Market Timing Strategies for Crypto currencies using Kalman Filters

Rakesh Sharma P. R

WorldQuant University, Master of Science Financial Engineering
rakeshsharma.pr@gmail.com

Abstract

Machine learning algorithms power the modern stock trading strategies, with complex modeling and noisy data prediction of stock prices remains a challenge. Algo trading strategies rely on the accuracy of prediction and introduction of the Kalman filter is in this direction. In this paper I introduce Kalman Filters for forecasting stock prices. The Kalman Filter algorithm is used in dynamic systems for finding the next state of a robot or autonomous vehicle and is only dependent on the previous state. In Kalman Filter the previous state captures all the information of the history and removes the need of rolling features that make the model deterministic. Kalman Filters updates its current state based on the incoming data or observation and makes for an aggressive trading strategy on stock prices. I will be using crypto prices for forecasting and will be using other ML algorithms such as Linear Regression and LSTM for comparison of the performance. Kalman filter forecasting will be used for back testing of trading strategies which allows for combination of ML algorithms for better prediction outcomes. The code experiments are available in a jupyter notebook and at the github repo https://github.com/rakeshsharma14/WorldQuant-Capstone. *Kalman Filter methods have shown promise in this paper, with the RMSE metric it can be seen that they do well with short term predictions.*

Keywords: *Kalman Filters, Machine Learning, Crypto Trading, Timing Strategies, Pair Trading, ML Algorithms, Linear Regression, LSTM, Kalman Forecast, Stock Price Forecast*

1. Introduction

Machine Learning algorithms are changing the way we are trading stocks for both buy and sell strategy. It allows financial managers to engage in newer ways to generate the alpha to beat the market. Financial data is inherently very noisy and this makes it very challenging for accurate forecasts. Moreover, there are numerous other factors that lead to information flow affecting stock prices, making data collection and identification of their correlations a tedious task and traditional models such as ARIMA, ARMA and GARCH fail to capture this [Diego Villarino, 1]. Deep learning techniques can be trained on large historical data sets and have found application in the financial industry for a long time. They tend to outperform traditional Machine learning models [O S Alamu, 5]. Financial data is noisy and there are a lot of external factors that cause information flow and the subsequent fluctuation of prices. Machine learning is a companion of a modern trader as stock price prediction can now follow data more closely than ever and need not be driven by a mean price or intuition [O S Alamu, 5]. The search for a new algorithm prompted the need for accuracy with the noisy data.

Linear Regression is a popular technique used to predict against a trend. In many cases they have given surprisingly well. LR is easy to model and explain and works on the minimization of difference between actual predictions and actual prices. Linear Regression models suffer from overfitting when worked on rolling features [Dengxin, 2]. One problem with the Machine Learning model is that machine learning techniques suffer from frequent retraining of the data due to the data drift with the new trends of stock. For aggressive pricing financial managers need to take into factor retraining costs. There is a general tradeoff between increase in data points, retraining cost and improvement of the accuracy. If the frequency of trading is daily or hourly then this becomes a bottleneck, as training a whole dataset with just new additional data points is not very efficient.

Other techniques that have gained attention recently are Kalman Filters, and have come from other fields such as Aerospace engineering and robotics [Claudio Urrea, 4]. Kalman Filter can overcome the bottleneck of frequent training as it makes use of the difference between priori prediction and current observation in the new state to be predicted [Claudio Urrea, 4]. Kalman Filter based trading strategies allow financial managers to participate in long and short selling and can make alpha better than market returns. The problem with time series data is that it is noisy, is hard to predict and fails to generalize and will overfit for their prediction. Kalman Filters is a filtering algorithm that can be used for noisy time series data. Time series data is unique in that it is a linear combination of signal, trend and seasonality. The challenge with time series for modeling is to separate out the underlying signal from the noise. The Kalman filter in finance is also known by the name alpha – beta – gamma system. It is based on the average update trick, which means as an estimate at any time n will be the average of all the previous estimates [Qiang Li, 3]. Calculating the average for the previous values will be computationally costly.

2. Scope and Objective

The objective of this research paper is to explore the use of Kalman Filter for crypto price prediction for Bitcoin and Ethereum for short term trading strategies and compare the results with traditional Machine Learning methods - such as Linear Regression and Deep Learning architecture LSTM. Crypto data is highly volatile and it is highly non-linear in its underlying trend [Jingyang Wu, 12]. Kalman Filter on the other hand works well with dynamic systems which need real time data [Qiang Li, 3]. Using constant real time updates for crypto prices could work as an effective trading strategy for financial managers. Frequent training of data for short term prediction is not feasible in the financial world.

3. Background - Kalman Filter

Origin of the Kalman Filter algorithm can be attributed to Rudolf E. Kalman, who introduced the term Kalman Filter in 1960 in his famous paper “Describing A Recursive Solution to the Discrete-Data Linear Filtering Problem” for the estimation of non-observable state form events that may have error [Qiang Li, 3]. Kalman filters are used in systems that collect multiple data from various sensors such as GPS systems trying to estimate velocity and location. The problem with such systems is that the signals represent the hidden states and are used to predict the future system states. In such systems there is no direct way to measure state of the system and all the signals are indirect, leading to the noise in the system [Qiang Li, 3]. It is also used in where there are multiple sources of signals and is a powerful tool for combining information in the presence of uncertainty. It can be used in any system where there is uncertainty and one has to make educated guesses [Camile J.J. Beckers, 10]. Our stock price prediction is such a system and Kalman Filter makes an ideal choice for this application. There are several strengths of using Kalman Filter in addition to the uncertain nature of the system. It is a linear model and is not computationally expensive. It is light on memory and as the previous state of the system captures the history and so it need not keep any historic states like a long-term memory system [Camiel J.J. Beckers, 10]. Due to this nature, it is extremely fast and is used in real time systems. Kalman systems assume that the state of the current system can be dependent on any independent variables, which are Gaussian distributed. Each variable has a mean μ and is the center of the distribution or the likely state [Camiel J.J. Beckers, 10]. The variance of the distribution or σ^2 is the uncertainty. At any point in time the future stock price can be any combination of these variables picked from the distribution but certain values are more likely than others, this is the very essence of the Kalman Filter Algorithm. With Kalman Filter updates stock price can be constantly updated with the incoming real time data and makes it easy to change or build trading strategies in search of superior α.

4. Literature Review

There are several papers that this research allows me to explore and careful study tells us that there is a gap in the methods used in prediction of stock prices with the currently used Machine Learning methods. ARCH, ARMA and GARCH models are popular for modeling and interpreting time series data. The ARCH model was proposed by Engel to explain clustering and persistence of the stock market. Bollerslev extended the ARCH model to GARCH, which could also model variance of error [Ningyi Li, 6]. On the similar lines ARMA is suitable for short term prediction and is suitable for the study of stationary stochastic processes. These models require time series to be random and stable [Ningyi Li, 6]. The paper from Ningyi Li et al clearly brings out the need for combination of ARMA, ARCH and GARCH for various lag levels (p, q) parameters to capture the time series data such as stock index yield, interest rate market risk, exchange rate volatility etc. [Ningyi Li, 6].

In the paper by O S Alamu and Md Kamrul Siam in "Stock Price Prediction and Traditional Models: An Approach to Achieve Short-, Medium- and Long-Term Goals" concluded that deep learning methods such as LSTM outperform traditional methods as they are able to capture complexities and nonlinear pattern of data [O S Alamu, 5]. These methods also make it less interpretable and require greater computational resources. Daily trading strategies using these models involve trade-offs on accuracy and computational complexity. Another paper from John Pahn and Hung-Fung Chang on "Leveraging Fundamental Analysis for Stock Trend Prediction for Profit" employs an CNN-LSTM model that uses technical analysis details to achieve greater accuracy [John Pahn, 9]. The paper indicates that feature engineering is an integral part of stock price prediction and there is value in incorporating external factors for improving accuracy.

In another paper from Dengxin Huang, he studied Apple stock for stock market forecasting and uses Fama French 3-factor, Linear Regression, Random Forest Regressor and Gradient Boosting Regression model for comparison is critical to performance improvement of stock price predictions. Linear Regression emerges as the best performing model in that case study [Dengxin, 2]. Simplicity of Linear Regression and interpretability made a strong case for it to be used for stock price prediction with feature engineering techniques on stock data.

There is various research that has highlighted the use of the Kalman Filter algorithm is used for state estimation. Notable among those is one using Kalman Filter to estimate the state of storage battery capacity. The paper by Camiel J.J. Beckers et al uses this technique to estimate the aging of batteries and its economic lifetime on the uncertainty of real-world data [Camiel J.J Becker, 9]. The research uses a combination of joint Extended Kalman Filter, combined with Recursive Least Squares to estimate the impedance. The algorithms converge quickly to the trend of the capacity and resistance in this case. The cost of computation is very minimal in the process the data is updated in the real time driving scenario in this case. Kalman Filter has been in the use in industry for many years and is known for its robustness, this explained in detail in a paper by Claudio Urrea et.al in their paper "Kalman Filter: Historical Overview and Review of Its Use in Robotics 60 Years after Its Creation" and inspires our insearch for Financial market use cases.

4.1 Competitor Analysis

The literature review above brings our focus on the new method to improve stock price prediction and why Kalman Filter can be a robust method for stock price prediction due to its noisy and dynamic nature of the financial data. A competitor analysis of the current methods helped identify where current research fits in. Kalman Filter updates capture the history in the current state and represents the highest likelihood of the stock price value, which is the future state predicted. This offers an opportunity to predict more on real time data and allows frequent predictions like hourly data. Traditional ML methods would be computationally costly and to train on real time data for higher frequencies. The longer the prediction we do, the more static the prediction will be, such models will also suffer from overfitting. The strength of this algorithm allows for high frequency trading strategies, the Kalman Forecast signal can also be combined with other data for more robust prediction and can be a scope for further research. It is clear from many of the research papers we have studied Linear Regression with feature engineering works better than some of the Deep learning methods for equity stock prediction. CNN-LSTM based architecture increases the computational complexity and may not be ideal for prediction based on real time updates. Kalman Filter applications are more suitable for real time updates, which is the case with stock price prediction and forecast will have a shorter window. Traditional forecasting algorithms are suitable for forecasting for longer periods and executing long term strategies.

At the same time, crypto currency trading has been gaining momentum in recent years. Even though adoption of crypto as a legitimate currency is debated by many and it has seen vehement opposition by many states and its banking institutions, it has increased in valuation and is now accepted as an alternate asset to hedge equities or other forms of assets [Bingqia Luo, 13]. Crypto is characterized by high levels of volatility compared to other stocks. It is also considered as a high-risk investment. Predicting crypto prices using algorithms is a challenge due to these factors. The global value of Crypto currency was at \$3 trillion in 2021, however next year the value dropped down to \$1 trillion, such is the volatility and risk inherent in crypto trading [Duy Thien An Nguyen, 10]. Research has concluded that multivariate Convolutional LSTM gives better performance for Crypto Price prediction during Covid-19 period where the volatility was high. All these strengthen the need to find alternate methods such as Kalman Filter as a replacement for Crypto price prediction in this research.

5. Methodology

In this section we would be describing the design and process of the research done to understand the performance of the Kalman Filter algorithm. The methodology follows guidelines and pitfalls highlighted in the research paper on Machine Learning Pitfalls [Michael A. Lone, 12]. We have used best practices for supervised learning followed in industry for the comparison. The below sections are various steps performed before reaching the research outputs. All the coding is done in python language and we have relied on verified python libraries to implement the Kalman and other ML algorithms. Kalman Filter implementation is done using *pykalman*, an open source package available at https://pykalman.github.io/

5.1 Data Collection - For this research we have collected crypto currency data, namely Ethereum and Bitcoin for the last two years from yahoo finance website. Due to its incompleteness, and incorrectness we have decided to read data from yahoo.com website. Other sources that we have considered are Kaggle and data available in Github. The observation window is from *2022-09-01* to *2024-08-31*. We have taken the data from the post Covid-19 event as that event has added a lot of volatility to the crypto data and wanted to remove watershed events from our study and observation. The data can be collected using the APIs from $yahoo\ finance\ package$ which can be called in python and read into the data frame data structure for $pandas$. This makes it easy to analyze the data in a $jupyter\ notebook\ environment$ without much hassle. The data points are at daily frequency, there are also lower frequencies available like hourly and its applicability of Kalman algorithm for real time updates could be a study for the future. The reason for choosing ETH and BTC are that they are the leaders in the market of more than 10,000 cryptocurrencies [Bigqiao Lu, 12]. They are also the oldest crypto currencies and represent information and nuances from this space such as NFTs and other emerging trends. Ethereum is traded with the ticker label ETH-USD and Bitcoin uses the ticker symbol BTC-USD. Both ETH-USD and BTC-USD are crypto linked to the US dollars for their value. Etherium was developed by Vitalik Buterin and Gavin Wood in 2013 and has the largest market cap after the Bitcoin currency [Bigqiao Lu, 12]. Bitcoin was developed by Satoshi Nakamoto in 2009 and is the most popular crypto currency [Bigqiao Lu, 12]. Ethereum and Bitcoin can be traded with peers or can be received from apps and exchanges [Buterin, 14].

The below figure shows the top five rows of Ethereum (ETH-USD) and Bitcoin (BTC-USD) in their dataframe structure.

	Open	Close	High	Low	Volume	Adj Close
Date						
2022-09-01 00:00:00+00:00	1553.756348	1586.176758	1593.082764	1520.188354	1520.188354	1586.176758
2022-09-02 00:00:00+00:00	1586.017944	1577.220459	1643.183228	1551.877930	1551.877930	1577.220459
2022-09-03 00:00:00+00:00	1577.213745	1556.872681	1579.454346	1541.672119	1541.672119	1556.872681
2022-09-04 00:00:00+00:00	1556.895874	1577.641602	1578.009277	1543.698853	1543.698853	1577.641602
2022-09-05 00:00:00+00:00	1577.884033	1617.183228	1621.661377	1559.781860	1559.781860	1617.183228

	Open	Close	High	Low	Volume	Adj Close
Date						
2022-09-01 00:00:00+00:00	20050.498047	20127.140625	20198.390625	19653.968750	19653.968750	20127.140625
2022-09-02 00:00:00+00:00	20126.072266	19969.771484	20401.568359	19814.765625	19814.765625	19969.771484
2022-09-03 00:00:00+00:00	19969.718750	19832.087891	20037.009766	19698.355469	19698.355469	19832.087891
2022-09-04 00:00:00+00:00	19832.470703	19986.712891	19999.689453	19636.816406	19636.816406	19986.712891
2022-09-05 00:00:00+00:00	19988.789062	19812.371094	20031.160156	19673.046875	19673.046875	19812.371094

5.2 Data Preprocessing (Cleaning) - The dataset has been used for multivariate modeling and has these 4 columns namely $open, high, low, close, adj\ close\ and\ volume$ as values. The following checks are done as part of preprocessing. Both the data series are checked for $null\ and\ not\ a\ number\ (NA)$ values in the dataset, for rolling properties, where the column values are null are eliminated. There is data availability for all the date indices for BTC and ETH. We have applied and scaled the data series for ETH and BTC. Standardization is the technique we have applied to scale the data. Post standardization the data series will have a mean of zero and standard deviation of one. It can be defined by the formula

$$z = \frac{X - \mu}{\sigma} =$$

where X is the data point, μ is the mean of the data series and σ is the standard deviation. We have used the python library $scikit - learn$ to achieve the scaling. The output of the data preprocessing is the training dataset.

5.3 Exploratory Data Analysis - Summary statistics and visualization of time series will help understand the trend, seasonality and noise with crypto data. Exploratory analysis gave the following conclusion about the crypto data. Crypto data is noisy and will have to use techniques to separate out trend and seasonality. Kalman Forecast separates out the signal from the noise. $Autocorrelation$ and $partial\ autocorrelation$ functions identified the lags significant for BTC and ETH data. Other columns will be used as independent variables. We have used the statsmodel package from python to plot the autocorrelation and have performed the Dickey fuller test to confirm the trend and seasonality.

We can see the summary statistics of the ETH-USD and BTC-USD here.

	Open	Close	High	Low	Volume	Adj Close	Target
count	729.000000	729.000000	729.000000	729.000000	729.000000	729.000000	729.000000
mean	2152.824434	2154.140681	2197.411646	2106.811007	2106.811007	2154.140681	2155.429632
std	761.452613	761.256721	781.630105	737.071901	737.071901	761.256721	761.089210
min	1100.107178	1100.169800	1136.442627	1081.138184	1081.138184	1100.169800	1100.169800
25%	1617.240234	1619.698486	1644.727539	1580.165527	1580.165527	1619.698486	1622.890625
50%	1865.594971	1865.636108	1887.705322	1845.719238	1845.719238	1865.636108	1866.564209
75%	2641.685303	2642.185303	2710.421875	2587.110596	2587.110596	2642.185303	2642.185303
max	4066.690430	4066.445068	4092.284180	3936.627197	3936.627197	4066.445068	4066.445068

	Open	Close	High	Low	Volume	Adj Close
count	730.000000	730.000000	730.000000	730.000000	730.000000	730.000000
mean	37464.146091	37518.319863	38162.137949	36780.942348	36780.942348	37518.319863
std	17678.426052	17685.270401	18083.910703	17220.739292	17220.739292	17685.270401
min	15782.300781	15787.284180	16253.047852	15599.046875	15599.046875	15787.284180
25%	23627.717285	23665.855469	24119.581543	23253.754883	23253.754883	23665.855469
50%	29403.917969	29412.204102	29845.836914	29113.966797	29113.966797	29412.204102
75%	55644.687500	55988.014648	57679.622070	54234.083008	54234.083008	55988.014648
max	73079.375000	73083.500000	73750.070312	71334.093750	71334.093750	73083.500000

5.4 Feature Engineering - I have used multivariate Linear Regression to train the model to predict Closing Price and Volume. Feature engineering is a process that involves transforming a variable x into a form $f(x)$ to be used in the modeling. We have defined the target variable for prediction in this research. The target variable is next day's $Adj\ Close\ Price$. We use the shift operator to generate this variable in the dataframe. In the case of Volume Prediction, we try to predict the next day's Volume as the target variable y. In the linear regression model we have used standardized $Opening\ Price, Closing\ Price, High\ Price, Low\ Price$ to predict $Adj\ Closing\ Price$. The $Adj\ Closing\ Price$ is the dependent variable y or and the rest of the variables are dependent variables or features $[x_1, x_2, x_3, \ldots x_n]$. Rolling means with various lags are also part of the feature set as part of the exploratory analysis, they have not been included for training. For Volume prediction, $Volume$ is the dependent variable y and standardized $Opening\ Price, Closing\ Price, High\ Price, Low\ Price$ are the features $[x_1, x_2, x_3, \ldots x_n]$. Kalman Filter uses an update algorithm and is different from supervised learning and will not involve feature engineering. It predicts the current state of stock price, which can be expressed here as ($x_{t|t-1}$) and uncertainty ($P_{t|t-1}$) at time step t based on the stock price and uncertainty at time step $t-1$ [Mario Filho, 18].

5.5 Training and Modeling - The dataset was divided into training, validation sets. The dataset obtained from preprocessing is divided into ratios of 8:2 to training, validation and testing set. Training set is the dataset used to train the model and learn the underlying pattern and coefficients (weights). Validation set is used to backtest the data. Trading strategies will be tested using back testing techniques to understand the overall return. For the Kalman Filter forecast, the dataset will be updated from the last point of the training set. The value will be updated for each observation in the test set sequentially. Models trained are LSTM, Linear Regression and Kalman Filter Methods. The comparison of the models using the performance metric is done on the validation set.

5.5.1 Model Analysis
We are employing 3 models here
Linear Regression - It is an algorithm that predicts the dependent variable based on independent variables. It identifies a best fit line or surface that minimizes the error difference between the predicted values and actual observations. It is used for forecasting values and understanding the relationship between the dependent variable y and independent variables [O.S. Alamu, 5].

LSTM - It is a type of Recurrent Neural Network that has become popular in the field of deep learning. RNNs are popular for learning the patterns of temporal dependencies. LSTM solves the problem of vanishing gradient by having a memory cell in the hidden layer. Each input of LSTM is capable of memorizing the input information for a long time in this architecture and hence the name [Jingyang Wu, 11].

Kalman Filters - The crux of this paper is around Kalman Filter which is used for estimating and predicting states of dynamic systems. Kalman Filters uses the Kalman gain update process to update the value of a point estimate for the future value. The future value is the most likely estimate of the point based on the history of the value [Camiel J.J. Backer, 9].

5.6 Prediction Metrics - We have used **RMSE**, **MSE**, **MAE** for comparison.

RMSE stands for root means square error - It is the average difference between values predicted and actual values in a regression model. It is given by the formula

$$RMSE = \sqrt{1/N \sum_{i=1}^{n} (\hat{y_i} - y_i)^2}$$

It represents standard deviation of residuals and value ranges from $0 \ to \ \infty$. Lower the value better the model. The choice of RMSE is due to the fact that it is an absolute measure of error and it can assess prediction precision.

MSE - Mean square error, it is the squared difference between predicted and actual values. It represents the residuals, the values which the regression model is not able to account for. It can be represented mathematically as below

$$MSE = 1/N \sum_{i=1}^{n} (y_i - y)^2$$

There are several advantages of using this metric, it eliminates the negative values and can be compared to variance. Irrespective of the direction this is always a positive value. For a perfect model the MSE value will be 0. Squaring magnifies the larger error than smaller ones and so penalizes such differences in the model and learns to keep differences small in training.

MAE - Mean absolute error, it is the sum of absolute error divided by the sample size. The sum of absolute difference is also called Manhattan distance. It is given as below formula

$$MAE = 1/N \sum_{i=1}^{n} |y_i - y|$$

Unlike squared errors, it gives equal weightage to each of the observation and prediction pairs and is useful for understanding the magnitude of errors. The reason for choosing this metric is its robustness to outliers, interpretability irrespective of direction of the prediction and its intuitive nature as it is Manhattan distance. Smaller the MAE better the prediction and ideally, we would like to get a MAE of 0 which means observation and prediction are the same.

5.5 Comparison

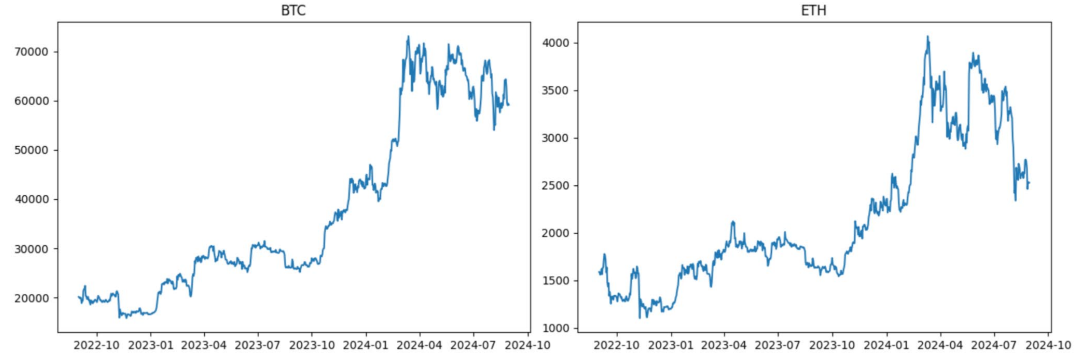
Comparison of Kalman Filter has been done with 2 other models, Linear Regression and LSTM. We have done comparisons for different periods of price forecast for different periods, $14, 25, 50, 100, 200$ days. We have also done a comparison with volume forecasts of the 2 cryptocurrencies for $14, 25, 50, 100, 200$ days. The comparison is done with the above defined metrics. Backtesting is also performed for past data for the forecasted period 14, 25, 50, 100 and 200 days to assess the potential of trading strategies. Trading strategy in crypto currency is defined as entry, exit for long and short positions with a positive position resulting in positive returns overall. Based on the time period of holding positions there can be long term and short-term trading strategies. For this research we will consider 14 and 25 days as short term and anything more than 50 days will be seen as long term. Comparison will also be done for long term and short-term periods for the 2 models.

6. Results and Discussion

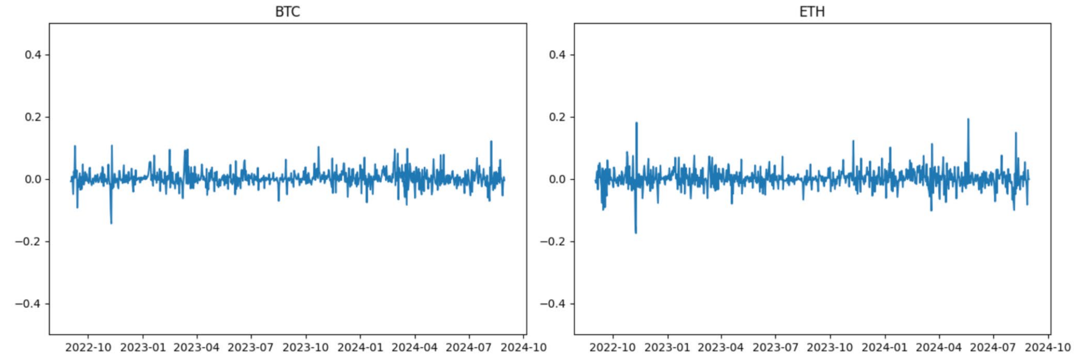
Before building any models, we should explore cryptocurrency data thoroughly to understand the underlying complexities, noise, and trends. In the below section we will look into details of the nature of the ETH-USD and BTC-USD and compare their performance with respect to other models.

6.1 Understanding Crypto Data

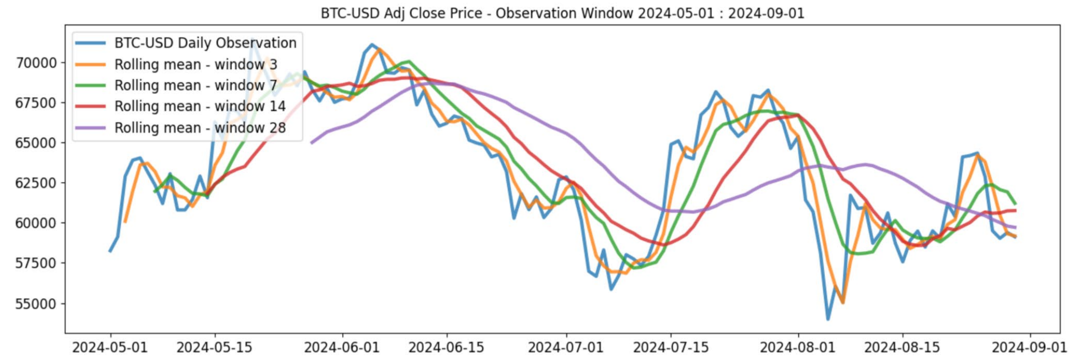
We will be modeling crypto data for the **Adjusted closing price** and **Volume.** These two factors contribute to our trading strategies.

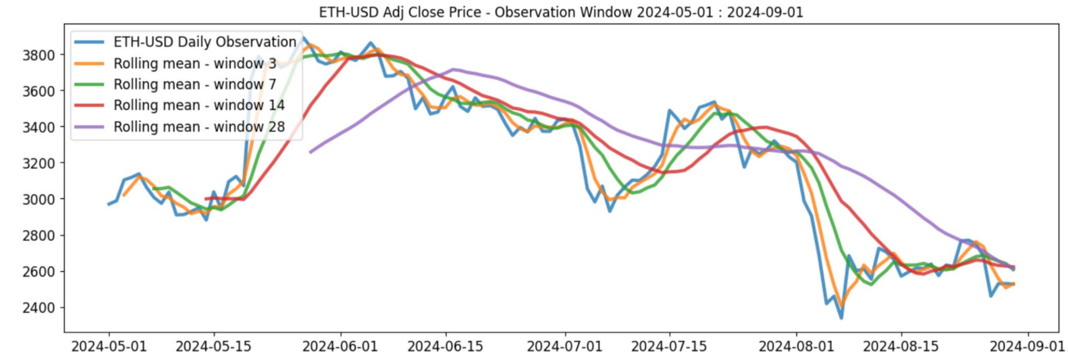


From the figure above it is clear that crypto data has an upward trend. BTC-USD has a higher value compared to ETH-USD for the same time period. Both the time series have a similar trend during the observation period.



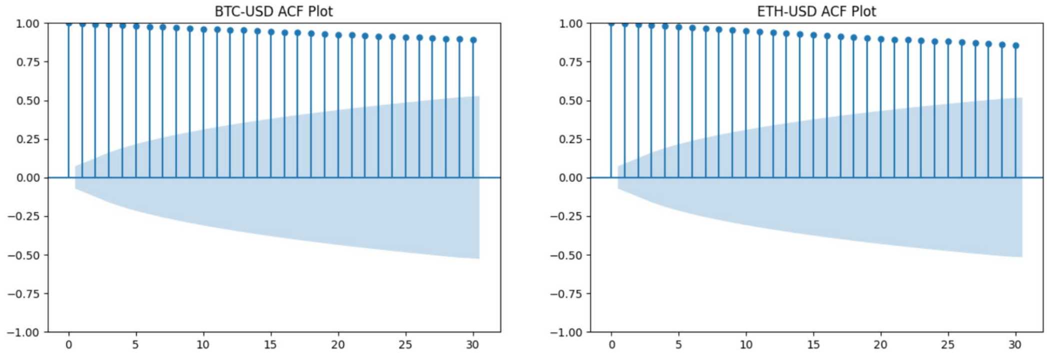
The plot for percent change in the above figure for the time period shows downward and upward jumps in the data. Both time series have indicated high volatility for the time period. We can apply some level of smoothing to separate the noise and understand the underlying trend as seen below.

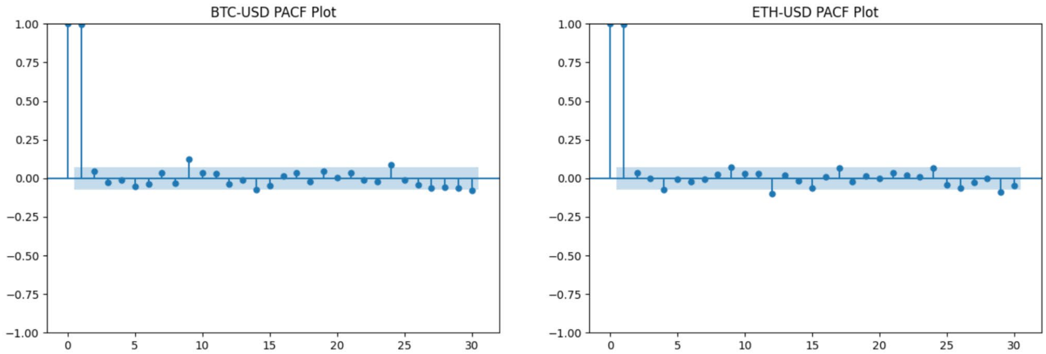




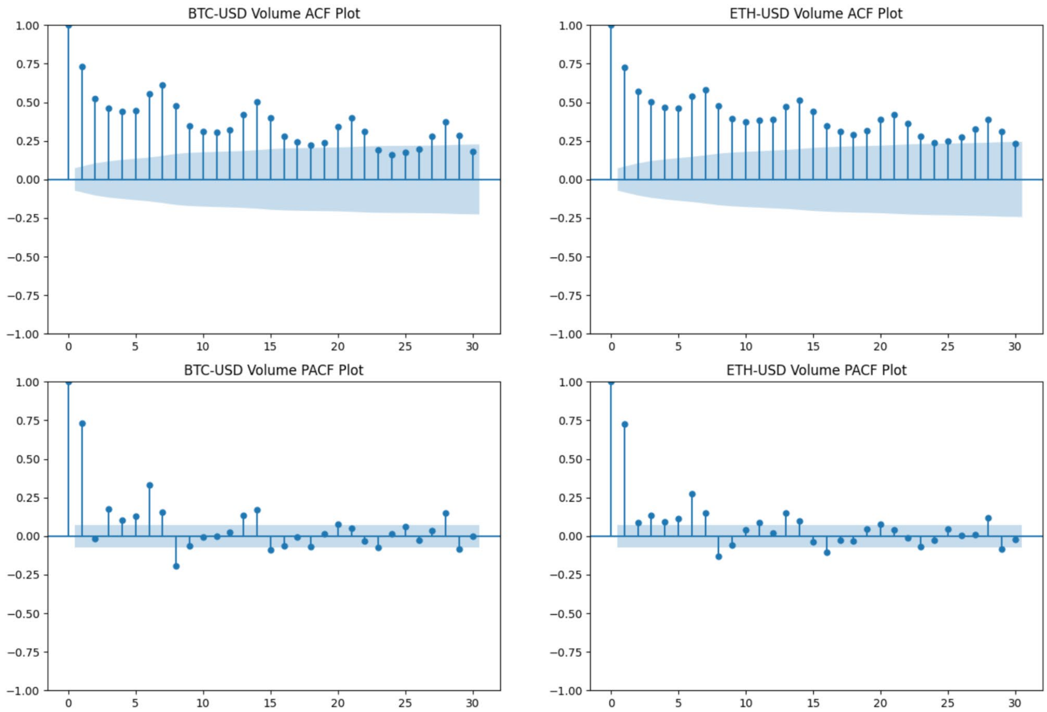
Overall Ethereum has a smoother trend for the rolling period than BTC-USD, for the observation period.

Let us look at the Volume shocks for the data. Smoothed Volume data shows less volatility and remains stationary. Volume data should be more suitable for prediction of strategies.





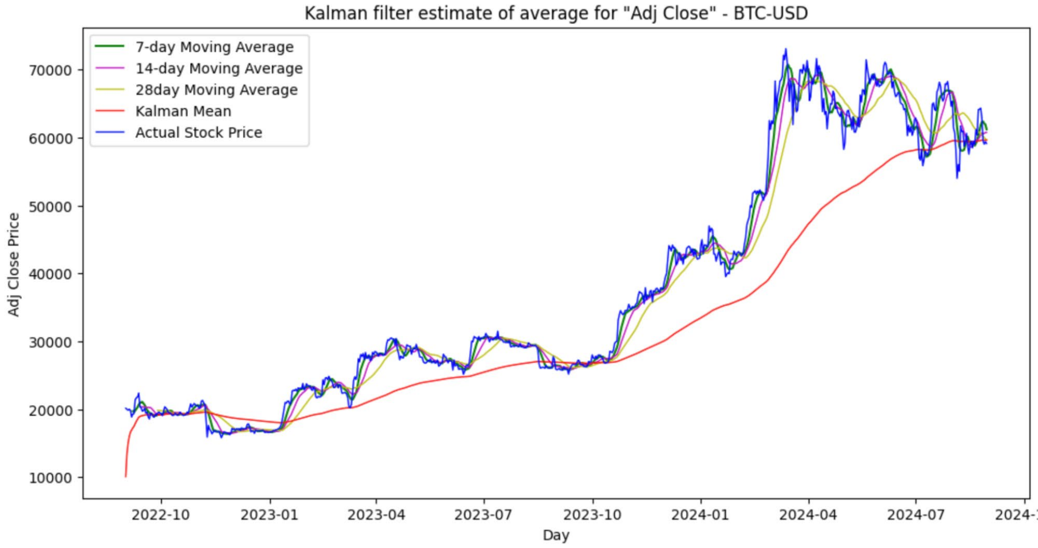
The plot shows significant autocorrelation, the previous value of the stock price determines day's closing price. From the PACF plot it is clear that the price of certain periods of time is more correlated than others, say for example price on day 4, 9, 12, 15 etc.

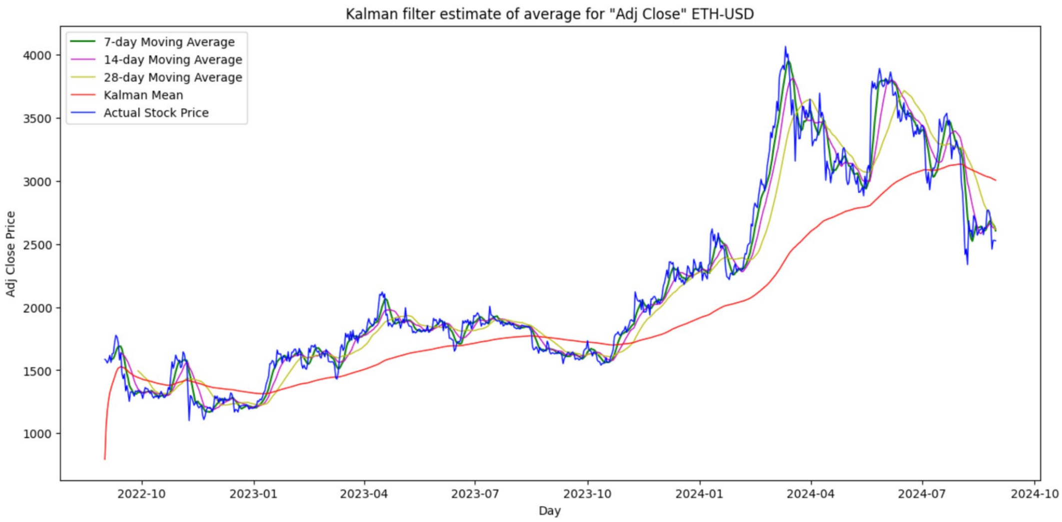


The ACF and PACF plots for Volume show significant correlation with observations at the previous time slots. This validates that regression on observation on previous time slots with a constant will be good to model this time series. This explains our choice of Linear Regression and LSTM model for prediction as it makes use of auto-regression.

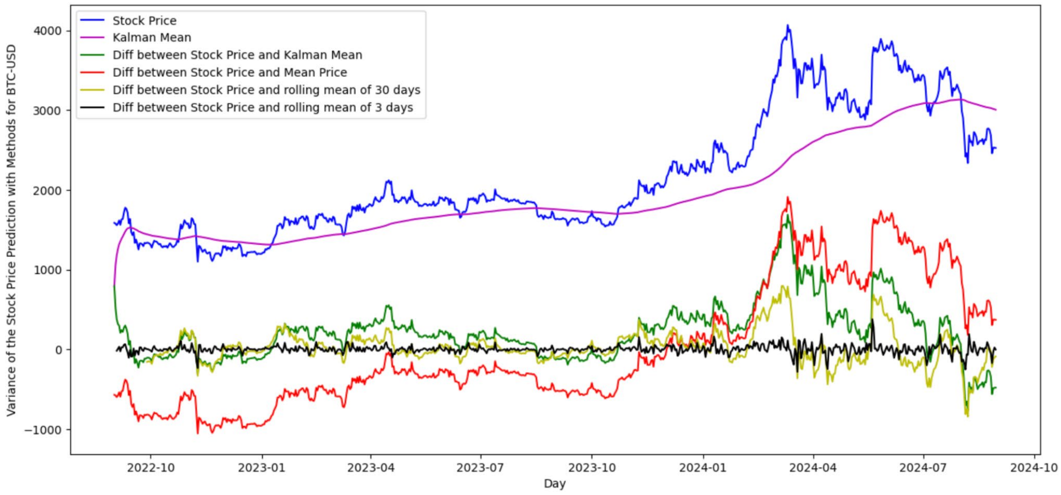
6.2 Kalman Filter

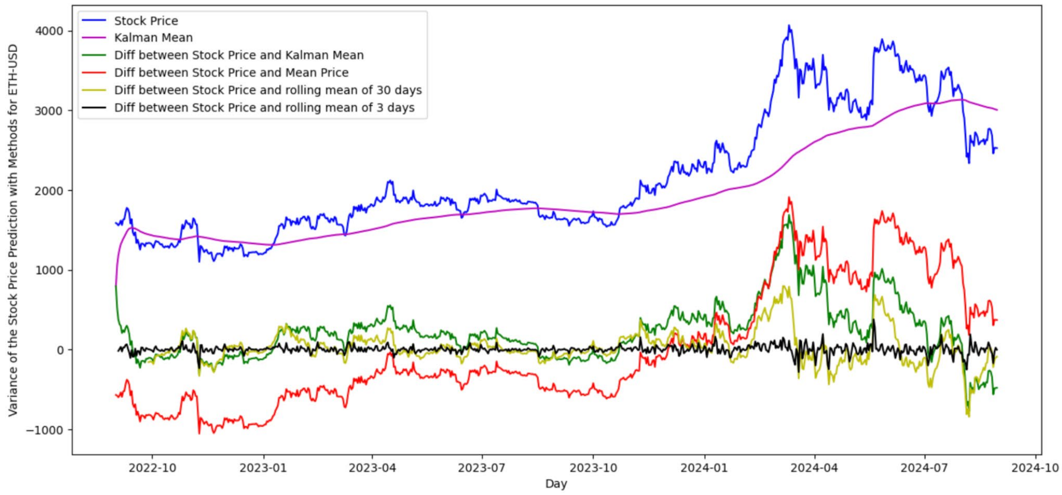
From the plots above it is inferred that crypto data is noisy, and highly volatile. There is an uptrend with strong autocorrelation with certain values at time spot $t - k$. In this section we will bring the observations from the Kalman Filter forecast.



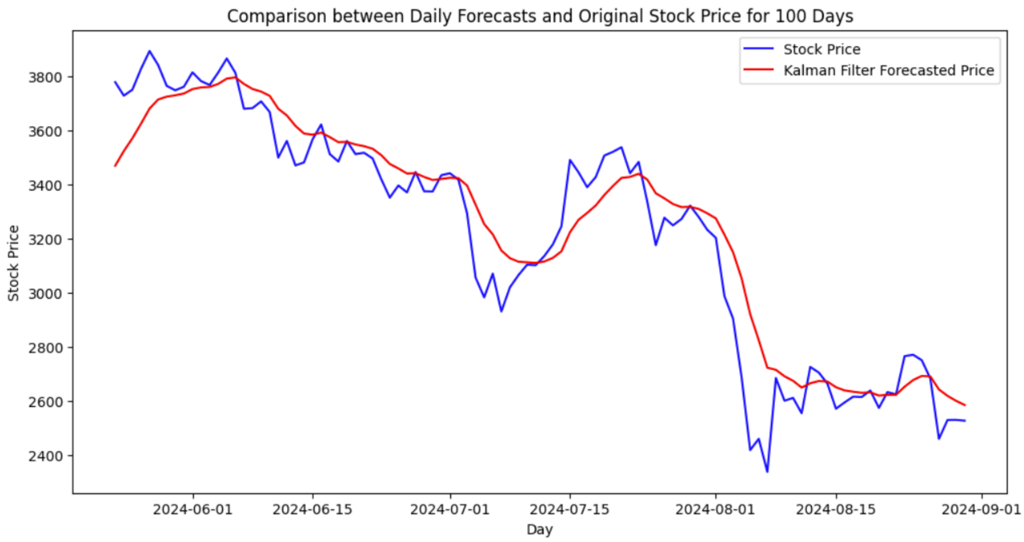


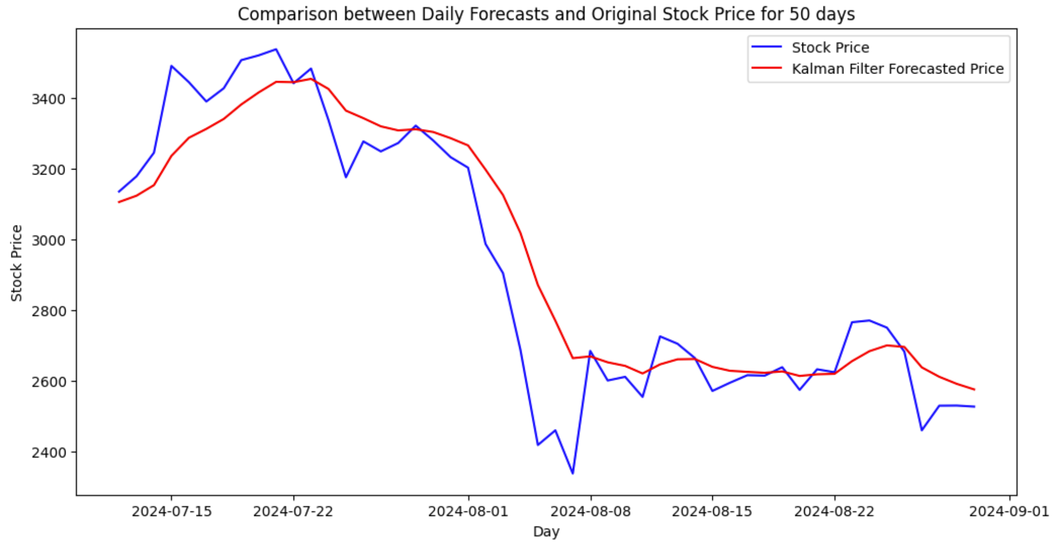
Kalman forecast is able to separate the signal from noise and is smoother than any of the rolling smoothing techniques to understand the trend. Bitcoin and Ethereum have an upward trend and are devoid of any seasonality. Kalman Filter forecasts do not suffer from overfitting compared to the rolling means. This will be clearer in the plots below where we plot the difference between Kalman Mean and stock price vis-a-vis the rolling means. As seen in the below plots the difference between Kalman Mean and stock price increase for the last 200 days.





Kalman forecasts for **“Adj Close”** of **ETH-USD** price have shown varying performance for different time periods. While forecasting we have used the mean and covariance from the observation period. Below we can see the plots for different time durations.





Overall Kalman forecasts do not overfit and follow the stock trend. We can compare the performance of Kalman Forecast for various time intervals with the metrics described in the methodology section.

Kalman Forecast ETH-USD Metrics

Metric	*100 Days*	*50 Days*	*10 Days*
R-Square	**0.90411**	**0.86984**	**0.3092**
MAPE	**0.0310**	**0.03312**	**0.02719**
RMSE	**134.7597**	**135.19253**	**89.1200**
MAE	**95.2399**	**93.3466**	**70.5711**

From the metric table above, we can see that Kalman forecast for a longer term shows better performance compared to shorter intervals like 50 and 10 days.

Linear Regression ETH-USD Metrics

Metric	*100 Days*	*50 Days*	*10 Days*
R-Square	**0.8852**	**0.8508**	**0.4464**
MAPE	**0.02024**	**0.0204**	**0.0125**
RMSE	**64.6897**	**60.526**	**26.6412**
MAE	**45.3442**	**42.9135**	**23.1240**

7. Conclusion and Future Work

Kalman Filter forecast for Adj Close Price prediction for ETH-USD and BTC-USD have show better performance for prediction window of 100 and 50 days compared to Linear

Regression methods. The performance has deteriorated for 10 days window compared to Linear Regression. The Kalman Parameters transition covariance value can be adjusted for more close-fitting predictions. We will explore this in detail in the final submission.

TBD – Comparison with LSTM and Volume prediction details.

8. References:

1. Diego Vallarino. Oct. 2024. arxiv.org. A Dynamic approach to Stock Price Prediction: Comparing RNN and Mixture of Experts Models Across Different Volatility Profiles. https://arxiv.org/pdf/2410.07234
2. Dengxin Huang. Apr. 2023. arxiv.org. Application of Machine Learning in Stock Market Forecasting: A case study of Disney Stock https://arxiv.org/pdf/2401.10903
3. Qiang Li, Ranyang Li. Kaifan Li. 2015. Kalman Filter and Its Applications. https://www.researchgate.net/publication/305871722_Kalman_Filter_and_Its_Application
4. Claudio Urrea, Ramon Agromonte. Sep. 2021. doi.org. Kalman Filter: Historical Overview and Review of Its Use in Robotics 60 Years after Its Creation https://onlinelibrary.wiley.com/doi/10.1155/2021/9674015
5. Opeyemi Sheu Alamu, Md Kamrul Siam. Sep. 2024. arxiv.org. Stock Price Prediction and Traditional Models.: An Approach to Achieve Short-, medium- and long-term goals. https://arxiv.org/pdf/2410.07220
6. Ningyi Li, Chennan Ju et al. Nov. 2023. Arxiv.org. Forecasting and Analysis of CSI 300 Daily Index and S&P 500 Index Based on ARMA and GARCH Models. https://arxiv.org/pdf/2312.14162
7. Vikram Krishnamurthy, Christian R. Rojas. Oct. 2024. arxiv.org. Slow Convergence of Interacting Kalman Filters in Word-of-Mouth Social Learning. https://arxiv.org/pdf/2410.08447
8. John Phan, Hung-Fung Chang. Oct. 2024. arxiv.org. Leveraging Fundamental analysis for Stock Price Prediction. https://arxiv.org/pdf/2410.03913
9. Camiel J.J. Beckers, Feye S.J Hoekstra. et al. Oct. 2024. arxiv.org. HiL Demonstration of Online Battery Capacity and Impedance Estimation with Minimal a Priori Parametrization Effort. https://arxiv.org/pdf/2410.03528
10. Duy Thien An Nguyen, Ka Ching Chan, Nov. 2024. International Journal of Information Management Data Insights. Cryptocurrency trading : A systematic mapping study. https://www.sciencedirect.com/science/article/pii/S2667096824000296
11. Jingyang Wu, Xinyi Zhang et. al. Jun. 2024. arxiv.org. Review of Deep Learning Models for Crypto Price Prediction: implementation and evaluation. https://arxiv.org/pdf/2405.11431
12. Micheal A. Lones. Jan. 2024. arxiv.org. How to avoid Machine Learning Pitfalls: A guide for academic researchers. https://arxiv.org/pdf/2108.02497v4
13. Bingqiao luo. Mar. 2024. arxiv.org. When Crypto Economics Meet Graph Analytics and Learning. https://arxiv.org/pdf/2403.06454v1
14. Vitalik Buterin. 2014. ethereum.org. A Next-Generation Smart Contract and Decentralized Application Platform. https://ethereum.org/content/whitepaper/whitepaper-pdf/Ethereum_Whitepaper_-_Buterin_2014.pdf
15. Mario Filho. Apr. 2023. forecastegy.com. Kalman Filter for Time Series Forecasting in Python. https://forecastegy.com/posts/kalman-filter-for-time-series-forecasting-in-python/

9. Appendix