifies the sentiment score. The proportion of the number of followers a user has to the total number of followers for every user on a particular day eliminates scams and irrelevant advertisements and includes tweets by important users that are more frequently viewed by the general public. The underlying premise is that tweets with a larger viewing have a better representation of public sentiment relating to Bitcoin price changes.

Table 1: Summary of Sentiment Counts for Bitcoin

Cryptocurrency	Positive Sentiment	Neutral Sentiment	Negative Sentiment	Total
Bitcoin	43,630	57,941	14,773	116,344

Source: Twitter API

Table 2: Examples of Each Type of Sentiment

Positive Tweet	Totally agree, same thrill I get from buying Bitcoin.
Neutral Tweet	They teach you how to trade your Bitcoin till you become a pro.
Negative Tweet	Who's dumping their Bitcoin and buying GME? It's now only up 12% this month!!

Source: Twitter API

After computing average sentiment scores relating to tweets about Bitcoin for each day, machine learning methods were used to predict the daily returns of a particular cryptocurrency. The six cryptocurrencies used for prediction are Bitcoin, Ethereum, Tether, Binance, Solana, and Cardano. As mentioned previously this paper is focused on predicting cryptocurrency prices over time and all input variables are lagged by a single day. Three different machine learning models are utilized in this study. They are an Ordinary Least Squares Regression, a Random Forest Regression, and an LSTM neural network. All three models have been traditionally used for time series modeling and forecasting.

As explained earlier in the literature review, an LSTM neural network is a type of recurrent neural network that overcomes the problem of gradient disappearance by adding a forget gate from the previous cell states. A traditional recurrent neural network has a structure containing a cell, an input gate, and an output gate. The addition of the forget gate allows the model to decide what information to keep and throw away from the previous cell states. For example, consider the sentence, “I lived in Germany for fifteen years, and I can speak fluent German.” From this information, it is difficult to predict that the language is German directly. The LSTM neural network is able to weigh the words Germany and fifteen more heavily while throwing away less relevant words such as can and for. As a result, the LSTM can remember relevant information even if the gap between important information becomes very large.

A Random Forest Regression is a machine learning model that applies the ensemble learning method which uses multiple learning algorithms to attain better predictive performance in comparison to using a single learning algorithm. More specifically, a Random Forest Regression utilizes multiple random decision trees each individually trained on subsets of the dataset. The subsets of data are constructed through the application of bootstrapping which is a process of random sampling subsets of a dataset over a specified number of iterations and dimensions. After the result of each individual decision tree is computed, these results are averaged to compute a robust, accurate overall prediction. Lastly, the Ordinary Least Squares Regression is a linear regression model that attempts to minimize the sum of the squared differences between the observed dependent variable and the predicted dependent variable from the best-fit linear function derived from the independent variables. The Ordinary Least Squares Regression describes the relationship between one or more independent variables and a dependent variable. While the Ordinary Least Squares Regression assumes there is a linear

relationship between the independent variables and the dependent variable, the LSTM neural network and the Random Forest Regression do not assume linearity and are able to find non-linear relationships between these variables. As a result, we will be comparing the prediction results for each model and observing whether models that do not assume linearity between the independent variables and the dependent variable are more accurate than models that assume linearity.

Previous research on cryptocurrency price prediction has shown that closing prices of previous days for a particular cryptocurrency are a strong predictor of the current-day closing price. Therefore, we employ a baseline time-series model of only daily price return time lags for a specific cryptocurrency in the past 60 days. The experiment model uses both daily price return time lags and average sentiment score relating to Bitcoin time lags in the past 60 days. A time lag of 60 days was selected as it was the time lag that generated the best model accuracy. Table 3 highlights the variables used to conduct the study. This paper addresses the question of whether adding social media sentiment relating to Bitcoin as an input variable improves the accuracy of return prediction models for each cryptocurrency, and places less emphasis on optimizing the results of each model. Lastly, the data is rearranged into chronological order and split into 80% training data and 20% testing data due to a limited number of observations. The training data consists of the earliest dates in the overall dataset while the testing data consists of the latest dates in order to ensure that our model will train in sequential order.

Table 3: List of Variables in the Final Dataset

Category	Source/Method	Relevant Metrics
Historical Prices	Yahoo Finance	Date, Close (used to calculate daily price returns)
Sentiment Analysis	Twitter/VADER	Date, Tweet Sentiment Scores-Mean

IV. Empirical Results

The empirical results section will be divided into two parts, an evaluation of the baseline model results and an assessment of the experiment model results. The first section of the empirical results displays the root mean squared error and the mean absolute error (MAE) for each of the three baseline machine learning models and a corresponding cryptocurrency. The second section details the accuracy metrics for each of the three experiment machine learning models and compares the results between the baseline and experiment model. In addition, both sections will compare the accuracy metrics between each machine learning model to identify the best-performing models for a particular cryptocurrency daily price return prediction. This section will focus on the prediction results of three of the six cryptocurrencies including a mainstream cryptocurrency, an altcoin, and a stablecoin. The prediction results of the other cryptocurrencies will be placed in the appendix.

1. Baseline Model Results

Overall, the prediction results of each model for Bitcoin without adding the effects of sentiment were not accurate with each model being on average 4.5 to 4.7 percent off based on the RMSE and 3.3 to 3.5 percent off based on the MAE. This result is expected as cryptocurrencies

tend to be more volatile than general stocks making it difficult for all three models to capture fluctuations in cryptocurrency prices. From Table 4, we can see that the Random Forest Regression yielded the lowest RMSE and MAE while the LSTM neural network had the highest RMSE and MAE. Furthermore, we can observe from Figure C1 that all three models tended to underfit the data and failed to correctly capture accurate trends in cryptocurrency daily price returns on the test dataset.

Compared to the prediction results for Bitcoin, all three models were less accurate in predicting the daily price returns for Solana, an alternative cryptocurrency to Bitcoin and Ethereum. From Table 5, we can see that the LSTM neural network yielded the lowest RMSE and MAE while the Ordinary Least Squares Regression had the highest RMSE and MAE for Solana. In general, the machine learning models that do not make the assumptions of a linear model were more accurate than the linear regression model. Nevertheless, the three models carried on the recurring theme of underfitting the data and failing to capture the volatility and rapid spikes for each cryptocurrency's daily price returns.

In comparison to the prediction results for Bitcoin and Solana, all three models were significantly more accurate in predicting the daily price returns for Binance, a stablecoin. This result is due to Binance experiencing considerably less volatility and price fluctuations compared to the mainstream and alternative cryptocurrencies as observed in Figure 2. From Table 6, we can see that both the Ordinary Least Squares Regression and the Random Forest Regression had a lower RMSE and MAE than the LSTM neural network for Binance. Even though all 3 models were more accurate in predicting the daily price returns for both cryptocurrencies, the models continued to underfit the data.

Table 4: Bitcoin

Model	RMSE	MAE
LSTM	4.679	3.522
OLS	4.466	3.355
RFR	4.458	3.345

Table 5: Solana

Model	RMSE	MAE
LSTM	7.770	6.168
OLS	8.593	6.820
RFR	8.209	6.328

Table 6: Binance

Model	RMSE	MAE
LSTM	0.086	0.078
OLS	0.043	0.031
RFR	0.041	0.031

2. Experiment Model Results

After adding the average sentiment scores relating to Bitcoin for each day to the daily price returns, the prediction results of each model for Bitcoin were relatively similar to the baseline model results. From Table 7, we can see that the Random Forest Regression yielded the lowest RMSE and MAE while the Ordinary Least Squares Regression had the highest RMSE and MAE. Both the LSTM neural network and Random Forest Regression had slight improvements in model accuracy after adding sentiment scores according to the RMSE as observed in Table 7 and Table 4. On the other hand, the Ordinary Least Squares Regression saw an increase in RMSE. Overall, the improvements in accuracy for the LSTM neural network and Random Forest Regression were marginal, meaning that social media sentiment relating to Bitcoin does not add much predictive value to Bitcoin's price compared to previous price trends.

Comparing the prediction results for Solana in Table 8 and Table 5, adding the average sentiment scores relating to Bitcoin for each day to the daily price returns significantly increased model accuracy for the LSTM neural network and the Random Forest Regression. More specifically, both models had an increase in accuracy of 0.4 percentage points. This result signals that social media sentiment relating to Bitcoin potentially affects the price of altcoins. Additionally, the LSTM neural network yielded the lowest RMSE and MAE while the Ordinary Least Squares Regression had the highest RMSE and MAE. Similar to the prediction results for Bitcoin, the Ordinary Least Squares Regression was less accurate after factoring in sentiment signaling that sentiment relating to Bitcoin does not correlate well with daily price returns for both Bitcoin and Solana.

In general, the prediction results for Binance, Table 9 and Table 6, had an increase in RMSE for the majority of the models after adding sentiment scores relating to Bitcoin meaning

that social media sentiment relating to Bitcoin does not add predictive value to Binance's price compared to previous price trends. Furthermore, the Random Forest Regression yielded the lowest RMSE and MAE while the Ordinary Least Squares Regression had the highest RMSE and MAE. Overall, Solana had significant improvements in model accuracy after factoring in social media sentiment relating to Bitcoin while Bitcoin saw minor improvements. Binance saw increases in RMSE after adding sentiment relating to Bitcoin suggesting that sentiment simply increases noise in the model and does not correlate with previous price trends. Additionally, the machine learning models that do not assume linearity between the independent variables and the dependent variable performed significantly better in the experiment model while all three machine learning models had relatively similar performances in the baseline model. This outcome signifies that adding sentiment scores to the daily price returns increases noise in the training data which makes it more difficult for a linear model to capture the relationship between the independent and dependent variables.

Table 7: Bitcoin

Model	RMSE	MAE
LSTM	4.503	3.410
OLS	4.568	3.463
RFR	4.441	3.328

Table 8: Solana

Model	RMSE	MAE
LSTM	7.411	5.680
OLS	9.777	7.873
RFR	7.816	5.946

Table 9: Binance

Model	RMSE	MAE
LSTM	0.065	0.053
OLS	0.200	0.159
RFR	0.046	0.036

V. Conclusion

After comparing the results between the baseline model and the experiment model for the three machine learning models (LSTM Neural Network, Ordinary Least Squares Regression, and Random Forest Regression) that were implemented, we can conclude that social media (Twitter) sentiment relating to Bitcoin does not add significant predictive value to forecasting daily price returns for each of the six cryptocurrencies analyzed in this study (Bitcoin, Ethereum, Solana, Cardano, Binance, and Tether). More specifically, the RMSE and MAE for both the baseline model and the experiment model were relatively similar and in a few cases, RMSE was worse for the experiment model. Additionally, the machine learning models that do not assume linearity between the current day price return and previous daily price returns combined with previous daily sentiment scores were more accurate than the model that assumes linearity. Therefore, future research within the cryptocurrency price prediction space should focus on optimizing the performance of nonlinear models.

Although this paper finds that social media sentiment relating to Bitcoin does not improve the prediction of daily price returns of other popular cryptocurrencies, there are further issues and areas to be addressed in future research. Due to the limited computing power and access to disk storage space, the Twitter dataset used in this study was relatively small. Increasing the number of observations in the dataset could potentially capture a more accurate representation of sentiment for a particular day and thus improve the testing error of our machine learning models. Furthermore, there are other methods of performing sentiment analysis such as employing a FinBERT transformer or a Naive Bayes classifier that may also improve the quality of capturing sentiment. Lastly, other machine learning models such as XGBoost or LightGBM can be used to improve the accuracy of predicting daily cryptocurrency price returns as both of

these models utilize ensemble learning boosting. Gradient boosting techniques employ multiple models with each successive model improving the accuracy of the previous model which could be useful when working with time series data.

VI. Bibliography

- Catelli, Rosario, Serena Pelosi, and Massimo Esposito. "Lexicon-Based vs. Bert-Based Sentiment Analysis: A Comparative Study in Italian." *Electronics* 11, no. 3 (2022): 374. <https://doi.org/10.3390/electronics11030374>.
- Dolan RJ. Emotion, cognition, and behavior. *Science*. 2002 Nov 8;298(5596):1191-4. doi: 10.1126/science.1076358. PMID: 12424363.
- Hutto, C., & Gilbert, E. (2014). VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1), 216-225. Retrieved from <https://ojs.aaai.org/index.php/ICWSM/article/view/14550>
- Ider, Duygu and Stefan Lessmann. "Forecasting Cryptocurrency Returns from Sentiment Signals: An Analysis of BERT Classifiers and Weak Supervision." (2022).
- Ko, Ching-Ru, and Hsien-Tsung Chang. "LSTM-Based Sentiment Analysis for Stock Price Forecast." *PeerJ Computer Science* 7 (2021). <https://doi.org/10.7717/peerj-cs.408>.
- Loewenstein, George. 2000. "Emotions in Economic Theory and Economic Behavior." *American Economic Review*, 90 (2): 426-432. DOI: 10.1257/aer.90.2.426
- Sidogi, Thendo, Rendani Mbuyha, and Tshilidzi Marwala. "Stock Price Prediction Using Sentiment Analysis." *2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2021. <https://doi.org/10.1109/smc52423.2021.9659283>.

Sohangir, Sahar, Nicholas Petty, and Dingding Wang. “Financial Sentiment Lexicon Analysis.”

2018 IEEE 12th International Conference on Semantic Computing (ICSC), 2018.

<https://doi.org/10.1109/icsc.2018.00052>.

Sonkiya, Priyank, Vikas Bajpai and Anukriti Bansal. “Stock price prediction using BERT and

GAN.” *ArXiv* abs/2107.09055 (2021): n. pag.

Trabelsi, Nader. “Are There Any Volatility Spill-over Effects among Cryptocurrencies and

Widely Traded Asset Classes?” *Journal of Risk and Financial Management* 11, no. 4

(2018): 66. <https://doi.org/10.3390/jrfm11040066>.

Valencia, Franco, Alfonso Gómez-Espinosa, and Benjamín Valdés-Aguirre. “Price Movement

Prediction of Cryptocurrencies Using Sentiment Analysis and Machine Learning.”

Entropy 21, no. 6 (2019): 589. <https://doi.org/10.3390/e21060589>.

Vo, Anh-Dung. “Sentiment Analysis of News for Effective Cryptocurrency Price Prediction.”

International Journal of Knowledge Engineering 5, no. 2 (2019): 47–52.

<https://doi.org/10.18178/ijke.2019.5.2.116>.

VII. Appendix

Appendix A: Additional Baseline Model Results

Table A1: Ethereum

Model	RMSE	MAE
LSTM	6.010	4.522
OLS	6.233	4.681
RFR	6.301	4.662

Table A2: Cardano

Model	RMSE	MAE
LSTM	7.629	5.367
OLS	7.648	5.512
RFR	7.581	5.382

Table A3: Tether

Model	RMSE	MAE
LSTM	0.184	0.156
OLS	0.098	0.053
RFR	0.101	0.053

Appendix B: Additional Experiment Model Results**Table B1: Ethereum**

Model	RMSE	MAE
LSTM	6.093	4.512
OLS	6.383	4.757
RFR	6.227	4.583

Table B2: Cardano

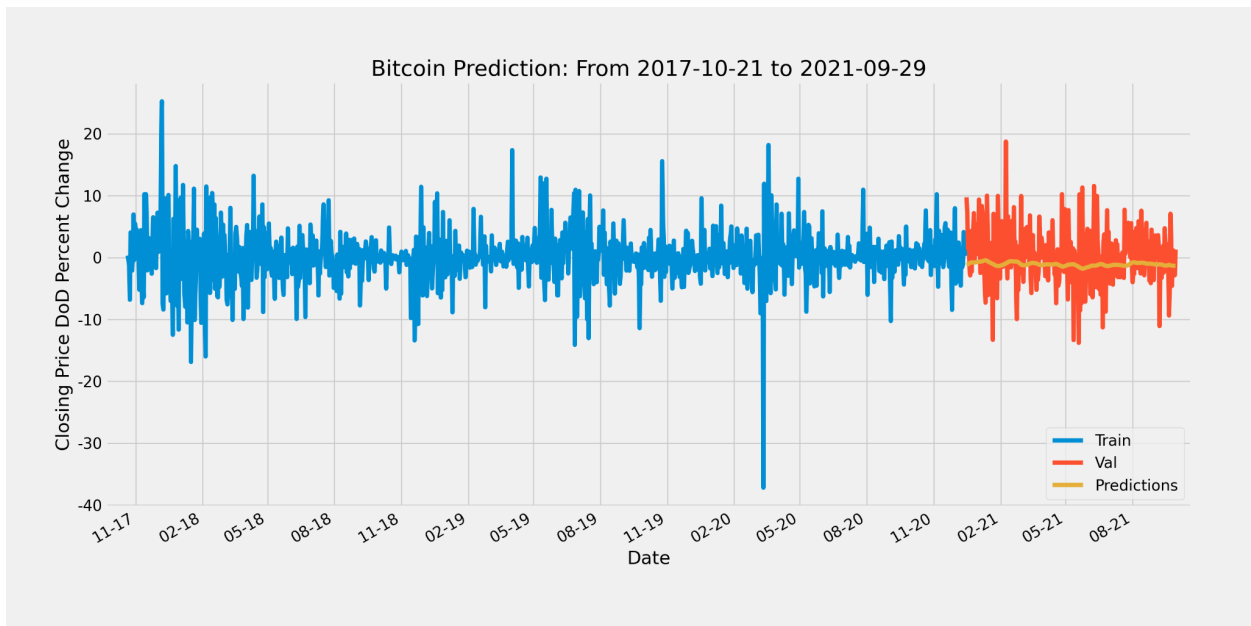
Model	RMSE	MAE
LSTM	7.506	5.310
OLS	7.649	5.579
RFR	7.558	5.358

Table B3: Tether

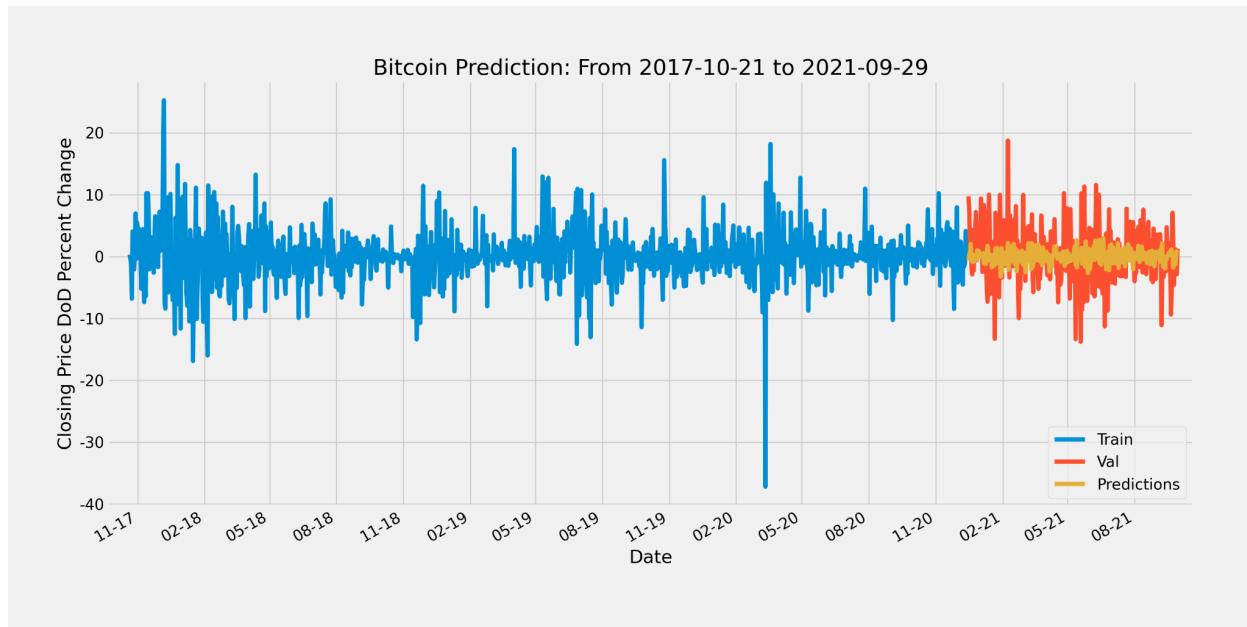
Model	RMSE	MAE
LSTM	0.227	0.206
OLS	0.136	0.100
RFR	0.104	0.057

Appendix C: Baseline Model Results Figures**Figure C1: Bitcoin**

LSTM



OLS



RFR

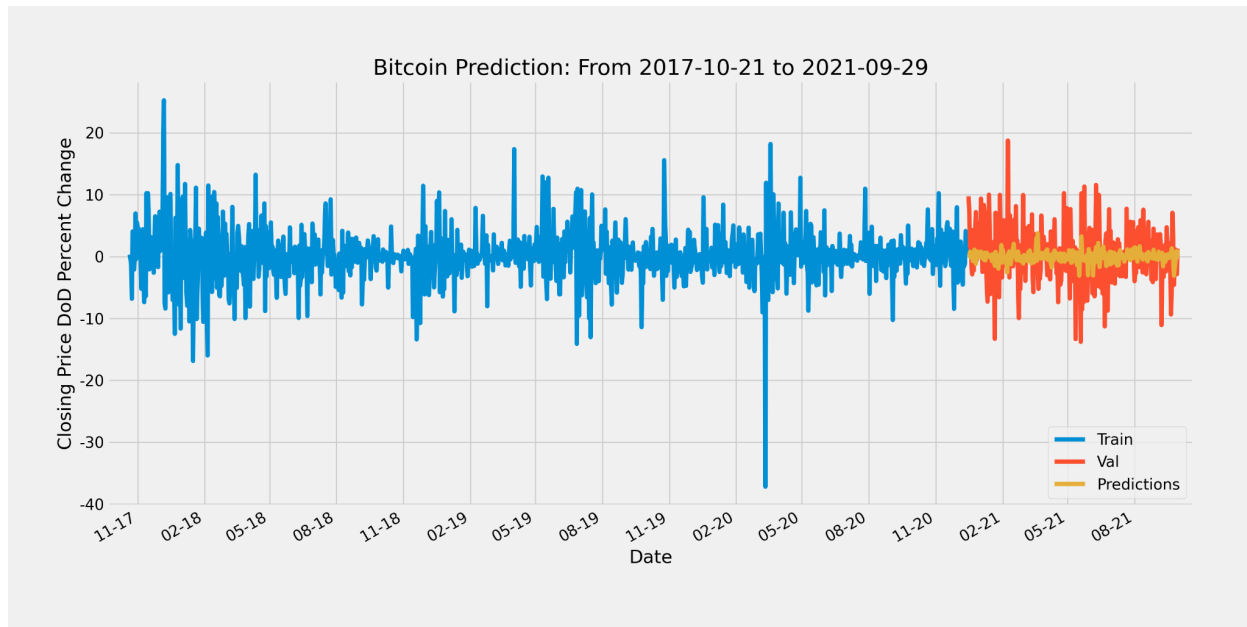
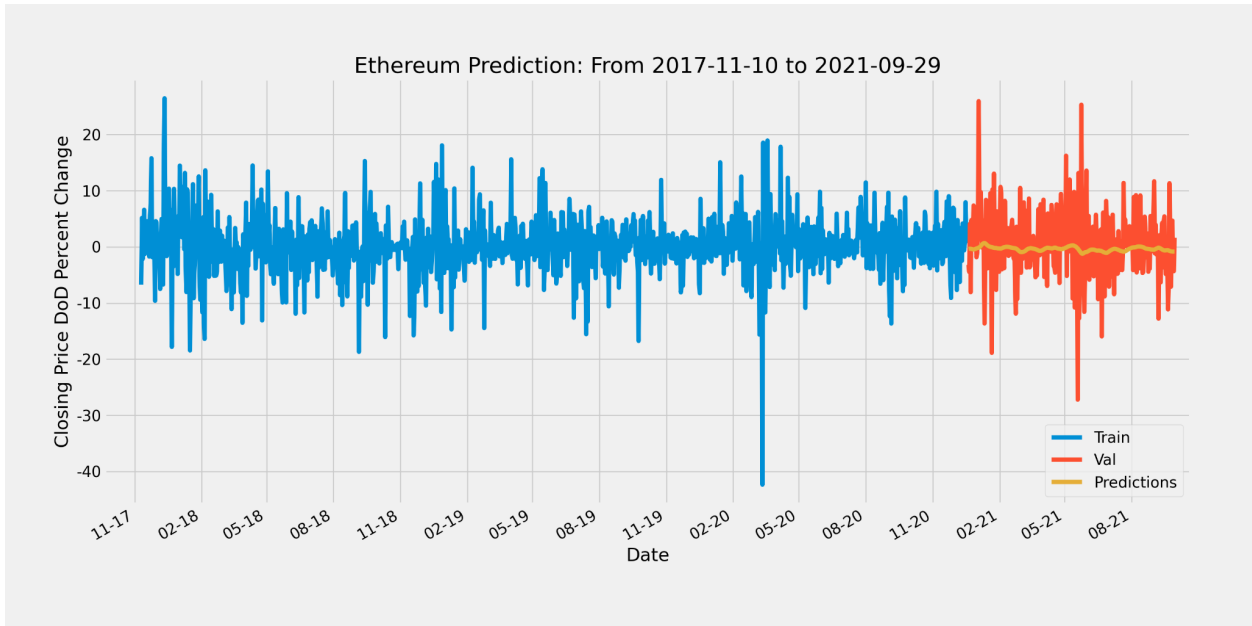
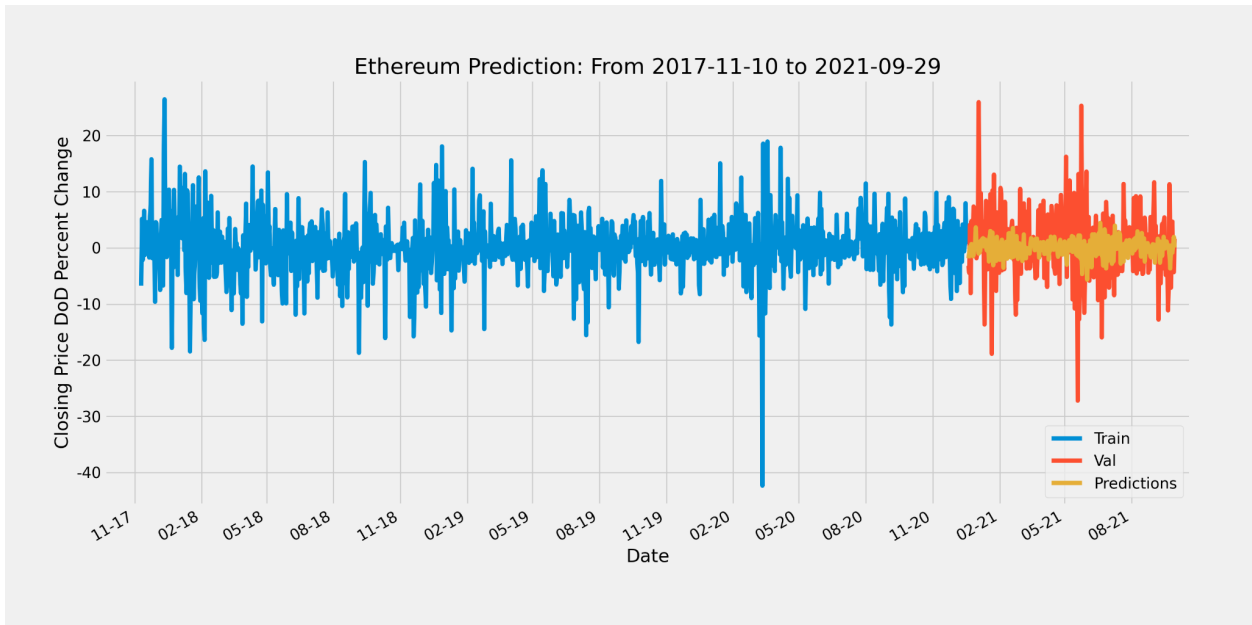


Figure C2: Ethereum

LSTM



OLS



RFR

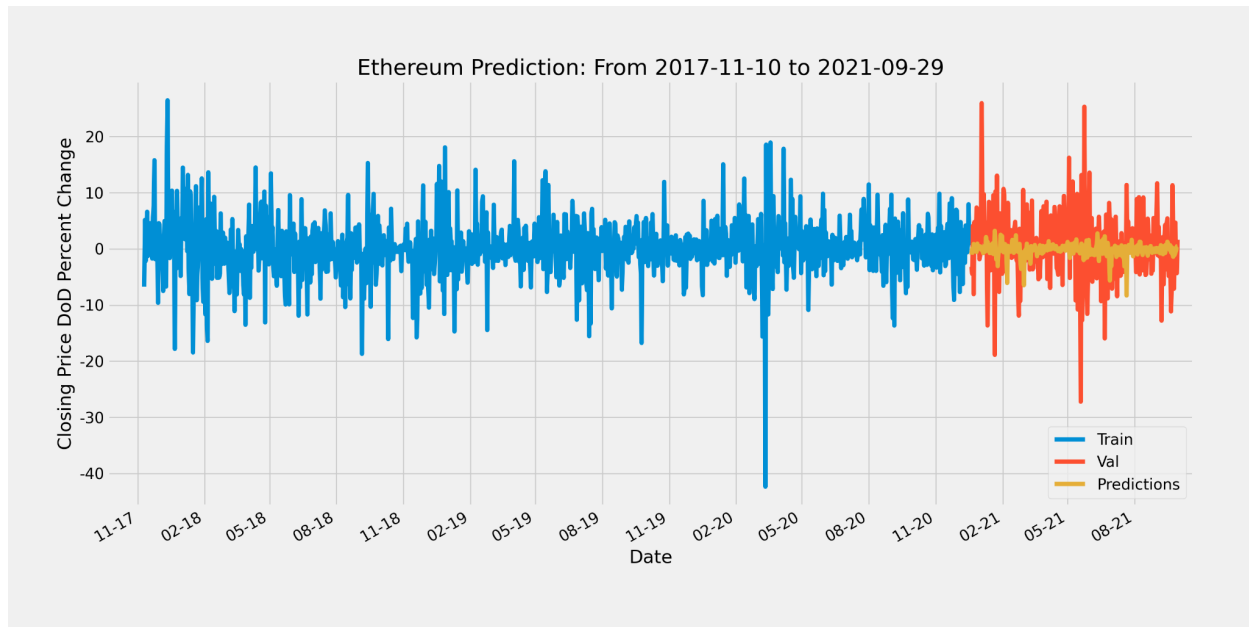
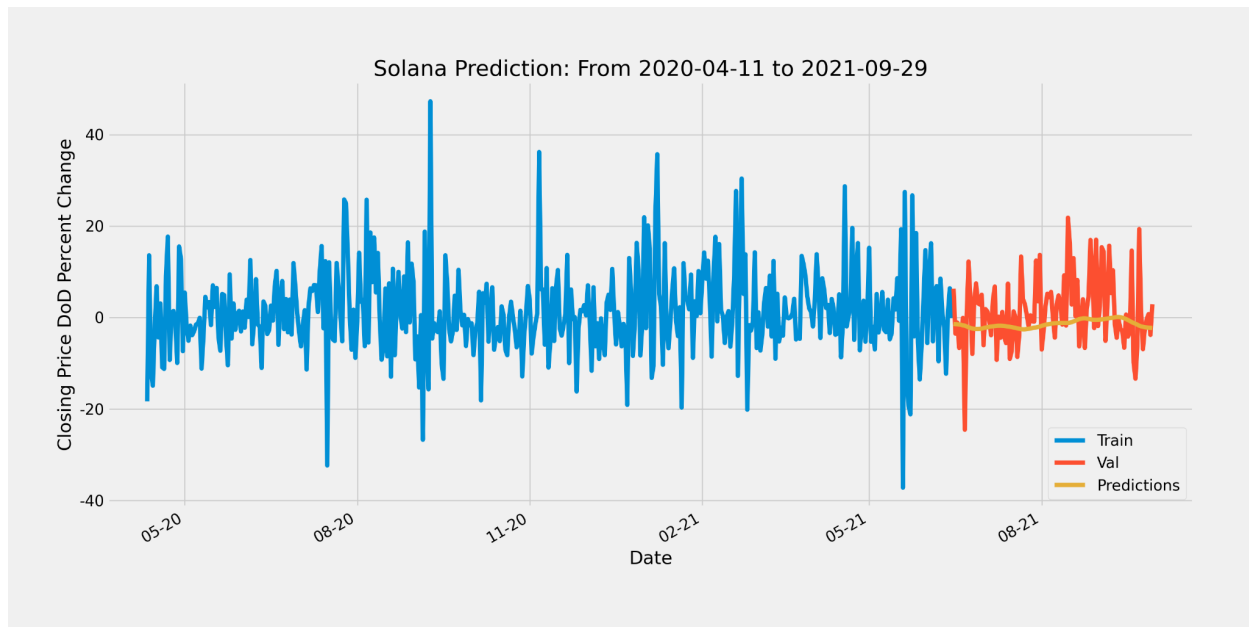
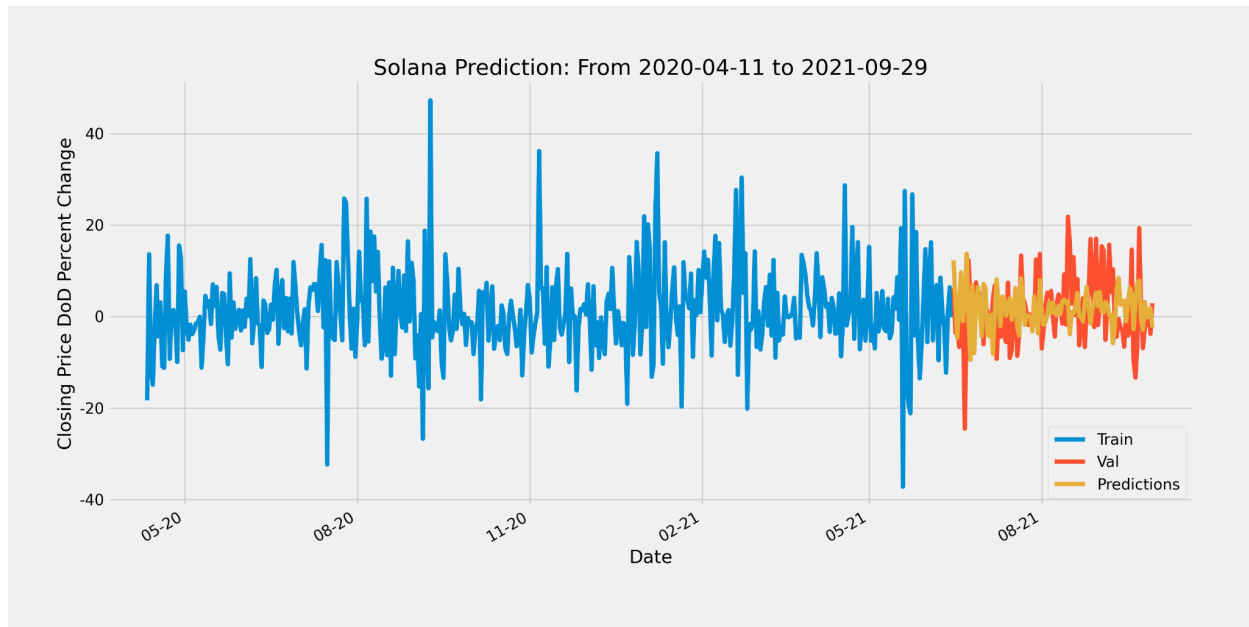


Figure C3: Solana

LSTM



OLS



RFR

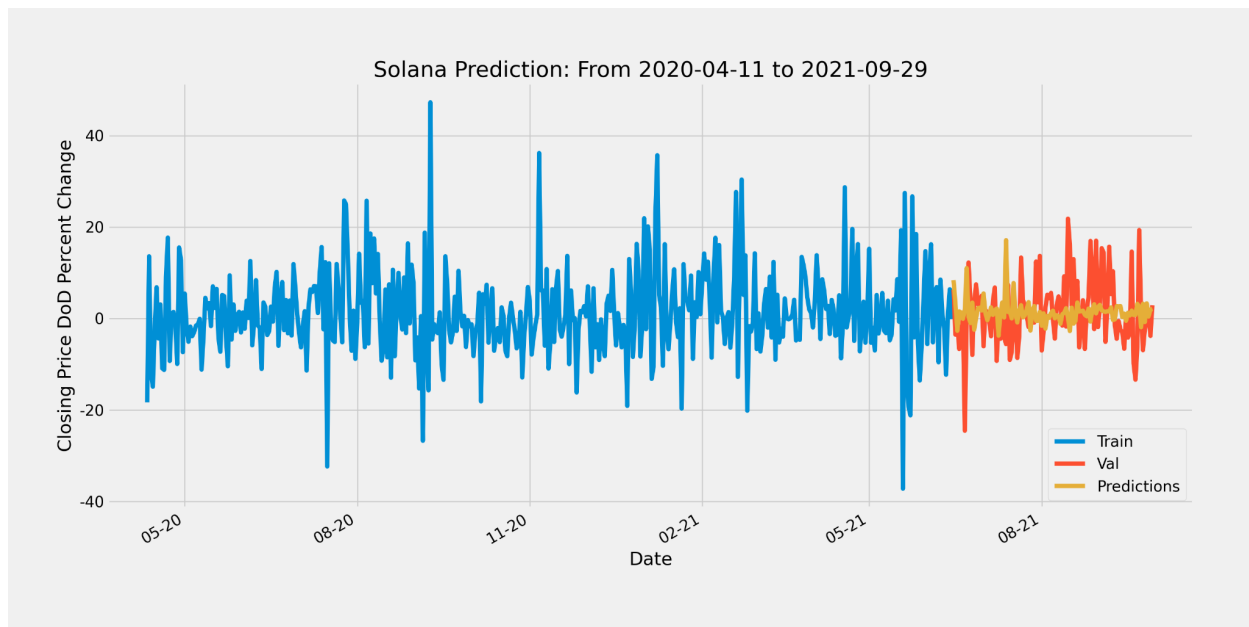
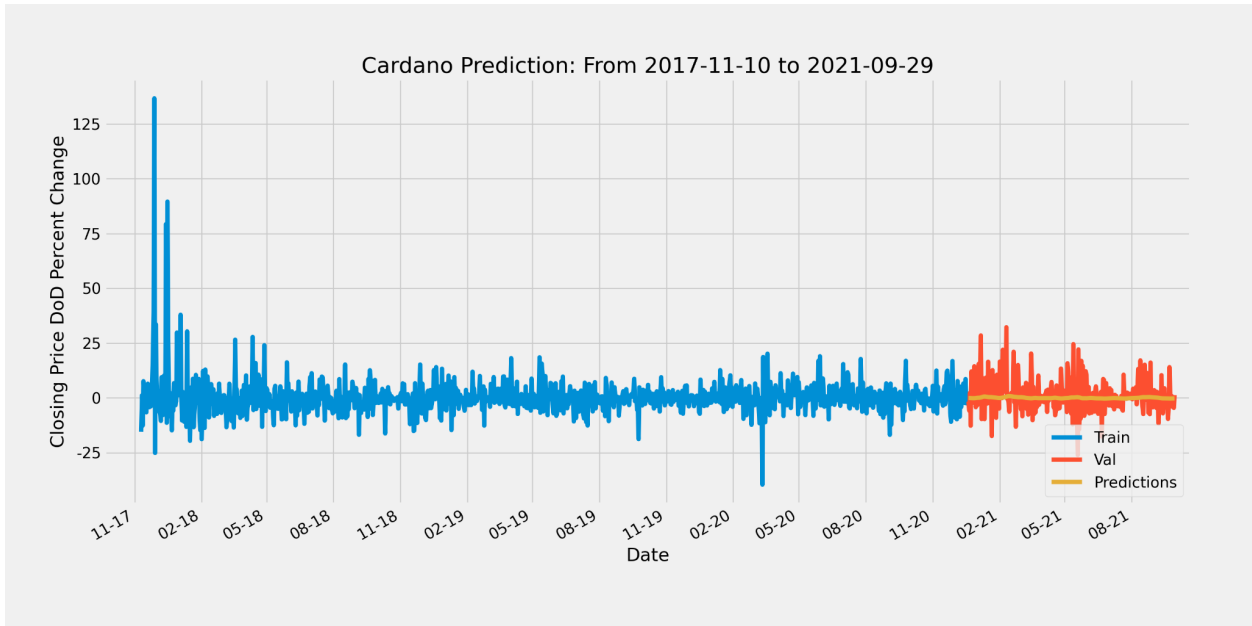
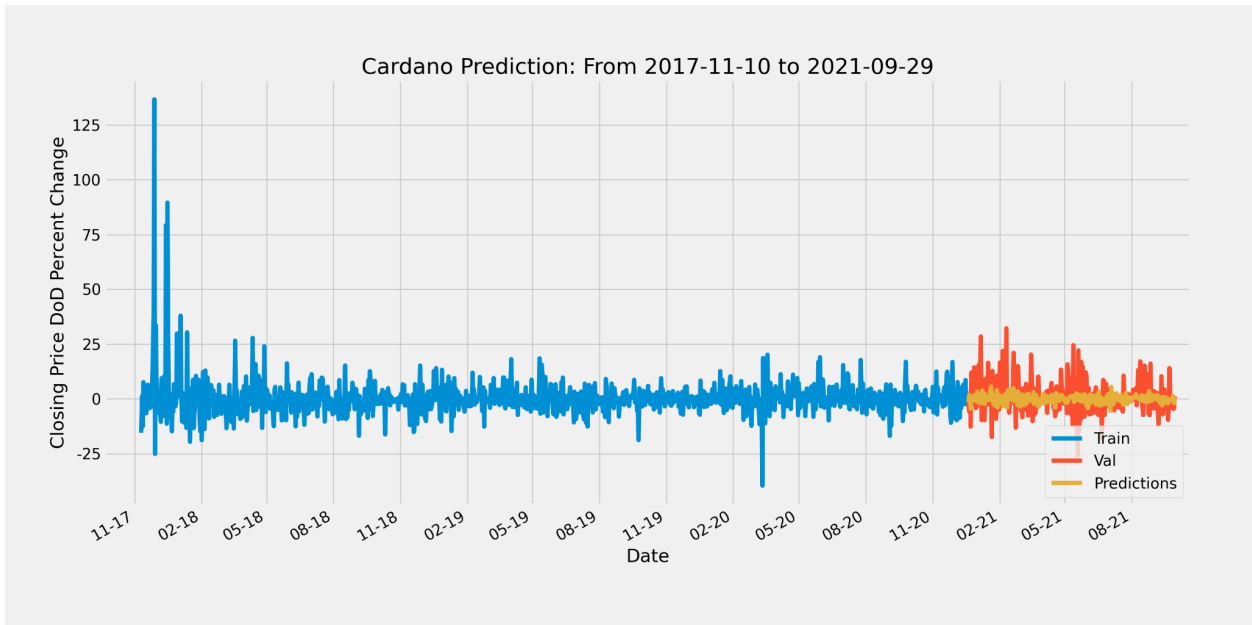


Figure C4: Cardano

LSTM



OLS



RFR

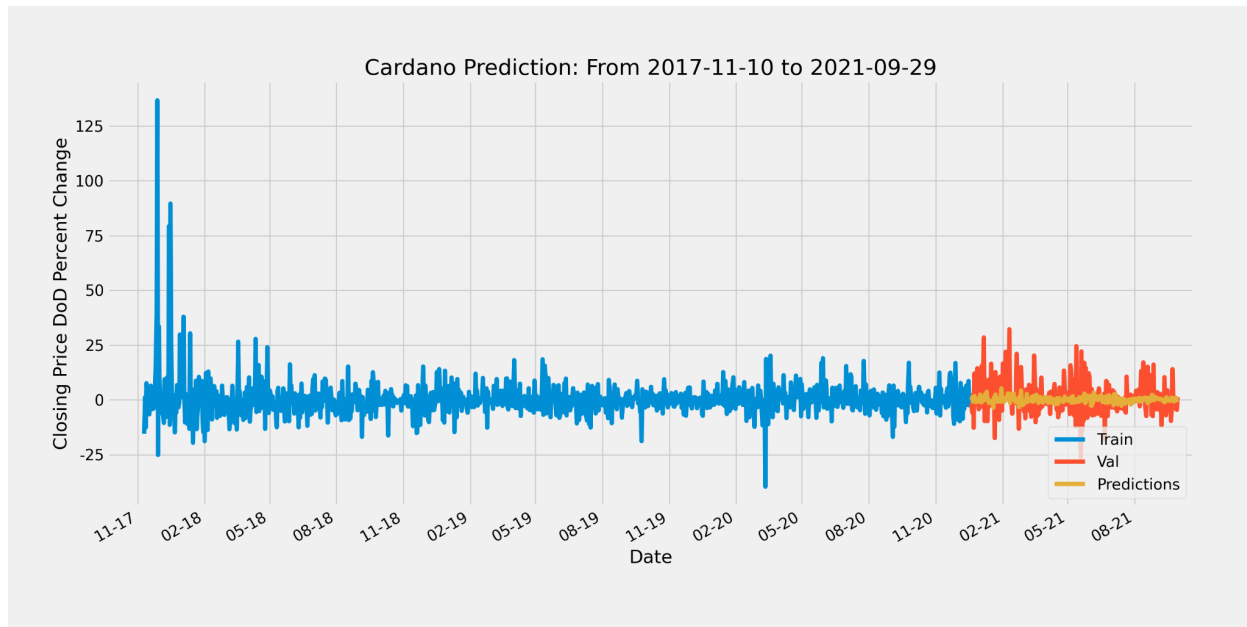
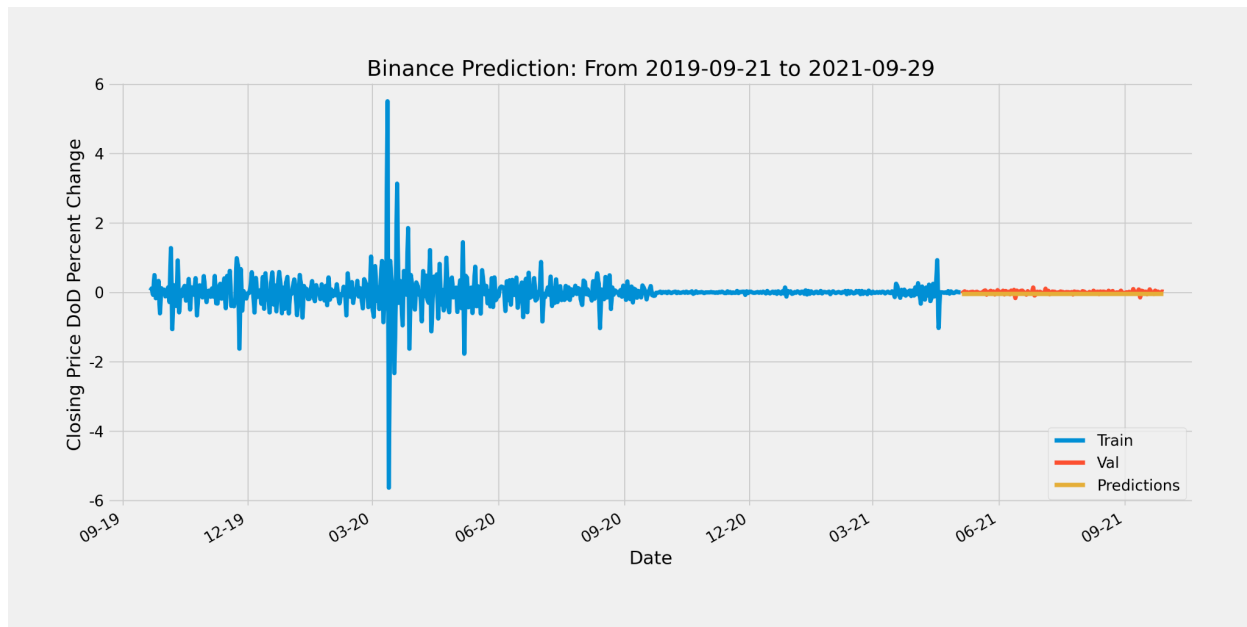
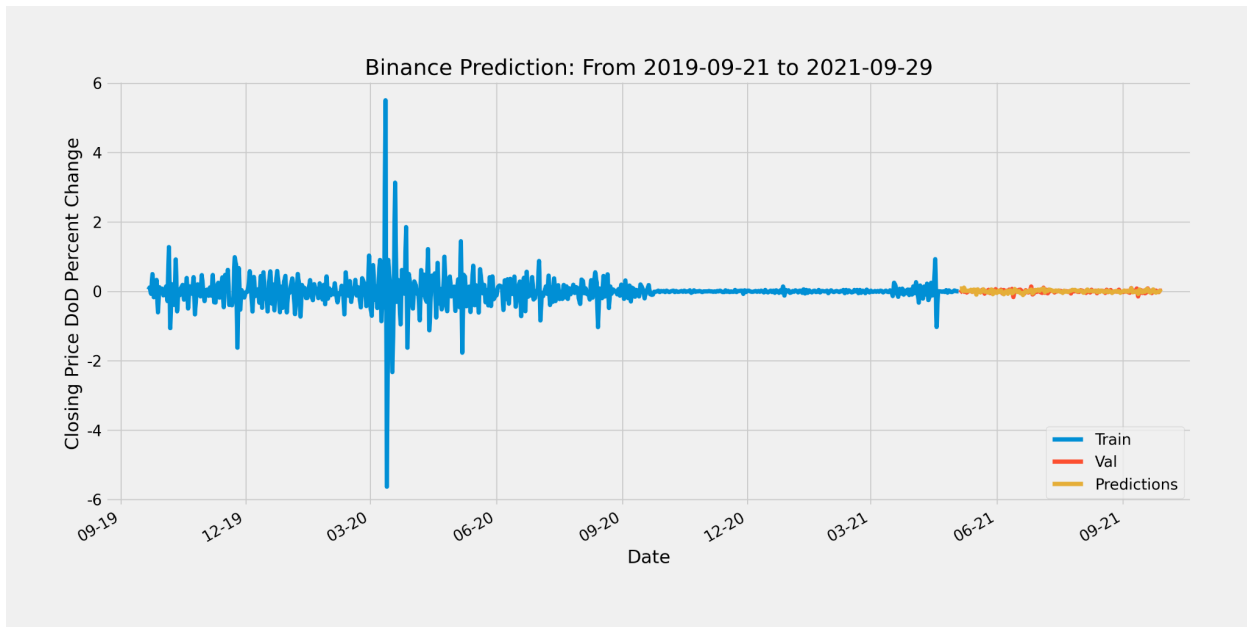


Figure C5: Binance

LSTM



OLS



RFR

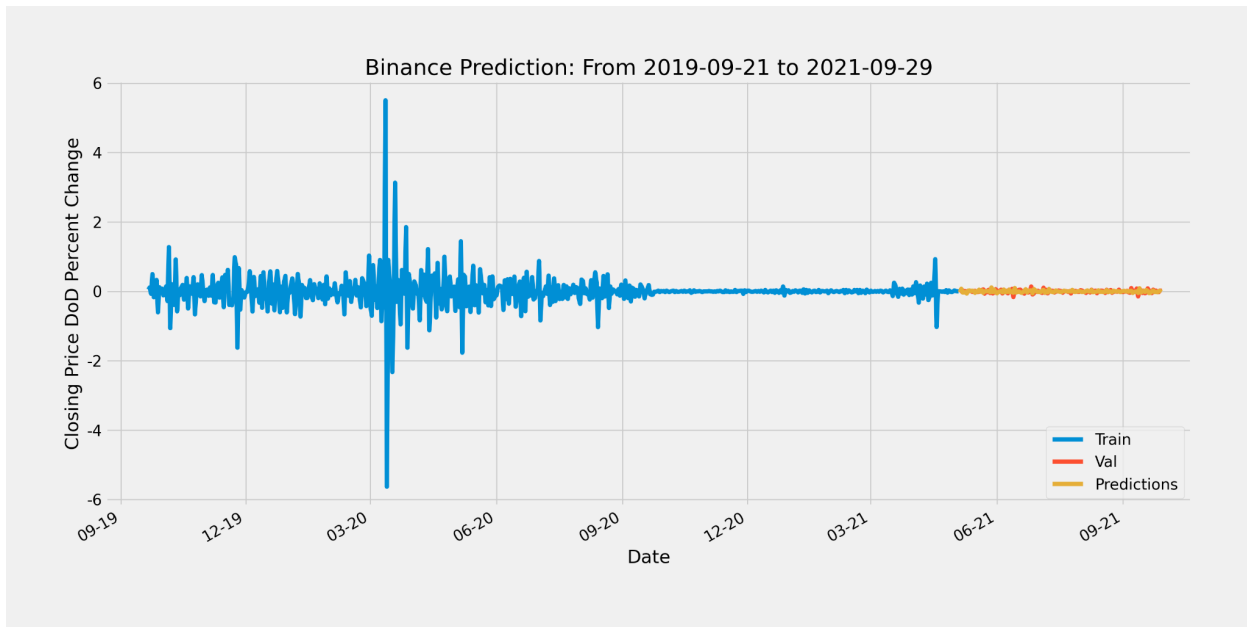
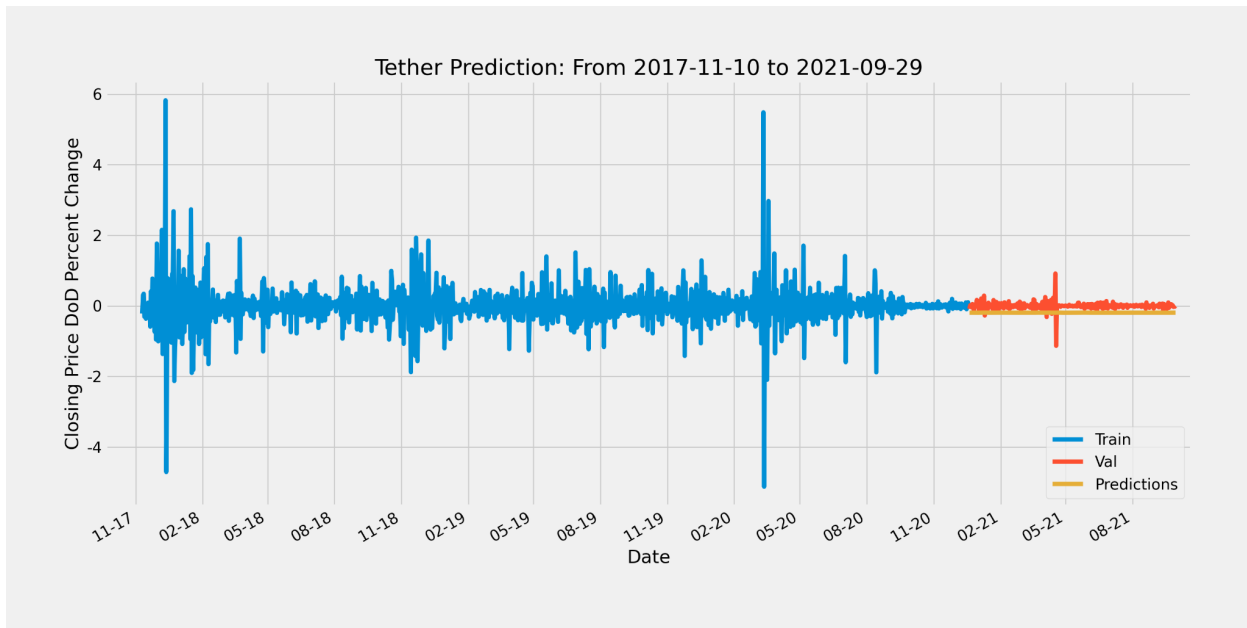
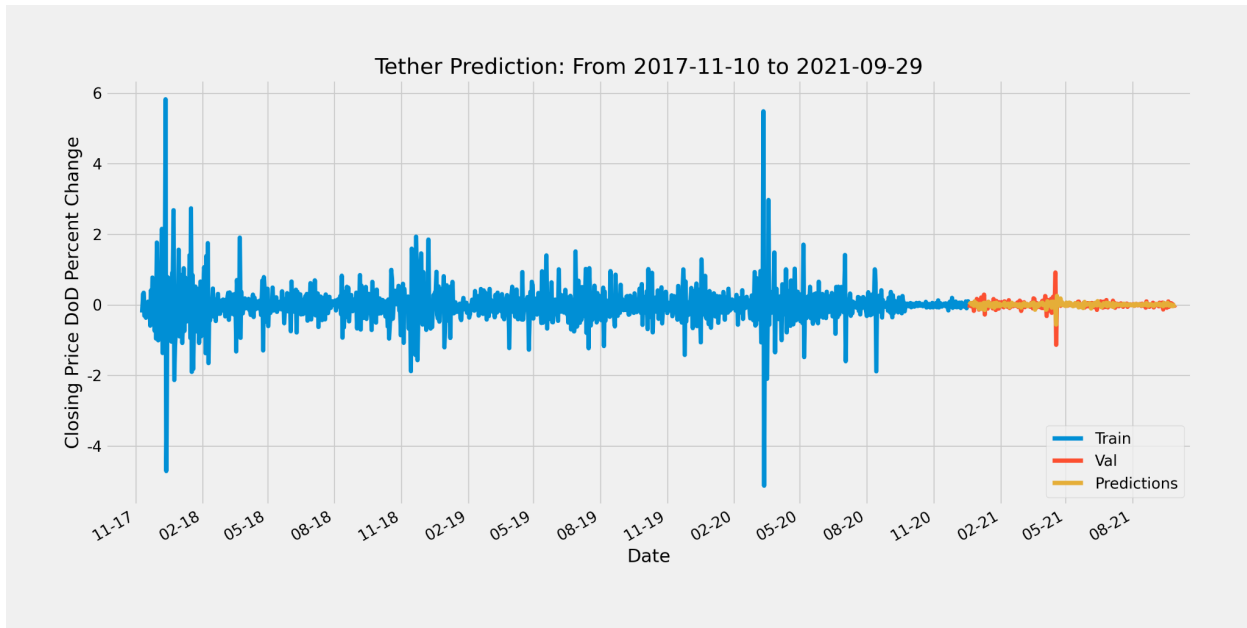


Figure C6: Tether

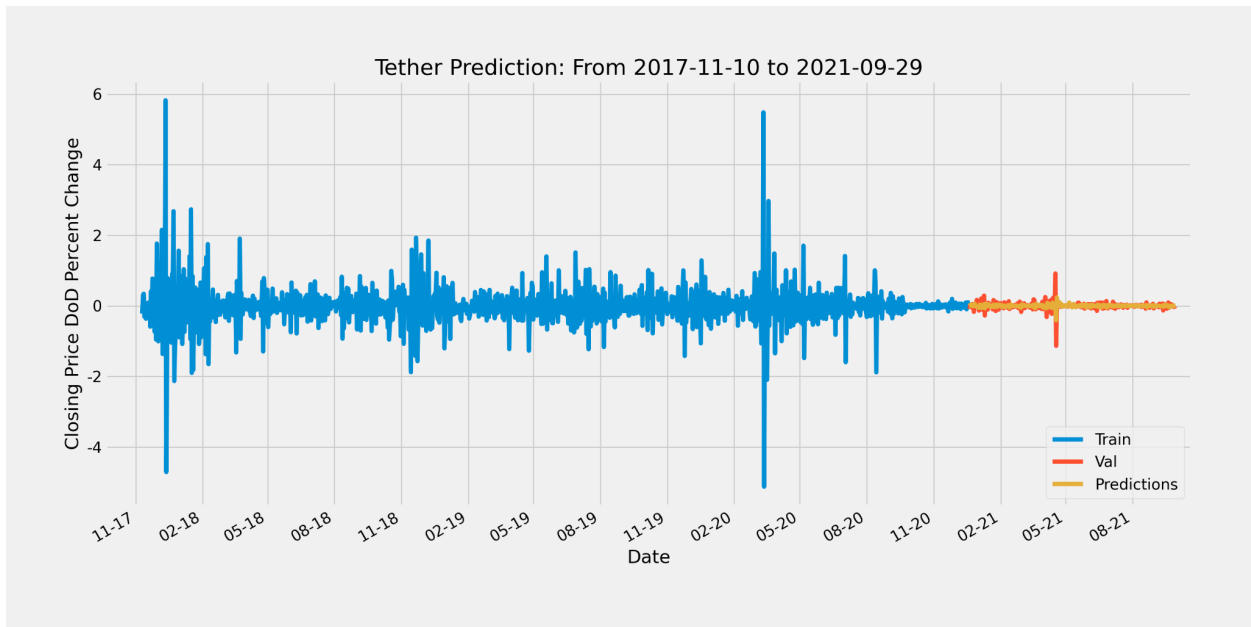
LSTM



OLS



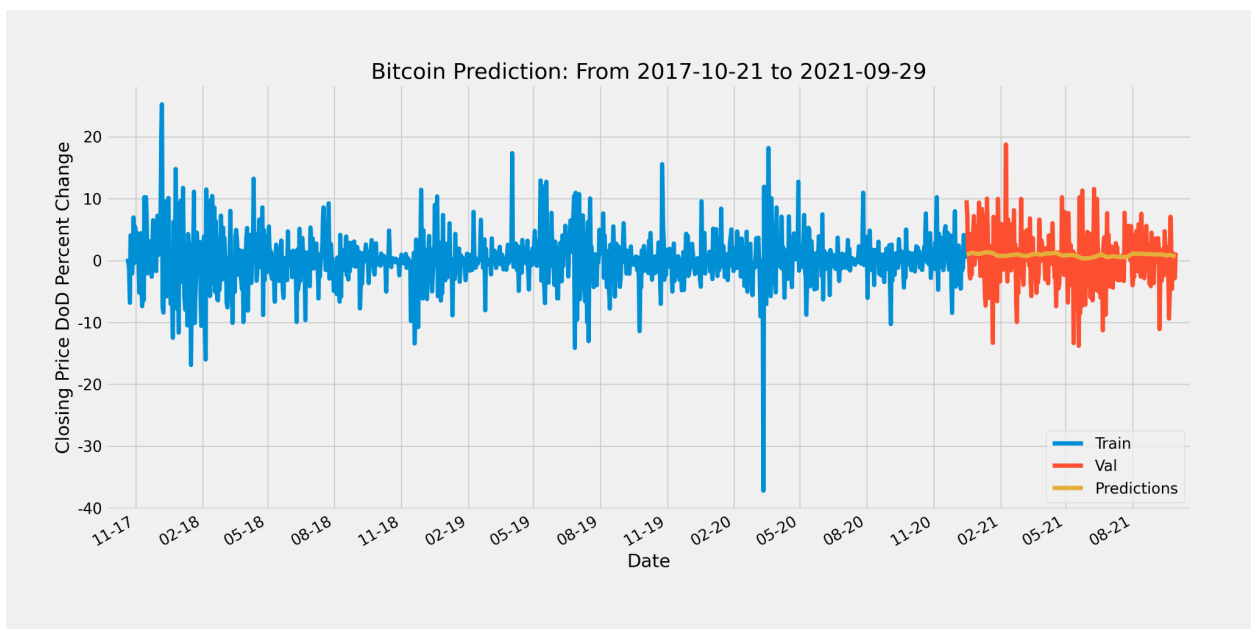
RFR



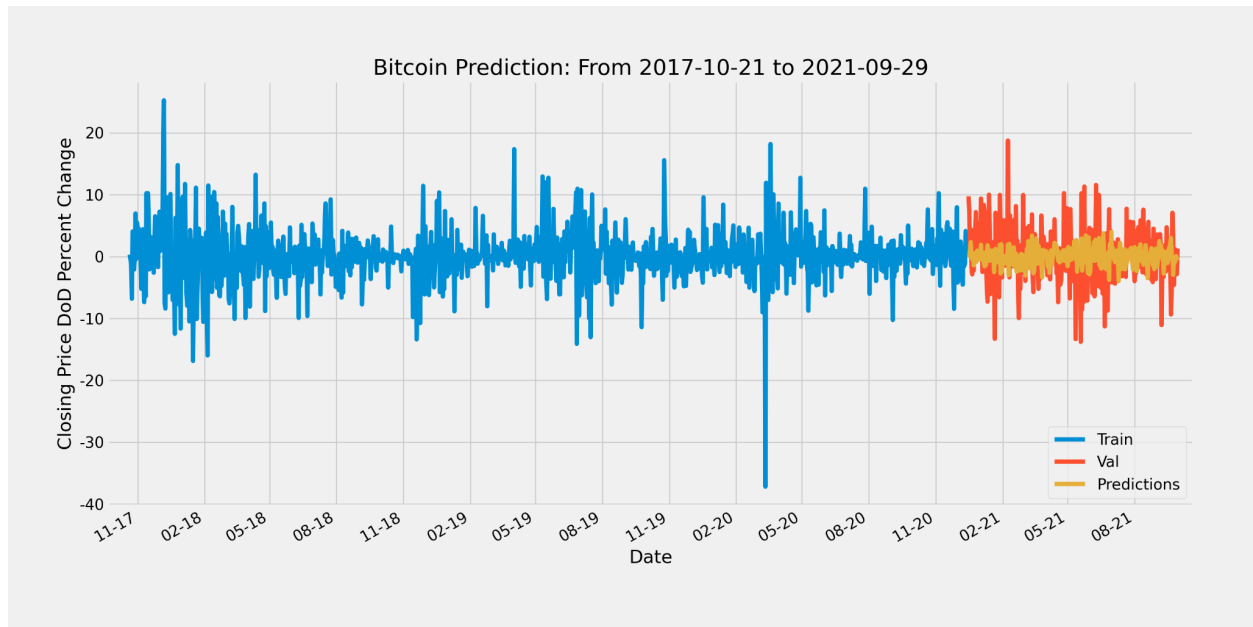
Appendix D: Experiment Model Results Figures

Figure D1: Bitcoin

LSTM



OLS



RFR

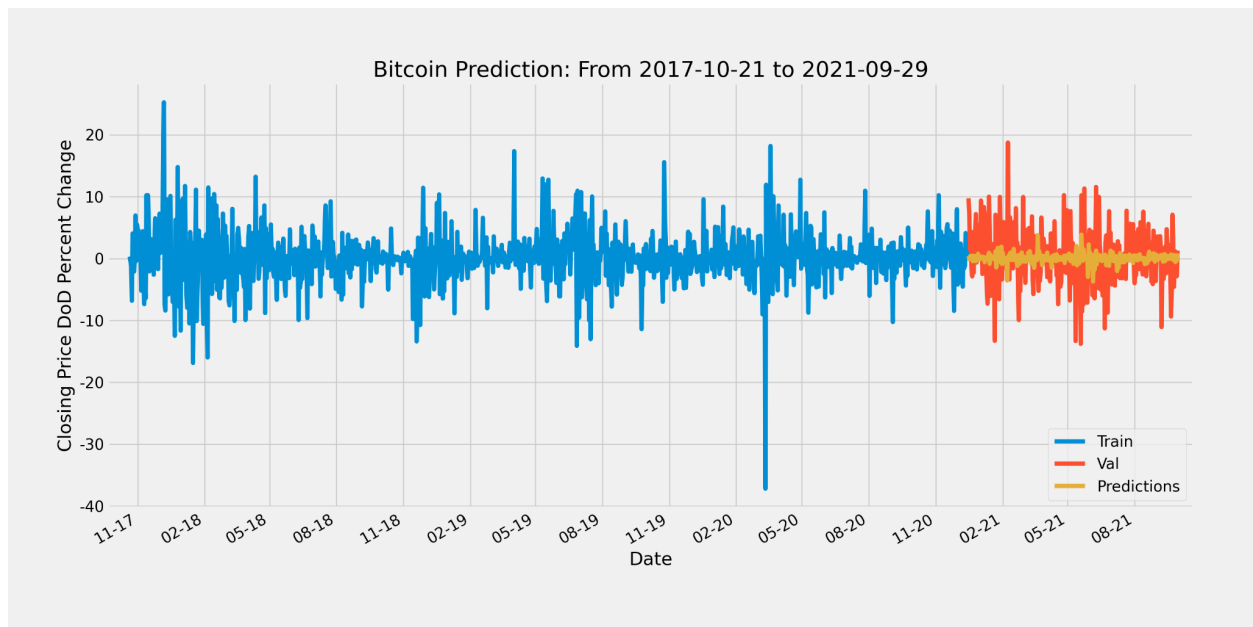
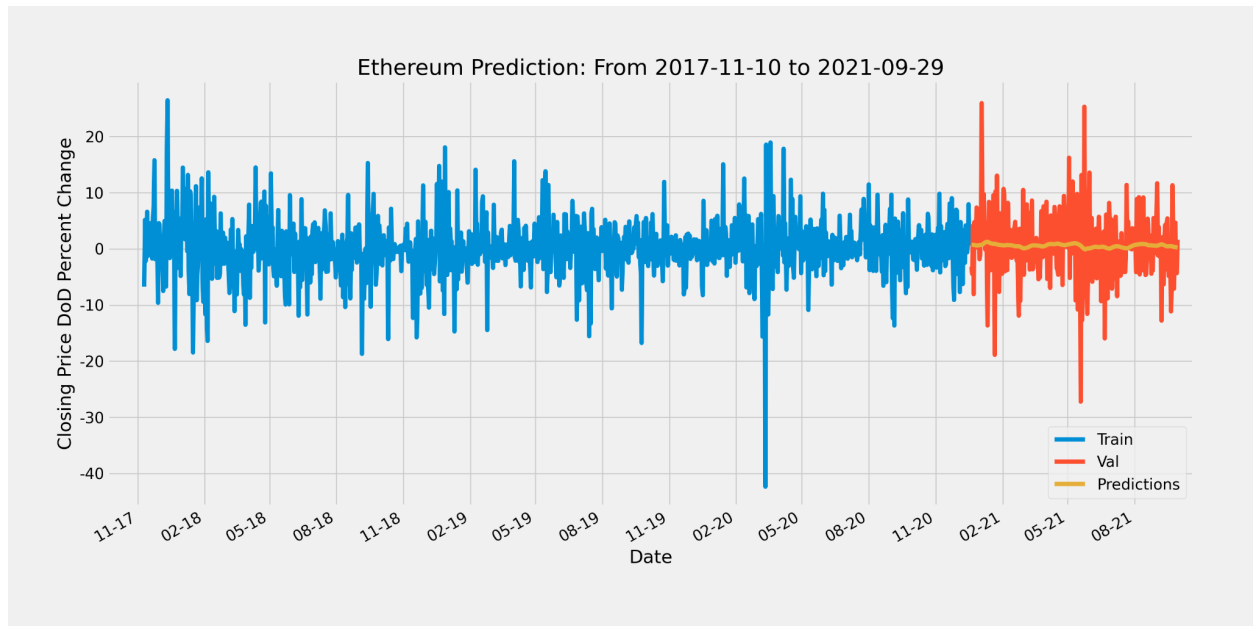
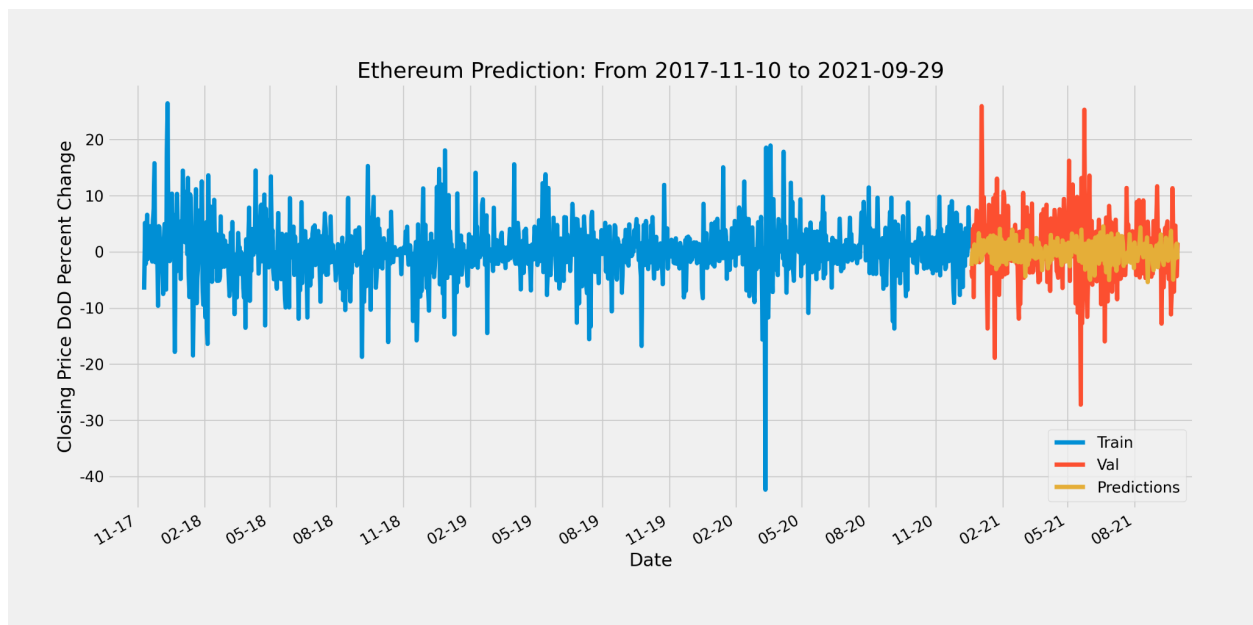


Figure D2: Ethereum

LSTM



OLS



RFR

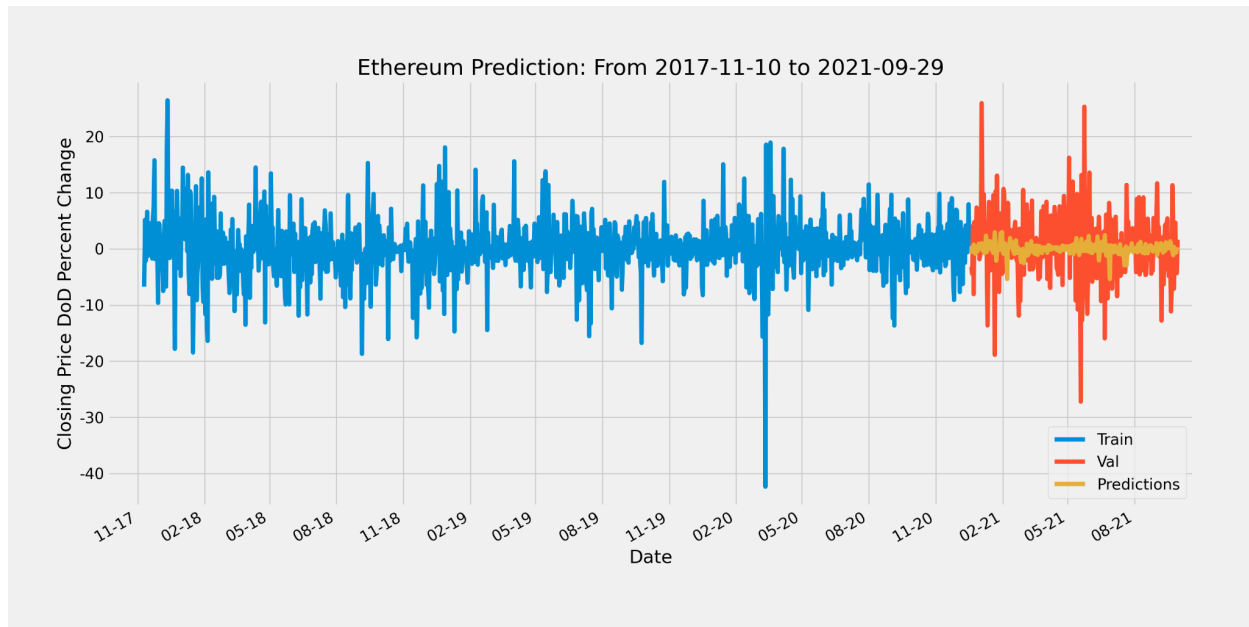
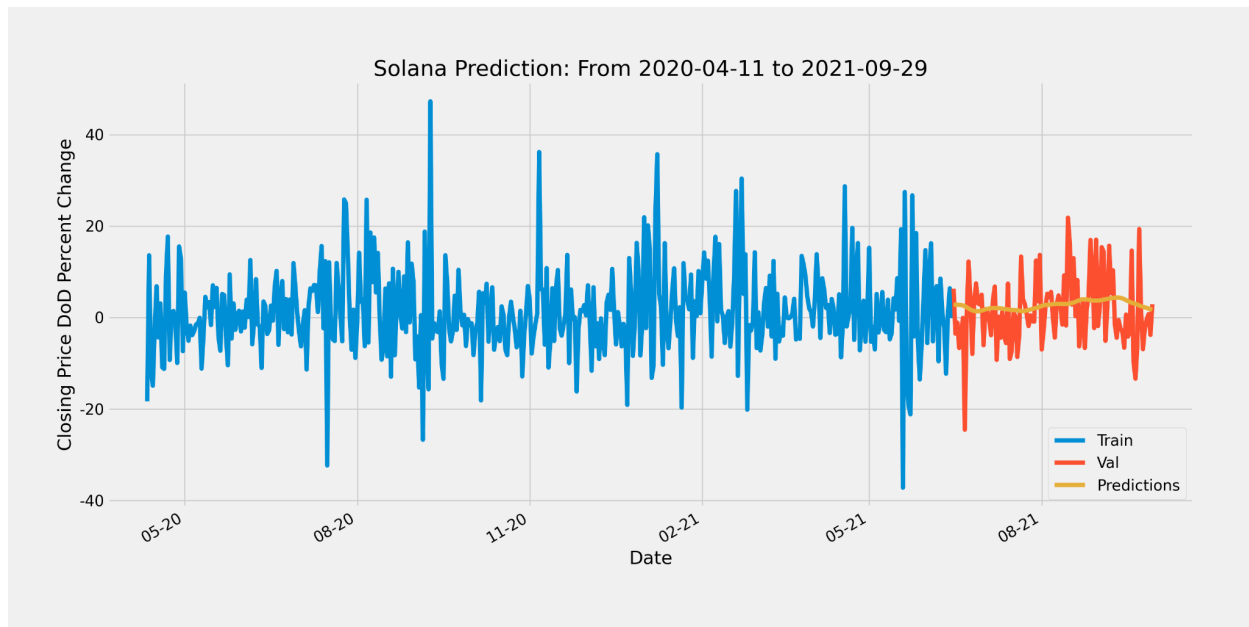
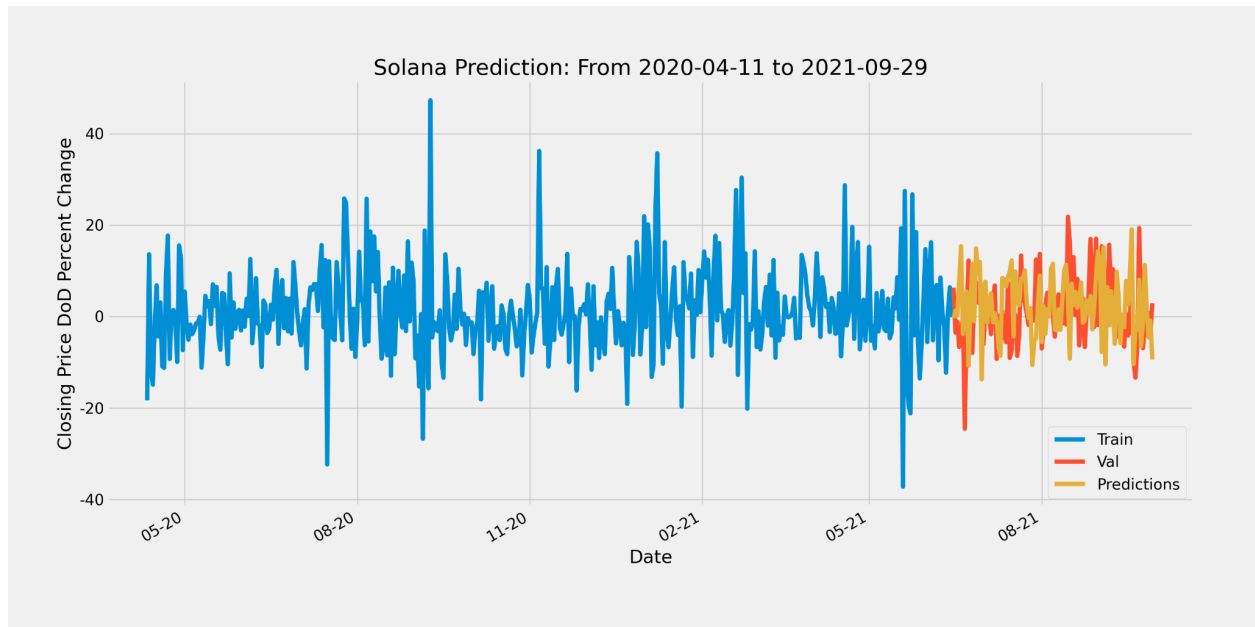


Figure D3: Solana

LSTM



OLS



RFR

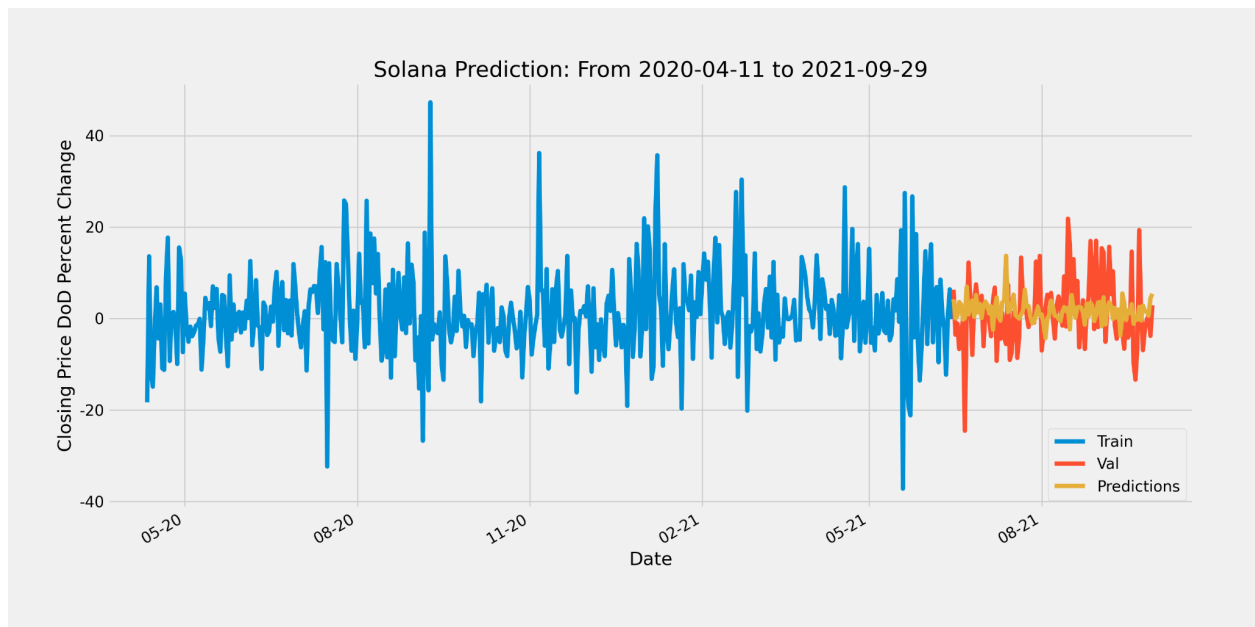
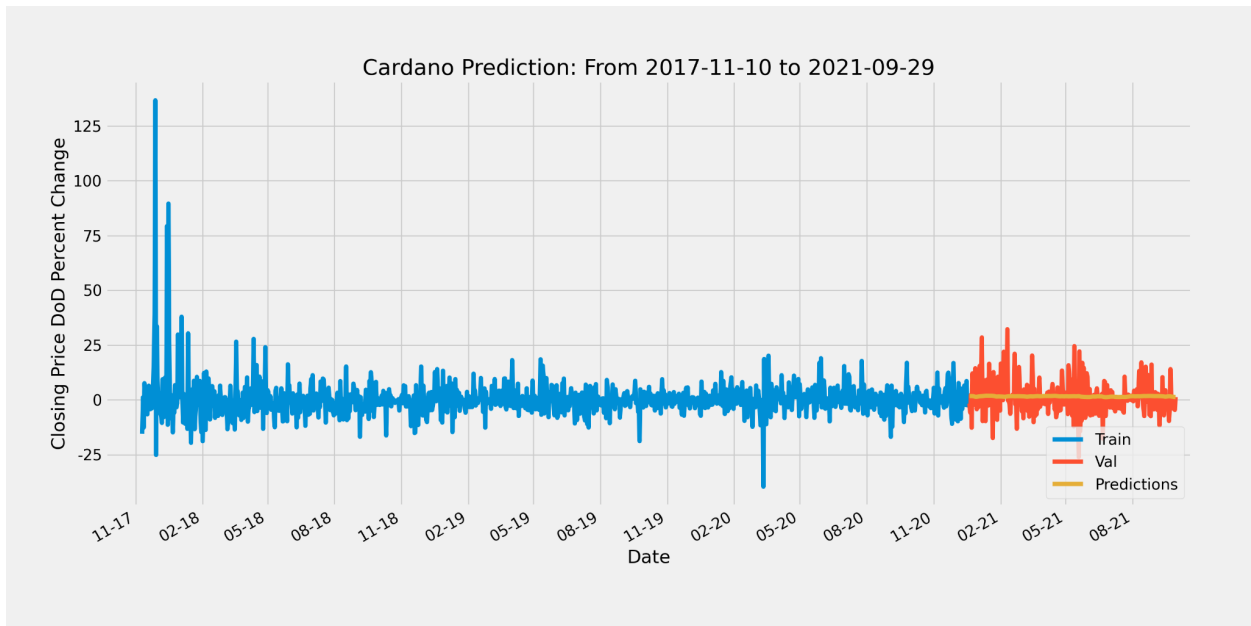
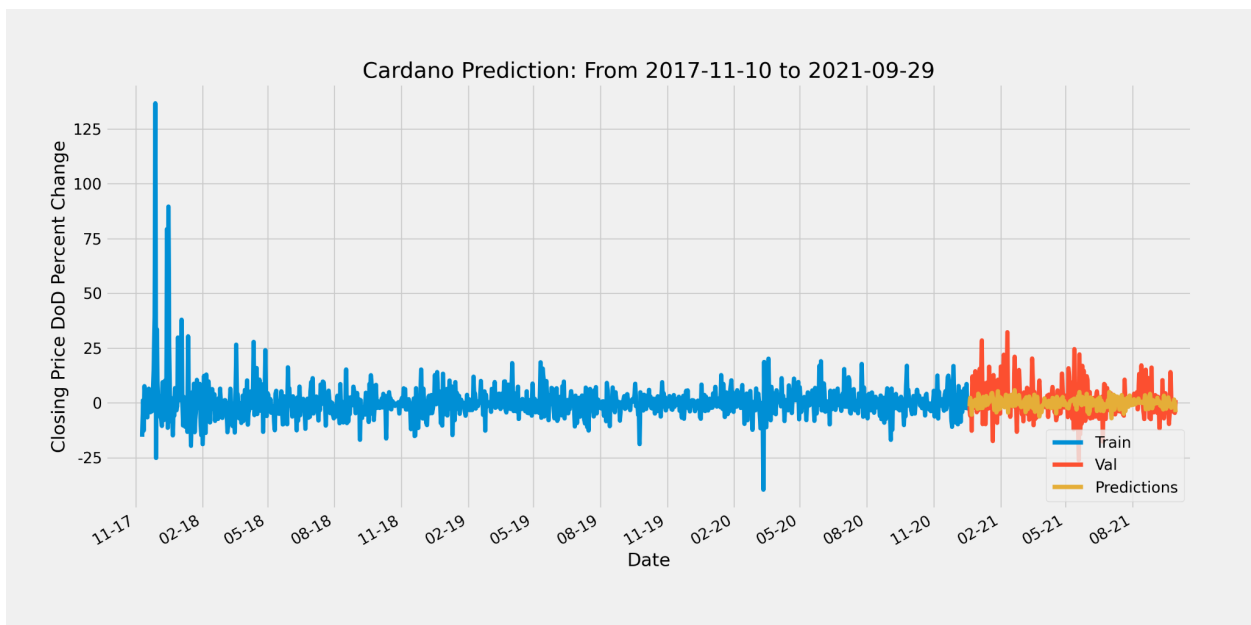


Figure D4: Cardano

LSTM



OLS



RFR

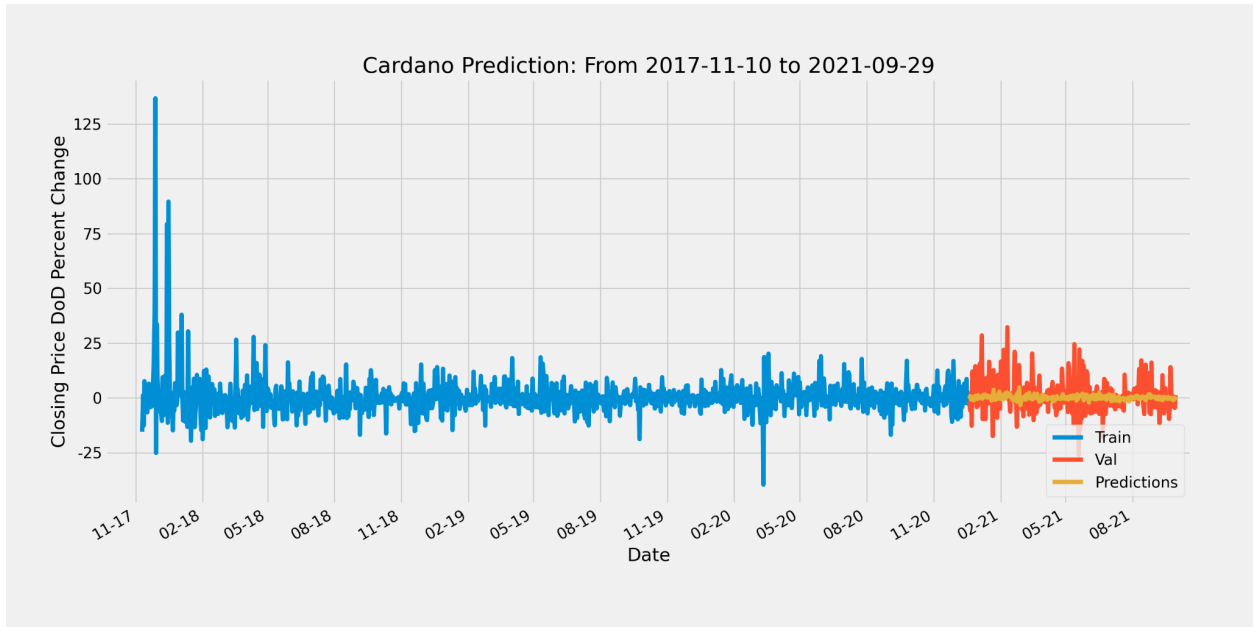
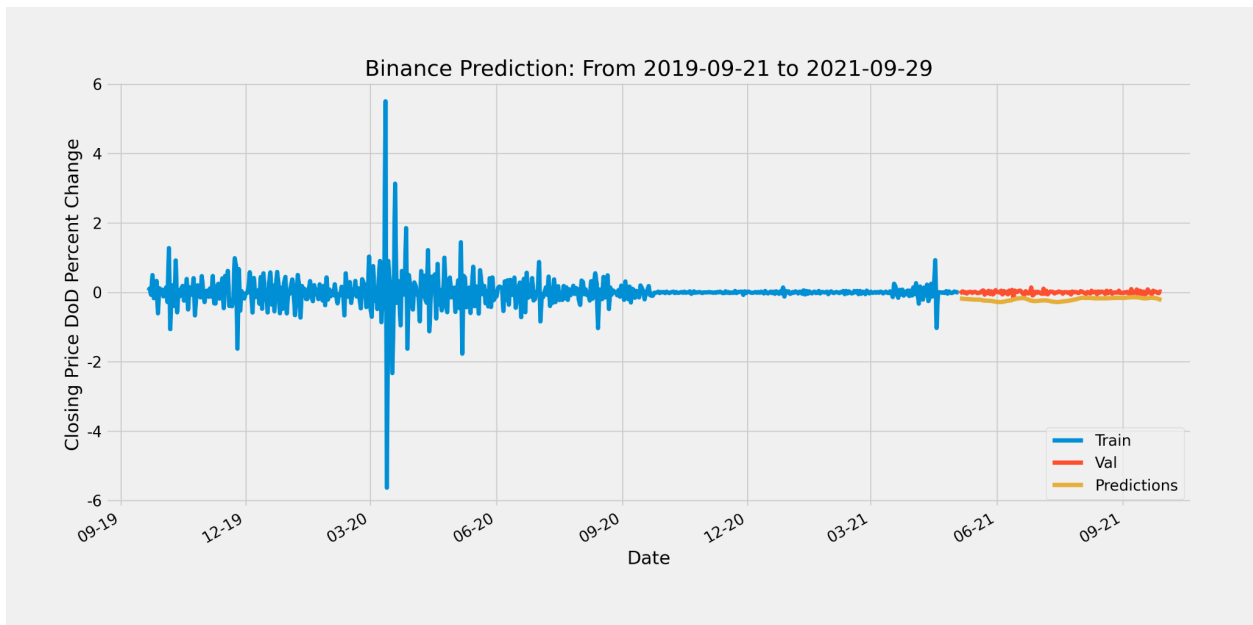
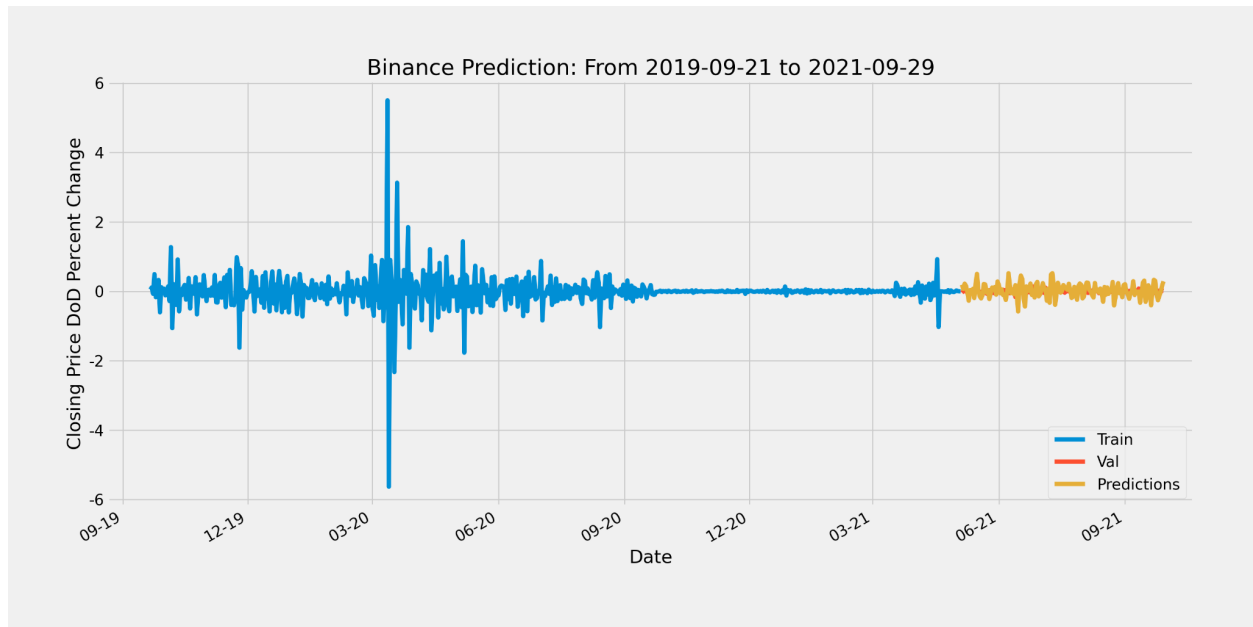


Figure D5: Binance

LSTM



OLS



RFR

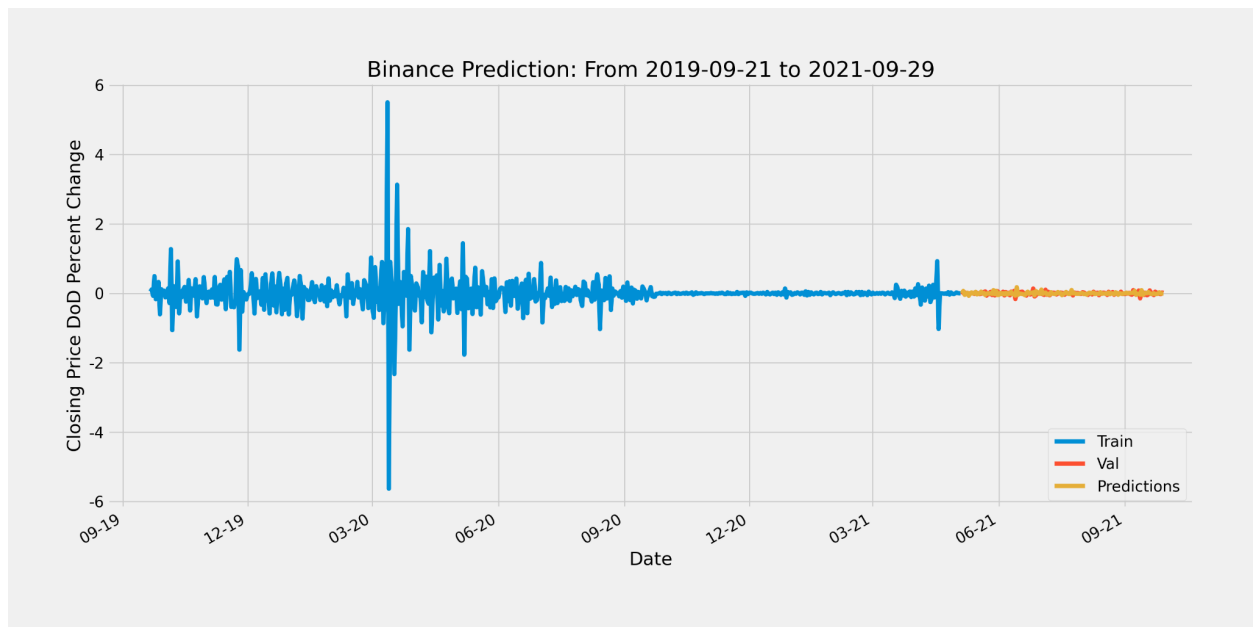
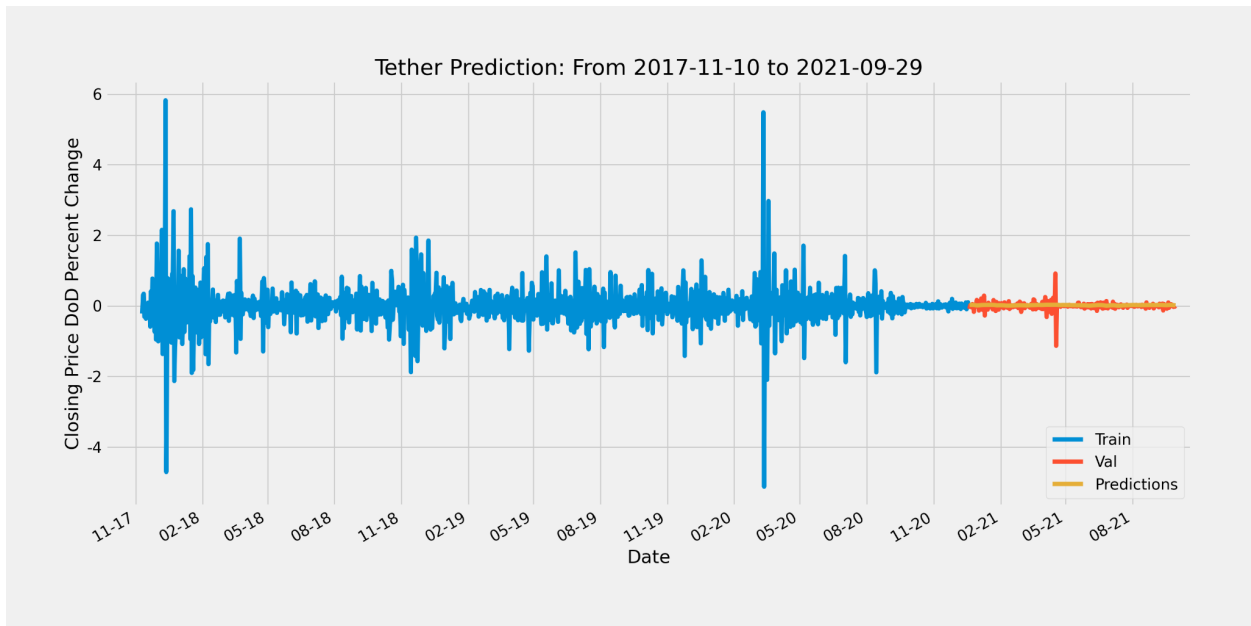
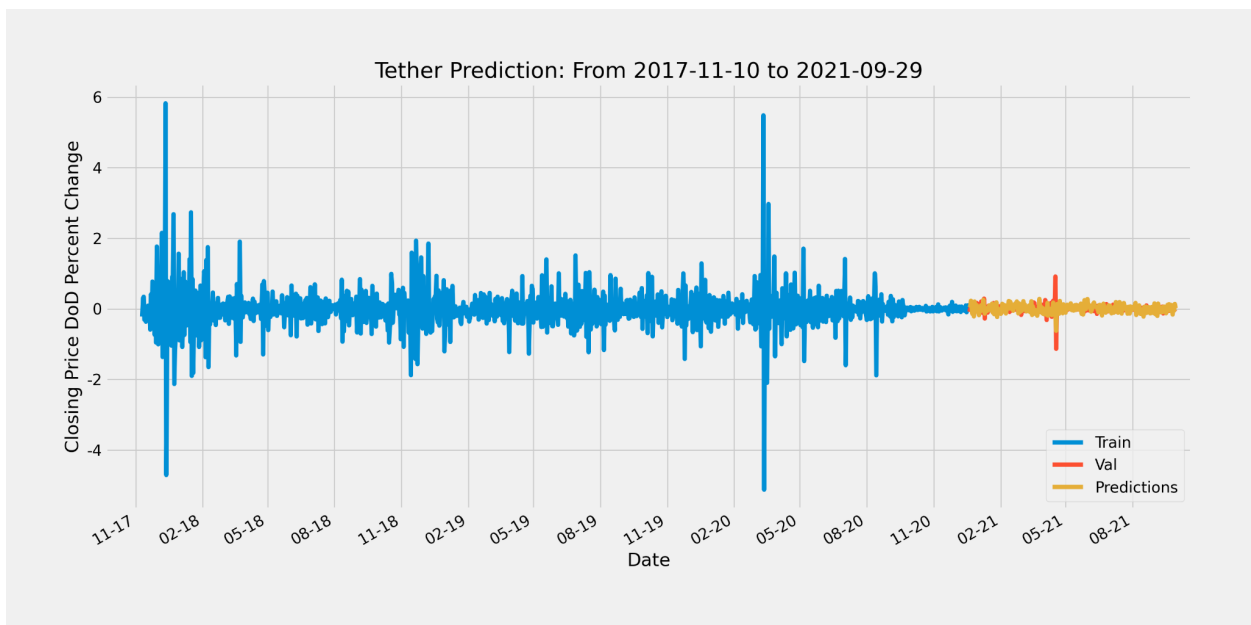


Figure D6: Tether

LSTM



OLS



RFR

