Predicting prices of decentralized finance cryptocurrency

Gaurav Gade[1]

Harrisburg University of Science and Technology

Abstract

This study reveals a detailed data analysis using multiple data analytics models like time series and machine learning to predict the price of Decentralized finance tokens Cardano, ChainLink, Polkadot and Solana in addition to Bitcoin. The predictive power of the models is tested using both statistical model and advanced machine learning models while considering the time series, seasonal trends, based on closing prices, high and low daily prices, and adjusted price along with monthly, quarterly and weekly prices. The ML techniques have been optimized with the addition of variables, layers and using other sophisticated algorithms like Long term Short term memory. PCA and Clustering analysis have been used to study the variance – covariance between the 5 tokens.

Keywords: DeFI, Cryptocurrency, Price prediction, Time-series, LSTM, Sentiment

Introduction

Decentralized Finance is a relatively new concept in crypto where users can borrow or lend tokens in order to earn interest or open new position. A majority of research studies and experiments have been conducted on top cryptocurrencies like Bitcoin and Ethereum. However, there has not been a conclusive study of price prediction for Decentralized Finance tokens which will be the main focus of my research. The techniques and methods that researchers have used to analyze Bitcoin and Ethereum can be expandedto Decentralized finance tokens which is why it is important to perform an in-depth study the results of these models. Accordingly, I performed a literate survey of over 30 articles published in established journals to prepare the foundation of my research study.

The first part of the paper covers the initial problem statement and hypothesis. The main goal here is to understand if data science models can be utilized to predict the prices of decentralized finance cryptocurrencies. Decentralized cryptocurrencies act like a bank where users can borrow and lend money. However, there is no middle-man or human interaction. There is a contract that specifies the terms of the lending or borrowing process. This includes a staking process where your cryptocurrency is staked (loaned) on the platform to earn interest specified by the platform usually in the form of an Annual Percentage Yield (APY). The most important decentralized finance tokens are Cardano, Solana, Polkadot and ChainLink. Therefore, I have considered these four tokens as part of my analysis. The data was sourced from Yahoo finance. The date range covers October 2020 to October 2021. The data in csv format was converted to time series. The main features are opening price, closing price, daily high, daily low, volume and market cap. The data was then merged after removing any NA or null values. For purpose of comparison, the price values was normalized and scaled. Training and test datasets were created.

For data exploration, visualization methods were used in addition to PCA for understanding the correlation between the 5 tokens. Cluster analysis was performed using K-means and hierarchical clustering. Cluster analysis will help us to identify whether the stocks are correlated or not and the extent to which they are correlated. This could help in price prediction.

For forecasting, I have used time-series modeling using random forests. Random Forest are an optimized form of bagged decision trees which splits nodes based on information gain received from each of the random features. This algorithm is robust as it does not get influenced too much by outliers and highly correlated variables. For example, market cap is a function of price multiplied by number of coins in circulation. Hence, these two variables would be highly correlated. Random forests are better suited to handle such scenarios.

Finally, recurrent neural networks utilizing long-term short-term memory was used using Keras and Tensorflow packages. LSTM was chosen for this type of analysis, because the cryptocurrency data is in a time-series format and LSTM can capture both long and short term seasonality from time series very well. In addition, LSTM can capture patterns and can handle non-linear relationships between parameters. The study reveals the price fluctuation and where the price seems to be heading in the future time periods as well as the correlation between these tokens that can be indicative of the direction that the market is taking.

Literature Review

Review of traditional statistical methods: Bayesian Regression and GLM/Random forests can be used to predict the price of Bitcoin. (Velankar, Valecha, & Maji, 2018). To run the model, the data needs to be segmented into consecutive intervals of sizes 180s, 360s and 720s. Then k-means clustering can be applied to determine the best clusters. Using Bayesian regression, the second set of prices to calculate the corresponding weights of features can be calculated. Price change at a specific time interval for time-series data can be used by running GLM/Random forest on two separate time series data to get two linear models which can then be used to predict the price change. The GARCH-MIDAS model allows us to link the daily observations on stock returns with macroeconomic variables, sampled at lower frequencies, to determine the impact on the stock volatility. (Hou, Asgharian, & Javed, 2013). Another study used this method for predicting the volatility the crypto market. According to (Fang, Su, & Yin, 2020), the study provides new evidence about the impact of the volatility in crypto. This paper investigates how the price of cryptocurrency is impacted by NVIX or News based implied volatility. Using the GARCH-MIDAS model, it was noted that NVIX has a significant negative effect on the price. This leads us to believe that investors in Cryptocurrency get affected by negative news about crypto. In the paper, (Ma, Liang, Ma, & Wahab, 2020), authors have predicted the accuracy of the realized variance of Bitcoin. They concluded that the novel MRS-MIDAS model showed an improved accuracy in the forecasting of the RV of Bitcoin for 2-week and 1 month time intervals. In this paper, (Conrad, Custovic, & Ghysels, 2018) have used Long term and Short term volatility to predict Bitcoin volatility. Using the GARCH-MIDAS model on S&P 500 companies, they inferred that volatility to have a negative and highly significant effect on Bitcoin. S&P 500 risk volatility has positive effect on Bitcoin's volatility that can be used to improve long-term forecasts. Per (Chen, Li, & Sun, 2020)), it is possible to predict Bitcoin price using traditional statistical methods like Regressions and LDA. However, the study included ML algorithms and based on the results, the researchers concluded that highly specialized ML techniques were better at performing

price prediction. Lastly, the paper - (Caporale & Plastun, 2019) studied the effect of certain days in cryptocurrency using traditional statistical tests including Student's t-test, ANOVA, Kruskal-Wallis and regression. The result determined that for Bitcoin, Mondays are statistically likely to provide higher returns than other days of the week. No effect was present for other cryptocurrency

Review of sophisticated algorithms like ML and Neural networks. According to (Chen, Li, & Sun, 2020)), using ML methods like random forests, XGBoost, Quadratic LDA, and SVM by using high-dimension features including property, network, trading, market, attention and gold spot price. The Logistic Regression and Linear Discriminant Analysis for Bitcoin daily price prediction with high-dimensional features achieved an accuracy of 66 while achieved 67.2% accuracy for the ML methods. Based on the findings, statistical methods proved to be an easy way to analyze time-series data. However, there are other algorithms that outperformed these traditional methods.

Since the data used to analyze cryptocurrency is in the form of time-series, I researched articles that used the time-series model to predict prices. (Kumar & Rath, 2020) has predicted the trends of price for Ethereum based on deep learning techniques particularly time-series. ARIMA model was found to be one of the easiest and effective machine learning algorithms where time-series data is involved. (Ji, Kim, & Im, 2019) However, ARIMA model for BTC price prediction resulted in large MSE values. Even so, it can be used for price prediction in sub-periods of the timespan by dividing the timespan over smaller subsets. Deep learning techniques such as multi-layer perceptron (MLP) and long short-term memory (LSTM) can help in predicting the price trends of Ethereum. Before building complex models, it is worth-while to check whether the inbuilt models available in Python and R libraries could be used for the prediction. Scikitlearn and Keras libraries have powerful in-built methods for price prediction of Bitcoin at 1 minute intervals.

Sci-kit, the 'Theil-Sen Regression' method and 'Huber Regression' method had a MSE of 0.000375 and 0.000373 respectively while the R2 was 99.2%. (Phaladisailoed & Numnonda, 2018). For

deep learning based regression models, Keras library can be used to create LSTM and GRU models with MSE 0.000431 and 0.00002 respectively. The R2 was 99.2% in both cases. (Phaladisailoed & Numnonda, 2018). Hence, we can utilize these methods to effectively predict prices. Several research papers have used neural networks for prediction. Per (Sin & Wang, 2017) ANN can be used to predict next day price movement (positive or negative) of Bitcoin with an accuracy of 64%. The paper explores the relationship between the features of Bitcoin and the next day change in the price of Bitcoin using an Artificial Neural Network ensemble approach called Genetic Algorithm based Selective Neural Network Ensemble using 5 Multi-Layered Perceptron (MLP). (Azari, 2019) has used time-series data to develop neural networks like recurrent neural networks, convolutional neural networks, and autoregressive integrated moving average (ARIMA) to develop a predictive model for Bitcoin price.

RNN's were found to significantly outperform ARIMA in predictive accuracy. However, according to a majority of articles, it seems that the LSTM models from machine learning were most effective in price prediction. Per experimental results conducted by (Ji, Kim, & Im, 2019) LSTM-based prediction models slightly outperformed the other prediction models for regression problems, DNN-based prediction models performed the best for classification problems. Per (Dutta, Kumar, & Basu, 2020), Sequence model can be applied with a fixed set of exogenous and endogenous factors for forecasting future crypto prices. The ML models like RNN and LSTM have been known to preform better than traditional time-series models. (Dutta, Kumar, & Basu, 2020). (Patel, Tanwar, Gupta, & Kumar, 2020) Utilizes LSTM (Long term Short term memory) a, Gated recurrent unit and Neural networks to predict crypto prices specifically Lite coin and Monero. The algorithm was able to predict with a high degree of accuracy the prices of these two currencies. Per (Pant, Neupane, Poudel, Pokhrel, & Lama, 2018) tweets related to Bitcoin can be fed to RNN model along with historical price to predict the price for next time frame. The accuracy for sentiment classification of tweets in two class positive and negative is found to be 81.39 % and the overall price prediction accuracy using RNN is found to be 77.62%. Per

(Jay et al., 2020) we can utilize a neural network model for the prediction of cryptocurrency. By using a layer based model for the observed feature activation of neural network to simulate the market volatility. The model used was MLP (multi-layer perceptron) and Long-term short memory. These models showed an improved accuracy than deterministic models to predict the price of 3 major currencies like Bitcoin, Ethereum and Lite coin. In (Sebastião & Godinho, 2021), researchers were able to develop strong and flexible techniques to discover predictability of major cryptocurrencies and formulate profitable trading strategies by using samples during the 'bear' market. The trading strategies were developed using the machine learning models and were validated and tested during the market fluctuations allowing the study to have an inference of its true predictions between the validation and test period. The success rates of individual machine learning models and the conclusive positive results with an achievable performance of major cryptocurrencies using model assembling have helped support the study.

Review of sentiment analysis of twitter and other social media platforms. According to the research conducted in (Abraham, Higdon, Nelson, & Ibarra, 2018), Google Trends and tweet volume were found to be highly correlated with price. When prices are falling, twitter sentiment was found to not be an effective indicator. Based on the research in (Stenqvist & Lönnö, 2017), VADER (Valence Aware Dictionary and Sentiment Reasoner) using Random Forest regression can be used to predict twitter sentiment and to identify relationships between different kinds of input. Twitter sentiments can be used in real-time for crypto price prediction. Using a Spark based architecture that can handle large volume of data, sentiment analysis can be performed in real-time to build and utilize a predictive model that can adjust the weights to accurately provide timely insights that can be used to make a decision. (Mohapatra, Ahmed, & Alencar, 2019). Another paper - (Misnik, Krutalevich, Prakapenka, Borovykh, & Vasiliev, 2018) estimates the price of Bitcoin using market data by analyzing social and time factors. Multilayer perceptron was used along with LSTM neural networks. For time-series, LSTM was considered to be the

best approach. (Misnik, (2018)). News and Social media can be used to effectively predict the price fluctuations of Bitcoin, Lite coin and Ethereum. According to (Lamon, Nielsen, & Redondo, 2017), the study utilizes traditional supervised learning algorithms for text based classification. Daily news and social media data was labelled based on price fluctuations. That way price fluctuation can be predicted without having to first predict the sentiment. The model can be used to predict the largest % increase or decrease for Bitcoin and Ethereum. Per (Wołk, 2020) tweet frequency had a high inverse correlation with crypto prices. Bad news especially caused an increase in post/tweet frequency. The study looks at price prediction for crypto currency using SVM, SGD, GBM, MLP Neural network, Least squares linear regression, ADAboost and Bayesian Ridge Regression, Decision Tree and ElasticNet. A hybrid model built on the mean of all the models was used to make the final prediction. In this article (Nizzoli et al., 2020) have discussed about the manipulation of cryptocurrencies by influencers like Twitter, Telegram and Discord and how topic modeling techniques were used to detect the fraud schemes. Major data set of messages was used as samples to find the sham schemes categorized in 2 sets 'pump and dump' and 'Ponzi' majorly inclining towards Telegram accounting to 20% of the total channels. In addition to this, out of total shared Twitter invite links for bot detection 93% were pointing to Telegram channels. The researcher predicts to fight this manipulation and cryptocurrency price abuse by using such records. (Kraaijeveld & De Smedt, 2020) used sentiment analysis approach to predict the price returns of largest cryptocurrencies. He has discussed about Twitter sentiment being used, along with financial data and granger-causality testing to predict the bullishness ratio and thus find the predictive power for some cryptocurrencies. Moreover this approach also led to findings of at least 1-14% of obtained tweets tweeted by 'bot' accounts on Twitter.

Review of Volatility of price and fluctuation. Several studies focused on predicting the volatility rather than price movement. The reason is that volatility based on the direction of price movement could result in huge gains or losses. In order to optimize the portfolio for risk mitigation, it is

important to understand how volatile the market it. In this research paper, (Catania, Grassi, & Ravazzolo, 2018) have studied the conditional volatility for Bitcoin, Ethereum, Lite coin and Ripple. They determined that the volatility in the crypto market was found to be similar in price movement to other financial time-series like foreign exchange returns. They were able to determine the effect of accounting for long-memory in the volatility process. In another study by (Yi, Xu, & Wang, 2018), researchers attempted to study the correlation of cryptocurrencies using static and dynamic volatility connectedness. The technique used was LASSO-VAR for estimation. 52 cryptocurrencies were found to be tightly interconnected. Finally, in the article, (Walther, Klein, & Bouri, 2019), the researchers have used a mixed data sampling approach to forecast the volatility of highly capitalized cryptocurrencies and its index. They have discussed the external factors driving this volatility major being Global Real Economic Activity factor which has been proved effective for the market fluctuations. Moreover, the derived low loss functions due to average forecasting shows us that the factors are time varying and that the model averaging approach broadens that approach.

Review of DEFI - decentralized finance. DEFI is a new way of borrowing or lending money in crypto without having to go to traditional banking methods like banks, brokerages or exchanges. DeFI can also provide good returns on investment. However, the DEFI space is still in its infancy and susceptible to risks. In the article, (Gudgeon, Perez, Harz, Gervais, & Livshits, 2020), researchers identify two weaknesses in DeFi protocols that are susceptible to price abuse in Crypto. First being over collateralization, author discusses how DeFi lending protocols' flexibility can be measured during the fall of its assets and show the speed at which DeFI protocol would become under Collateralized. Second being a governance attack, the author discusses how attacker is able to steal all the collateral. In this paper, (Corbet, Goodell, Gunay, & Kaskaloglu, 2021) found out strong correlation between Bitcoin and DeFi tokens. Ethereum or Bitcoin do not cause bubbles in the DeFI market according to their research. Mainly, the cryptocurrency Chain link and Maker are the main

drivers. Therefore, while perform price predictions on DeFI, they suggest treating DeFI tokens as separate asset classes.

Methods

In this study, the main topic is to analyze DeFI cryptocurrency. I started with statistical modeling tests on the top 4 Decentralized Finance tokens - ADA, SOL, LINK and DOT in addition to Bitcoin. Based on the literature survey, it is evident that neural networks and Machine Learning algorithms like long-term short-term memory models outperform other techniques to analyze crypto market. In addition to these models, I used time-series modeling and built random forest for the purpose of forecasting. I used in-built functions and libraries from R as well as some core libraries like Keras and Tensorflow as these have proved to be extremely effective in price prediction and easy to build and manage. [@phaladisailoed2018machine]. In addition to the time-series data, LSTM neural networks was implemented using recurrent neural network algorithm. For data preparation, the historic datasets available from sources like Yahoo finance or Coinmarketcap were converted to time-series data for processing. Information without proper actionable insights is not that effective. Therefore, I will be using data visualizations to showcase DEFI prices and predicted values. I have performed a comparative evaluation of the result.

Participants

The main requirement for my analysis is crypto stock data available over the internet. Based on the time-series data available, I can run my predictive models. In addition, the neural networks like convolutional neural network, recurrent neural network are memory intensive programs that may require processing power of a higher magnitude to run in a reasonable amount of time. I am therefore planning to use resources like cloud (Microsoft Azure or Amazon AWS). This will ensure that I can scale my program and deal with large datasets. LSTM models require Keras and Tensorflow libraries to be imported in R.

Procedures

The study would start with statistical modeling tests on Defi tokens like DOT, ADA, SOL and LINK. Then I would be plotting and developing proper visualization of the Defi prices. The data needs to be converted from character format to a time-series so that we can apply LSTM and Random Forest methods on it. Also, data imputation methods would be required as there are some outliers and NA values.

After performing the exploratory data analysis I will be performing clustering analysis and PCA on the data-set. Since this is time-series data, I ran time forecasting methods using random forests. Then, I have built a LSTM neural network to perform a price prediction. I used in-built functions and libraries from Keras and Tensorflow.

Material

To build this solution, I will be utilizing Keras, Sci-kit and Tensorflow libraries, Long-Term Short memory and Neural networks like CNN, RNN. ARIMA, Random Forest and other time-series forecasting models were used. The lubridate library was utilized to create additional features for the time variable (date).

Measures

For time-series modeling, the study will use measures like Autocorrelation values, partial autocorrelation, and forecasting accuracy using measures like MAPE to determine effectiveness. For LSTM and machine learning algorithms, I will be using a split dataset for validation. Accuracy of the model will be used to determine the best performing model. For running PCA, a single data frame will be generated by binding all the taken closing prices into a single data frame. Values were normalized to achieve standardization.

Data analysis

For running the time-series, the following analysis steps were performed:

1. Plotting the autocorrelation to see if there are many lags in the time series.
2. Creating validation test-datasets. Using time-intervals to capture maximum variance and volatility.
3. Building ACF and PACF for determining effect of lags or previous time intervals on price prediction.
4. Running Random Forests.

For LSTM modeling, I performed following steps:

1. Create training and test values, scaling data and assigning training values.
2. Create the model with appropriate parameters like optimizers, epochs.
3. Fit the model to the training data.
4. Assign test and predicted values and then plotting the results.

For PCA, I have performed following steps:

1. Building the covariance matrix using Pearson co-efficient on the data.
2. Using the princomp function to build the PCA model.

For Cluster Analysis, I have performed the following steps:

1. Find optimal number of clusters using elbow method.
2. Calculate the average silhouette distances for 2 to n-1 clusters.
3. Perform the K-Means cluster analysis.
4. Visualize cluster.
5. Build hierarchical cluster.

Analysis/Results

Exploratory Data Analysis:

Data import and preparation: the data for the 5 cryptocurrencies was imported into R, then converted to time series and then normalized. In addition, the column values were converted from string to numeric.

Normalized value of the 5 tokens and plotting the value of $1000 investment from August 2020

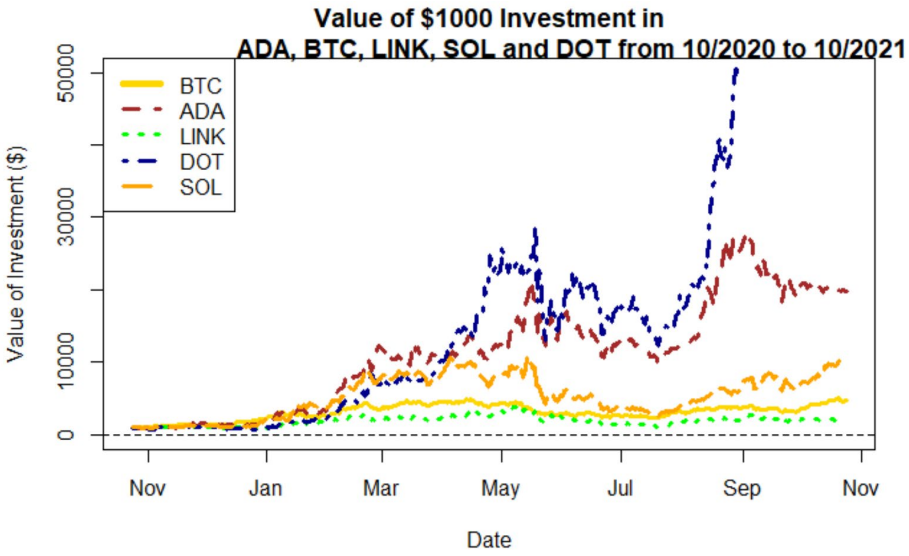
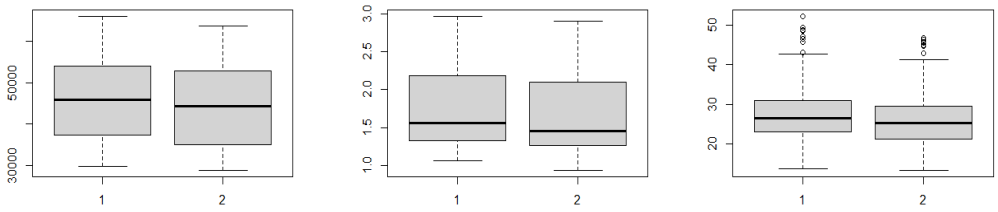


Figure 1: Value of $1000 investment in the 5 token from 10/2020 to 10/2021

As we can see from the above visualization, a $1000 investment would have given significant payoff in 1 year. The best performing asset was Polkadot, Cardano, Solana, and Bitcoin. ChainLink performed the worst among the 5 cryptocurrencies.

Boxplots:



Polkadot Solana

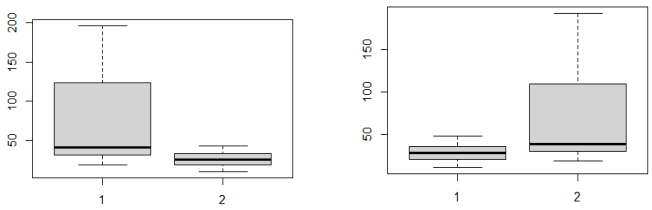


Figure 2: Boxplots for BTC, ADA, LINK, DOT and SOL

From the boxplots above (Figure 3), we can infer that there is less fluctuation between the closing prices for BTC, ADA and Link for the daily High vs daily low price. However, there is significant variance for DOT and SOL.

Correlation of cryptocurrency with BTC:

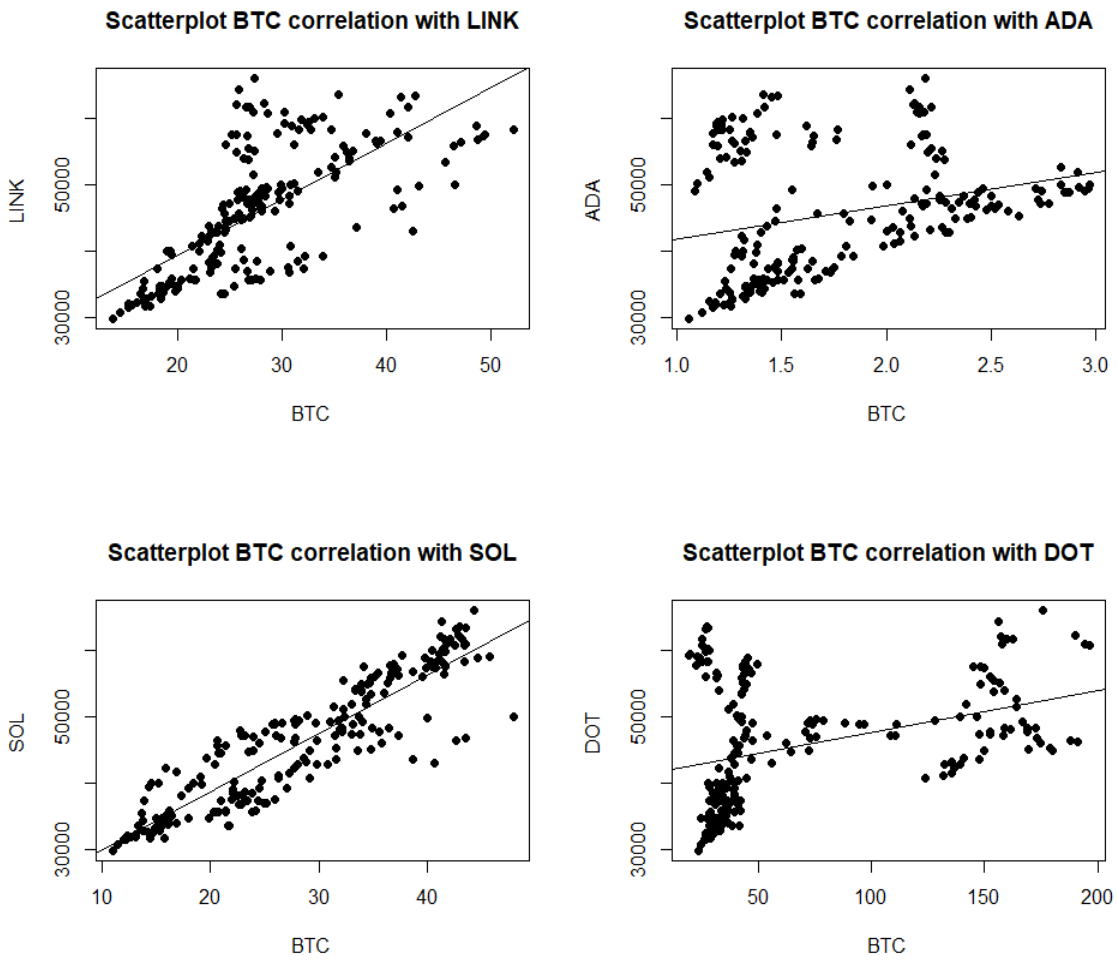


Figure 3: Scatterplots between BTC and other 4 tokens

By looking at the scatterplots in Figure 4, we notice that LINK, ADA and SOL follow BTC price fluctuation closely while it not so much for DOT.

Principal Component Analysis:

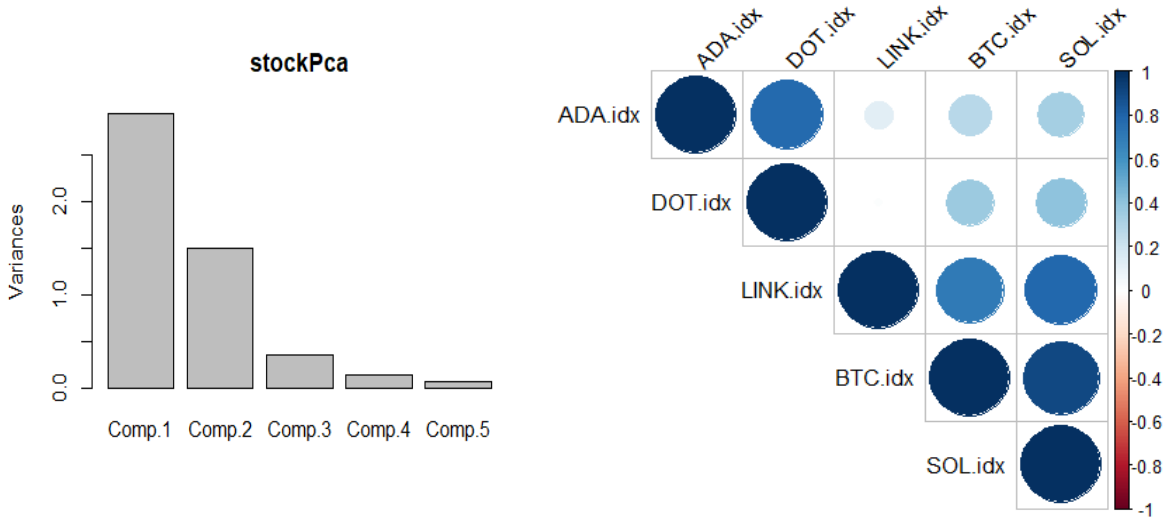


Figure 4: PCA variance and Correlation plot.

We can infer from the PCA analysis and the correlation plots (Figure 5) that Cardano and Polkadot are strongly correlated. Solana and Link are strongly correlated. Link and Bitcoin are moderately correlated. Cardano and ChainLink are not correlated.

Scree plot: Standardized PCA

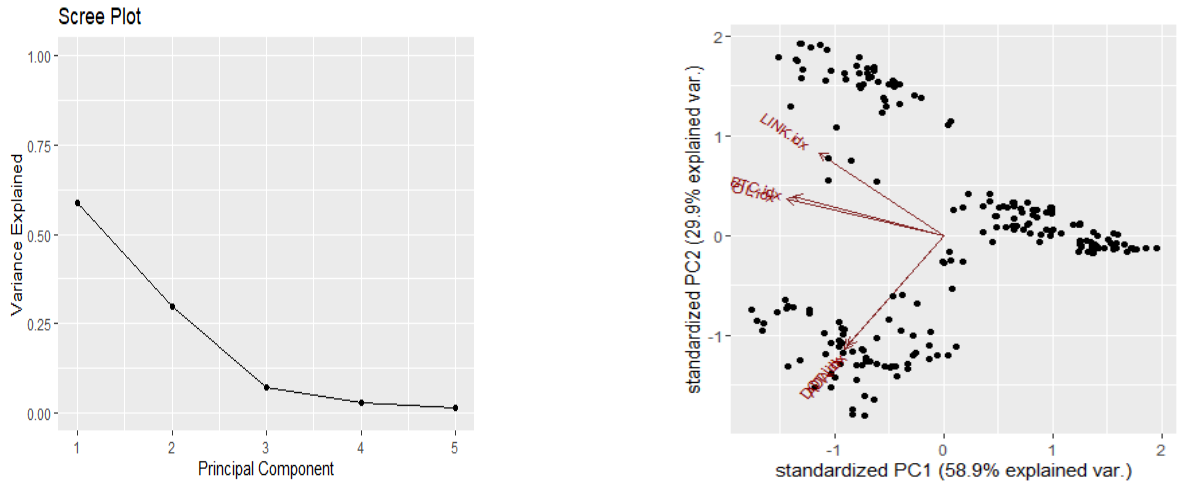


Figure 5: Scree Plot and Standardized PCA

As we can see from the Figure 6, PCA plot, BTC and SOL are correlated. (29.9% of variance explained).

Ideally, the PCA1 component explains .58 of the variance while for PCA2 it is .299.

Therefore, if we look at the cumulative % then 88% of the variance is explained by considering the PCA2 model. This helps us to interpret each Decentralized crypto currency by understanding the magnitude and direction of the coefficients. If the absolute PCA value of the coefficient is high, then its corresponding variable is more important in calculating the component. That will indicate which cryptocurrencies are strongly correlated and which are not. This may explain some of the future price prediction as well as the price variance or fluctuation.

2) Time Series forecasting using Random Forest model

Since the data is a time-series format, it would be worthwhile to run a predictive model. I decided to use a model using the random forest technique as it uses an ensemble technique for classification & regression. The dplyr, caret, forecast, randomForest, TTR and lubridate library were used to build the random forest model. The date and all variables are in character format which needs to be changed to date format and numeric format respectively. To convert the times series data to machine learning, I have created additional features from the date column using lubridate (). The formats used were year, quarter, day, and month. The data was partitioned into training and test data using 80:20 split. Then the model was built trying to predict the closing price of each crypto. For each cryptocurrency, a random forest model was built and evaluated on basis of the test data.

The metric used to evaluate the model is MAPE. Lower the MAPE value, better is the forecasting power of the model.

Results of the 5 forecasting model for each of crypto is as follows:

Cryptocurrency	MAPE – Training	MAPE - Test
ADA	0.78182	3.6201
BTC	0.5531	2.3907
LINK	1.018	1.5099
DOT	1.0727	2.5335
SOL	1.1977	6.2583

Table 1: MAPE values for Training and Test datasets

The values of MAPE for test are not that higher than the training dataset. This way we can concluded that the dataset is not overfitting the model to a large extent. Except for Solana, where the MAPE was slightly higher at 6.2583, the other values were relatively small.

CARDANO

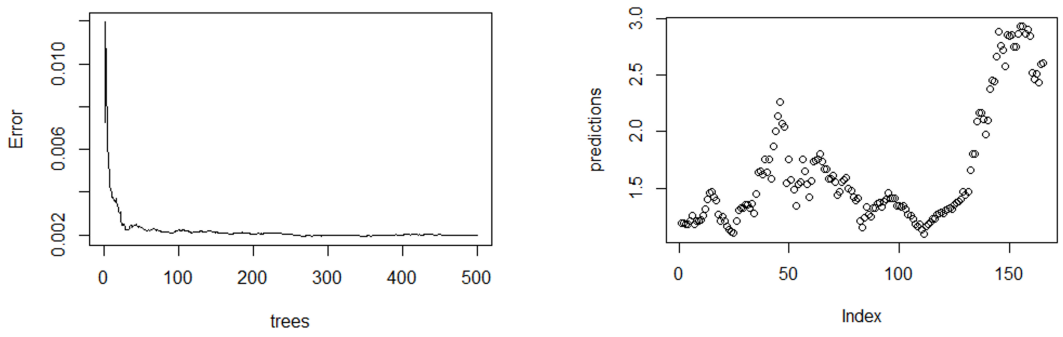


Figure 6: RF Error Curve, Importance of variables, and Predictive Curve

As we can see from the Figure (Cardano) for the random forest model the MAPE is 3.62. We can see that ADA (Cardano) exponentially increases in price during the April/May timeframe – with the value of Cardano reaching $3. The ACF and PACF shows the time series autocorrelation and partial autocorrelation lag. There is a spike for partial correlation at Lag = 10.

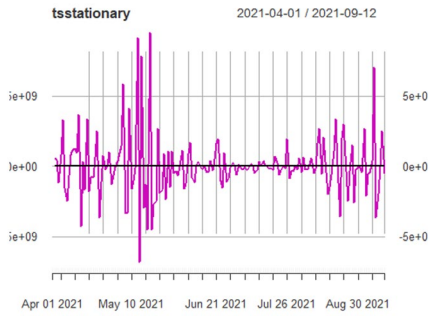


Figure 7: stationary graph for Cardano

The stationary graph shows that there was maximum volatility during the month of May. On May 10 2021, there seems to a huge spike that coincides with the All-time high for Cardano. The other spike is noticeable on August 30 2021. We can therefore predict that every 3 months, there is a major movement in ADA price. The next movement will be therefore be expected end of November 2021 and December 2021.

Random time series for Bitcoin:

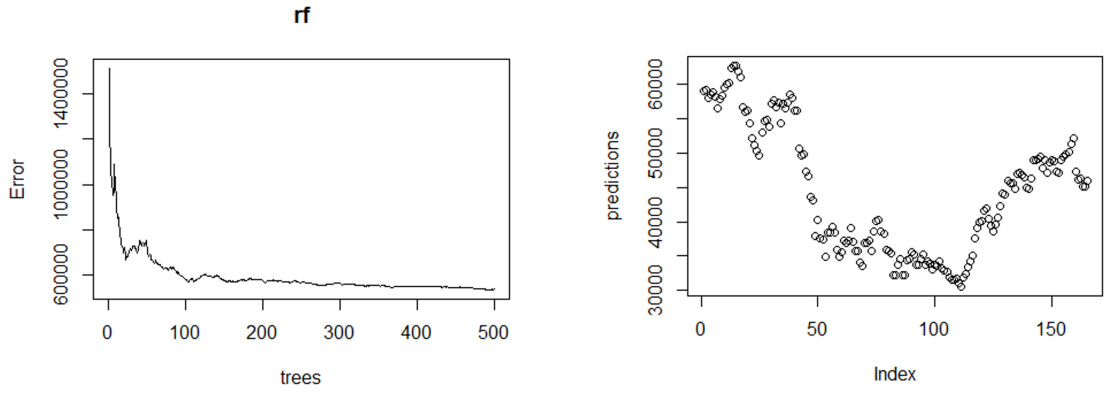


Figure 8: RF Error Curve, Importance of variables, Predictive Curve, ACF and PACF for BTC

For BTC, I observed that the MAPE was 0.5531 for training set and MAPE = 2.3907 for test data. The ACF and PACF shows the time series autocorrelation and partial autocorrelation lag. There is significant autocorrelation lag. From the ACF graph, it is quite evident that BTC prices depends significantly on previous time periods. The stationary graph for BTC shows that there was maximum volatility during the month of May.

LINK:

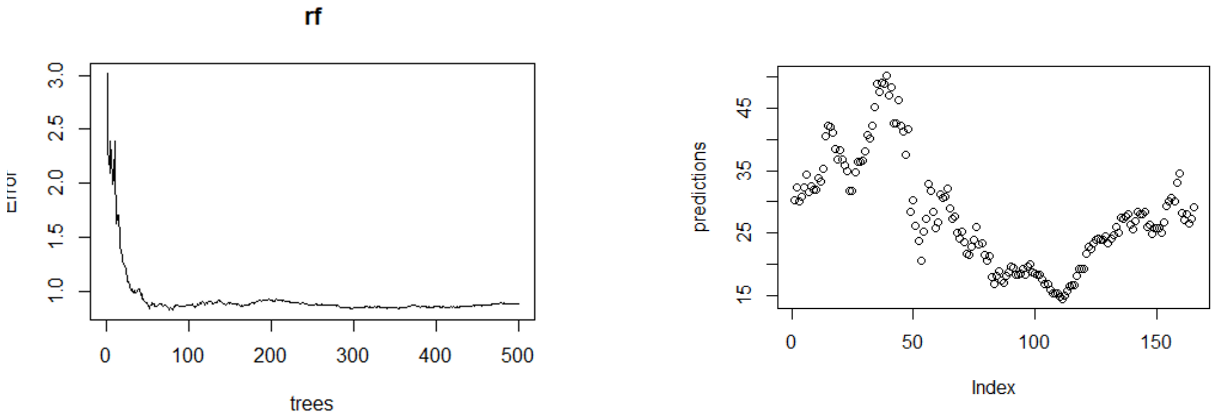


Figure 10: RF Predictions, ACF and PACF for LINK

For LINK, The ACF and PACF shows the time series autocorrelation and partial autocorrelation lag. There is significant autocorrelation lag. From the ACF graph, it is quite evident that LINK prices depends significantly on previous time periods.

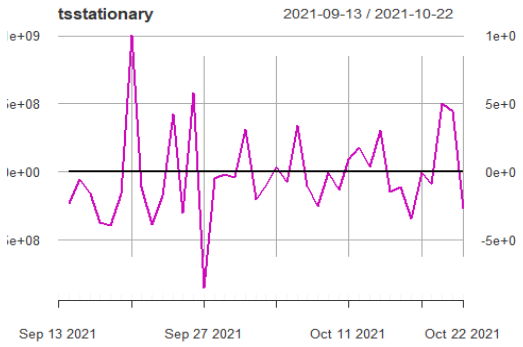


Figure 11: Stationary time series for ChainLink.

The stationary graph for LINK shows that there was maximum volatility during the month of September (27^{th}).

Polkadot:

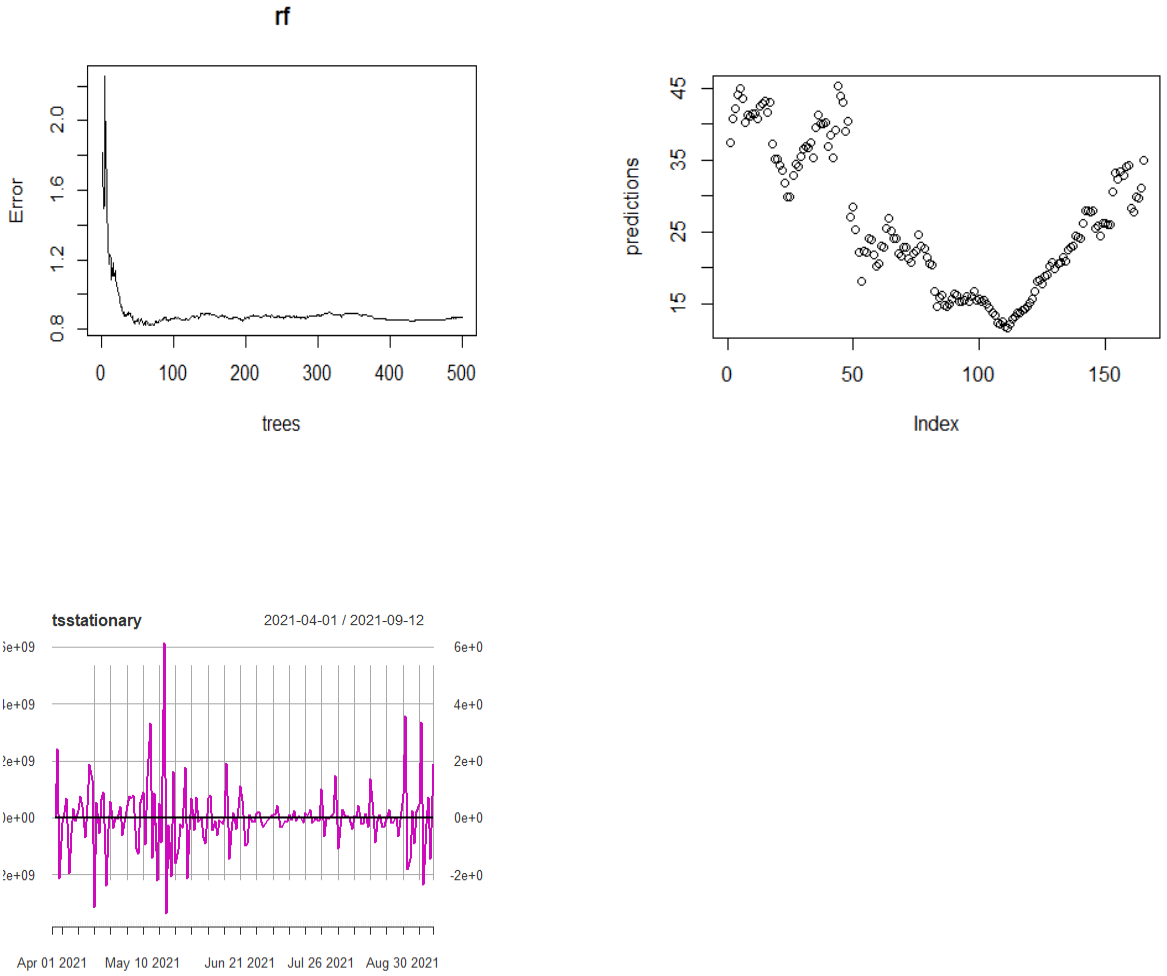
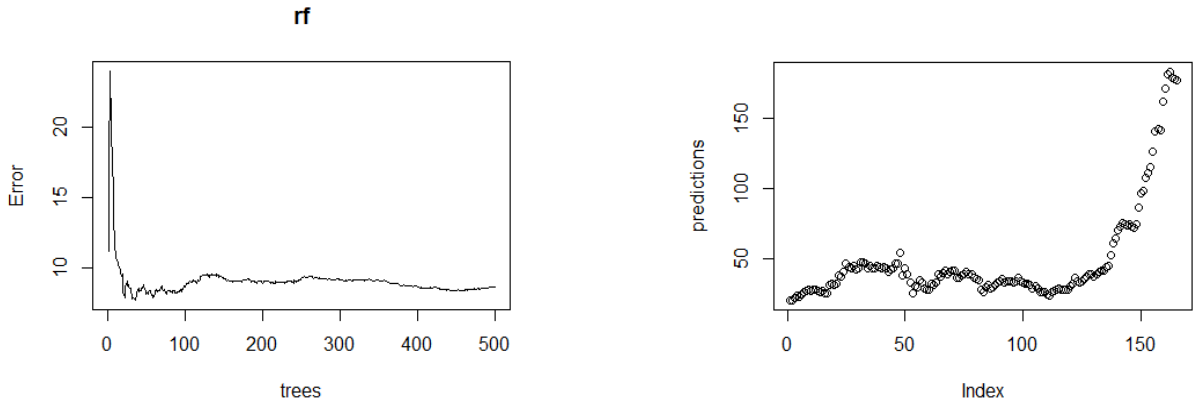


Figure 13: Stationary time series for ChainLink

For Polkadot, the ACF and PACF shows the time series autocorrelation and partial autocorrelation lag.

There is significant autocorrelation lag. From the ACF graph, it is quite evident that DOT prices depends significantly on previous time periods. The stationary graph for DOT shows that there was maximum volatility during the month of May (May 10th) and August (August 30)

SOLANA:



For SOL, I observed that the ACF and PACF shows the time series autocorrelation and partial autocorrelation lag. There is significant autocorrelation lag. From the ACF graph, it is quite evident that SOL prices depends significantly on previous time periods.

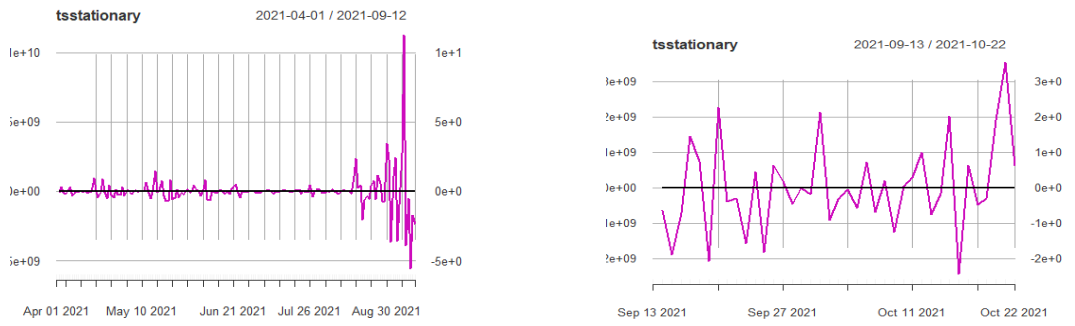


Figure 15: Stationary time series for ChainLink

The stationary graph for BTC shows that there was maximum volatility during the month October

3) Long-term Short memory using recurrent neural networks:

Using Keras and Tensorflow packages in R, I built the LSTM model using recurrent neural networks. LSTM can help in time series forecasting as they involve autocorrelation. From the previous model, I noticed that all 5 tokens are auto correlated. Therefore, using LSTM makes sense to use and it can recognize patterns for the entire duration of the time series.

The scaled data was first converted into a matrix. Then the data was lagged 11 times and arranged into columns each corresponding to 1 month. Therefore, we will look at the lagged value for 12

previous months. Each column in the scaled training set corresponds to a lagged version of the previous one. After converting the data into an array for training, I repeated the same process for the test data set.

Model building: The LSTM model was built using keras_model_sequential function. Layers were added on top of the model. After supplying the LSTM model with the required input shape, activation function, loss function and type of optimizer (Adam optimizer in this case), the model was built.

For ADA, the model result was as follows:

Total parameters were 30,651. 2 dropout layers. For evaluating the model and performing the prediction, the shuffle variable was turned to False.

```
## Model: "sequential"
##
_________________________________________________________________________
___
##  Layer (type)                        Output Shape                    Param
#
##
=========================================================================
===
##  lstm_1 (LSTM)                       (1, 12, 50)                     10400
##
##  dropout_1 (Dropout)                 (1, 12, 50)                     0
##
##  lstm (LSTM)                         (1, 12, 50)                     20200
##
##  dropout (Dropout)                   (1, 12, 50)                     0
##
##  time_distributed (TimeDistributed)  (1, 12, 1)                      51
##
##
=========================================================================
===
## Total params: 30,651
## Trainable params: 30,651
## Non-trainable params: 0
```

Figure 16: LSTM model parameters:

Prediction results for the 4 Decentralized Finance tokens were as follows for Oct 2021 to Oct 2022.

Cardano:

Closing price for Cardano after 10 months in 2022 is expected to be $2.86. Currently, it is $1.60 at the time of writing this paper (Dec 3 2021).

ChainLink:

Over the next 12 months, ChainLink is expected to increase 10x in value with the price going up to $225.86. Currently, it is $25.

Polkadot:

Over the next 12 months, Polkadot is expected to increase 13x in value with the price going up to $423. Currently, it is around $32.

Solana:

Over the next 12 months, Solana is expected to increase 6.5x in value with the price going up to $1389. Currently, it is around $212.

Discussion

Overall, in this paper I have used clustering, PCA, time-series forecasting and Recurrent Neural Networks (using Long Term Short term memory) for the analysis. Based on the results, it can be inferred that while cryptocurrencies may experience short term volatility, in the long-term it is an asset that will rise exponentially. The MAPE of the model was low (<4%) and the p-value was less than 0.05. The time-series forecasting as well as the LSTM neural network method both predicted high gains for all the 5 Decentralized tokens. The MAPE for training and test sets was low (below 4%) in a majority of cases which indicates that the model did not suffer from over-fitting. Predicted values for September and October generally were accurate to what happened in reality in the market.

On basis of the exploratory data analysis, it can be inferred that Polkadot has grown the highest among the 5 tokens over the last year. It can be seen from the line graph that all cryptocurrencies increased in value multiple times over the last 12 months. From the box plots, it can be inferred that BTC, LINK and ADA are good for long-term investments while SOL and DOT are good for both long-term and short-term trading. This is due to the significant variance in daily high and low prices for Solana and Polkadot. From the cluster analysis and PCA, it can be inferred that Cardano and Polkadot are strongly correlated and Solana and ChainLink are also strongly correlated.

On evaluating the time-series model and LSTM, it was predicted that all the cryptocurrencies would increase in value in the next 12 months. The predicted value of ADA was $2.86 which was the lowest growth among all the tokens. It should be noted that it is not even predicted to reach its all-time high price of $3.10 that it reached in 2021. For ChainLink, the prospects seem a lot better as it is forecasted to reach $225 by October 2022. That is

approximately an 8-9x return on investment from its current price ($25 in Dec). For Solana, the forecasted price was $1389. That is extremely impressive as that would mean a 6-7x return on investment from its current price ($220). In October 2020, Solana was trading at $4. If Solana reaches its predicted price in October 2022, it would mean 350x return on investment in just 2 years which is extremely good.

The best performer among the 4 DeFI tokens was Polkadot which is predicted to reach $423 in October 2022. That would mean almost 10-12x the current price (#35).

However, the prediction using this method should by no means be considered as absolute and final. There are several other factors that must be taken into consideration while making an investment. The paper delves purely on past performance of the currency with respect to the financial information available. However, market factors are complex to study and difficult to analyze as real-time trends come into picture. For example, the recent outbreak of the omicron coronavirus variant affected the stock market as well as cryptocurrency market globally as fears grow about the effect of another restriction on travel and daily life. The closure of borders, travel and trade between countries affects supply chain that has compounding effect on all sectors of the economy. Another factor that may show its effect long-term is the passing of several trillion dollar incentive stimulus packages that might increase inflation thus reducing the purchasing power of the USD. Hence, stock and crypto may be a good hedge to protect your savings rather than parking them in a financial institution like a bank. From the results perspective, the best gainer of decentralized cryptocurrency is Polkadot with over 13x gains projected over 12 month period. However, it is advisable that while investing in cryptocurrency the risk may be spread across multiple assets. Decentralized finance has shown tremendous potential and has resulted in massive profits over the last few years. As the concept is relatively new, this space will likely gain further traction as more people become aware of its benefits. For example, if I need to take a $20,000 loan

over a $10,000 collateral, a traditional banking system may require 2 weeks to process the claim. Also, it may require lot of documentation and paperwork. In DeFI, if you own $10,000 in Polkadot or Solana, then you can get the loan of $20,000 in under 2 minutes without any questions asked or signing any paperwork. The ease of borrowing and lending money is attractive to people from all age groups. Hence, we can conclude that the decentralized finance is here to stay for the future and it is recommended to invest at least a small amount of your investment in this space.

One limiting factor that I realized about this paper was that cryptocurrency prediction was considered in its own space (silo). There have been studies that show a correlation between cryptocurrency prices and the stock market or gold prices. If the stock markets tumble, then I have observed that crypto markets also tend to get hit. Whether the hit is long-term or not, is a different issue but there is an impact which could be statistically studied and analyzed using scientific methods. If this correlation can be tied to the predictive algorithm in real-time, then there may be a way to leverage this algorithm for short-term or even day-trading. As we have seen via the box plots of Polkadot and Solana that there is a huge difference in the daily high and low prices which presents an opportunity for short term trading which may even include leveraged trading or shorting the market for making quick profits in case there is a signal that there could be a crash due to any political or economic event taking place globally. Another aspect that can be explored is whether a specific business group or area impacts the crypto market more than others. For example, cryptocurrency is considered to be in technology sector. Tech stocks rising or falling may impact cryptocurrency more than the stocks from the Retail and Consumer goods sector. This is just an assumption and would require careful statistical assessment before such a conclusion can be inferred. However, this is beyond the scope of this paper and can be considered in future studies.

Conclusion

In this paper, I have proposed a clustering method for analyzing the top 4 decentralized finance token namely Solana, Polkadot, ChainLink and Cardano along with the market leader Bitcoin. The clustering method used KNN and hierarchical clustering to find out patterns between these 5 tokens and determine any correlation. Principal component analysis was used (PCA) in the clustering method for detecting the variance-covariance structure for the closing prices of these 5 tokens through linear combinations. By summarizing the result, it is observed that the Standard Deviation for Comp1 is 1.716 while it is 1.223 for comp.2. Comp.1 explains 58.93% of the proportion of variance while Comp.2 explains an additional 29.919%. The correlation plot shows a .8839 correlation between LINK and DOT. The PCA shows BTC and SOL components to be closely aligned.

Random forest technique using time-series was used for predicting the future prices of all the 5 tokens. I have performed ACF and PACF tests on all the 5 tokens to determine the effect of previous time periods. It was concluded that all 5 tokens exhibited high levels of autocorrelation. We also forecasted the prices for the next 12 months for these token and determined that all of the DeFI tokens would be profitable by a large margin. In fact, Polkadot emerged as the clear winner among all the tokens with its estimated price increasing by 12 to 13 times its current value.

Long-Term Short term memory was used as an additional forecasting method that uses recurrent neural network for time series data. This was done because, we found out that all cryptocurrencies were auto correlated and the LSTM method was proven to be extremely effective in prediction of crypto prices according to the literate study.

The LSTM method forecasted the crypto currency prices for the next 12 months from October 2021 to October 2022. The error rate for LSTM was also low (<5%) and like the Random

Forest technique, it predicted exponential rise in the value of all the 5 tokens. As stated in the discussion section, there are many factors at play when applying these algorithms. These factors could be economic, political and environmental.

Market sentiment and trends also play a pivotal role in the price. However, it is hard to quantify these factors and scientifically analyze them as the volume of data generated via social media is extremely high. For example, if we were to include market sentiment in this study, millions of tweets, Reddit posts, Facebook comments, Telegram and WhatsApp messages would have to be analyzed to see if there are any positive or negative indicators for a specific cryptocurrency. Extracting all this information, storing and parsing it and then applying algorithms on it would be a much more complex task. Data preparation would also require tremendous time and effort. However, a good subset of this information can be used and analyzed. While not full-proof, almost all statistical studies are based on a reasonable sample of the total population. By using proper statistical tests and text mining methods, this social media sentiment can be integrated in this study as a future scope. This would require real-time information to be processed as cryptocurrency markets are extremely volatile as compared to traditional stock markets. The integration of the social media aspect with the models described in this paper could help improve the forecast and make it much more reliable.

References

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.

Aust, F., & Barth, M. (2020). *papaja: Prepare reproducible APA journal articleswith R Markdown*. Retrieved from https://github.com/crsh/papaja

Azari, A. (2019). Bitcoin price prediction: An ARIMA approach. *arXiv Preprint arXiv:1904.05315*.

Barth, M. (2021). *tinylabels: Lightweight variable labels*. Retrieved from https://github.com/mariusbarth/tinylabels

Caporale, G. M., & Plastun, A. (2019). The day of the week effect in thecryptocurrency market. *Finance Research Letters*, *31*.

Catania, L., Grassi, S., & Ravazzolo, F. (2018). Predicting the volatility of cryptocurrency time-series. In *Mathematical and statistical methods for actuarialsciences and finance* (pp. 203–207). Springer.

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.

Conrad, C., Custovic, A., & Ghysels, E. (2018). Long-and short-term cryptocurrency volatility components: A GARCH-MIDAS analysis. *Journal of Risk and Financial Management*, *11*(2), 23.

Corbet, S., Goodell, J. W., Gunay, S., & Kaskaloglu, K. (2021). Are DeFi tokens aseparate asset class from conventional cryptocurrencies? *Available at SSRN 3810599*.

Dutta, A., Kumar, S., & Basu, M. (2020). A gated recurrent unit approach to bitcoin price prediction. *Journal of Risk and Financial Management*, *13*(2), 23.

Fang, T., Su, Z., & Yin, L. (2020). Economic fundamentals or investor perceptions?

The role of uncertainty in predicting long-term cryptocurrency volatility *I*nternational Review of Financial Analysis, 71, 101566.

Gudgeon, L., Perez, D., Harz, D., Gervais, A., & Livshits, B. (2020). The decentralized financial crisis: Attacking defi. *arXiv Preprint arXiv:2002.08099*.

Hou, A., Asgharian, H., & Javed, F. (2013). Importance of the macroeconomic variables for variance prediction: A GARCH-MIDAS approach. *Journal of Forecasting*, *612*(July), 1–29.

Jay, P., Kalariya, V., Parmar, P., Tanwar, S., Kumar, N., & Alazab, M. (2020).

Stochastic neural networks for cryptocurrency price prediction. *IEEE Access*, *8*,82804–82818.

Ji, S., Kim, J., & Im, H. (2019). A comparative study of bitcoin price predictionusing deep learning. *Mathematics*, *7*(10), 898.

Kraaijeveld, O., & De Smedt, J. (2020). The predictive power of public twitter sentiment for forecasting cryptocurrency prices. *Journal of International Financial Markets, Institutions and Money*, *65*, 101188.

Kumar, D., & Rath, S. (2020). Predicting the trends of price for ethereum usingdeep learning techniques. In *Artificial intelligence and evolutionary computations in engineering systems* (pp. 103–114). Springer.

Lamon, C., Nielsen, E., & Redondo, E. (2017). Cryptocurrency price predictionusing news and social media sentiment. *SMU Data Sci. Rev*, *1*(3), 1–22.

Ma, F., Liang, C., Ma, Y., & Wahab, M. (2020). Cryptocurrency volatilityforecasting: A markov regime-switching MIDAS approach. *Journal of Forecasting*, *39*(8), 1277–1290.

Misnik, A., Krutalevich, S., Prakapenka, S., Borovykh, P., & Vasiliev, M. (2018). *N*eural network approximation precision change analysis on cryptocurrency price prediction.

Mohapatra, S., Ahmed, N., & Alencar, P. (2019). KryptoOracle: A real-time cryptocurrency price prediction platform using twitter sentiments. *2019 IEEEinternational conference on big data (big data)*, 5544–5551. IEEE.

Nizzoli, L., Tardelli, S., Avvenuti, M., Cresci, S., Tesconi, M., & Ferrara, E. (2020).Charting the landscape of online cryptocurrency manipulation. *IEEE Access*, *8*,113230–113245.

Pant, D. R., Neupane, P., Poudel, A., Pokhrel, A. K., & Lama, B. K. (2018).

Recurrent neural network based bitcoin price prediction by twitter sentiment analysis. *2018 IEEE 3rd international conference on computing, communication and security (ICCCS)*, 128–132. IEEE.

Patel, M. M., Tanwar, S., Gupta, R., & Kumar, N. (2020). A deep learning-based cryptocurrency price prediction scheme for financial institutions. *Journal of Information Security and Applications*, *55*, 102583.

Phaladisailoed, T., & Numnonda, T. (2018). Machine learning models comparison for bitcoin price prediction. *2018 10th international conference on information technology and electrical engineering (ICITEE)*, 506–511. IEEE.

R Core Team. (2020). *R: A language and environment for statistical computing*.

Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Sebastião, H., & Godinho, P. (2021). Forecasting and trading cryptocurrencies with machine learning under changing market conditions. *Financial Innovation*, *7*(1), 1–30.

Sin, E., & Wang, L. (2017). Bitcoin price prediction using ensembles of neural networks. *2017 13th international conference on natural computation, fuzzy systems and knowledge discovery (ICNC-FSKD)*, 666–671. IEEE.

Stenqvist, E., & Lönnö, J. (2017). *Predicting bitcoin price fluctuation with twitter sentiment analysis*.

Velankar, S., Valecha, S., & Maji, S. (2018). Bitcoin price prediction using machine learning. *2018 20th international conference on advanced communication technology (ICACT)*, 144–147. IEEE.

Walther, T., Klein, T., & Bouri, E. (2019). Exogenous drivers of bitcoin and cryptocurrency volatility–a mixed data sampling approach to forecasting. *Journal of International Financial Markets, Institutions and Money*, *63*, 101133.

Wołk, K. (2020). Advanced social media sentiment analysis for short-termcryptocurrency price prediction. *Expert Systems*, *37*(2), e12493.

Yi, S., Xu, Z., & Wang, G.-J. (2018). Volatility connectedness in the cryptocurrencymarket: Is bitcoin a dominant cryptocurrency? *International Review of Financial Analysis*, *60*, 98–114.

Appendix

```
data.DOT[c(1:3,nrow(data.DOT)),]

           DOT.Open    DOT.High    DOT.Low     DOT.Close   DOT.Adjusted
2020-10-24 "4.266853"  "4.430925"  "4.244643"  "4.355344"  "4.355344"
2020-10-25 "4.355344"  "4.447240"  "4.267938"  "4.335231"  "4.335231"
2020-10-26 "4.335231"  "4.738708"  "4.304171"  "4.727016"  "4.727016"
2021-10-24 "44.050823" "44.177582" "43.305199" "43.571114" "43.571114"
```

Figure 18

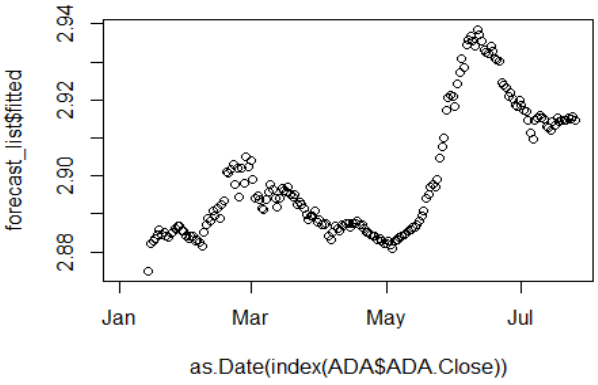


Figure 19: Forecasted price from LSTM for Cardano.

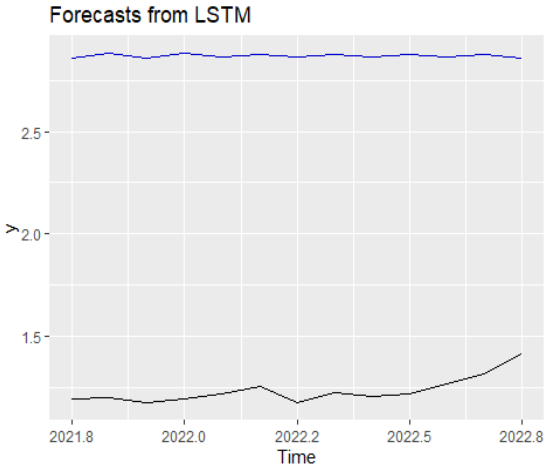


Figure 20: Forecasted price from LSTM for Cardano.

CARDANO

RF: Importance of Variables

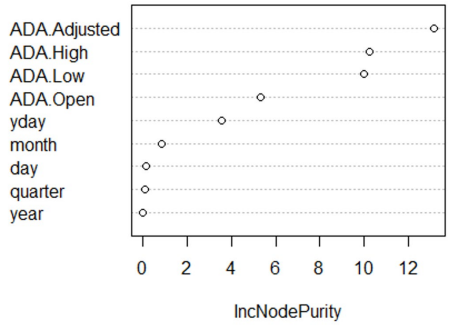


Figure 21: Importance Of variables by node Purity.

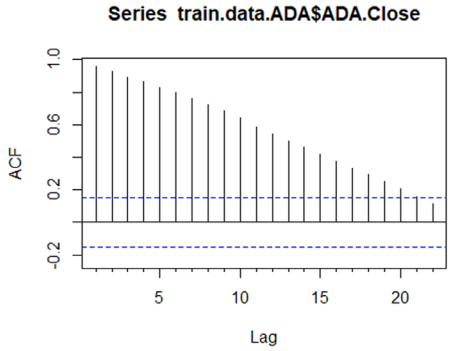


Figure 22: ACF for Cardano Closing price

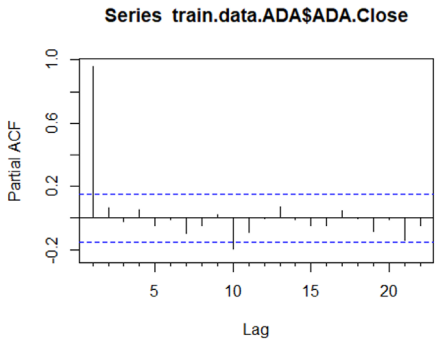


Figure 23: PACF for Cardano.

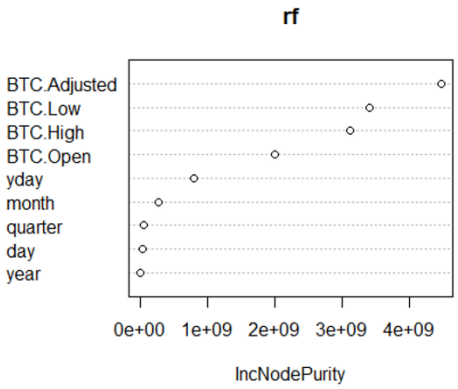


Figure 24: Importance Of variables by node Purity

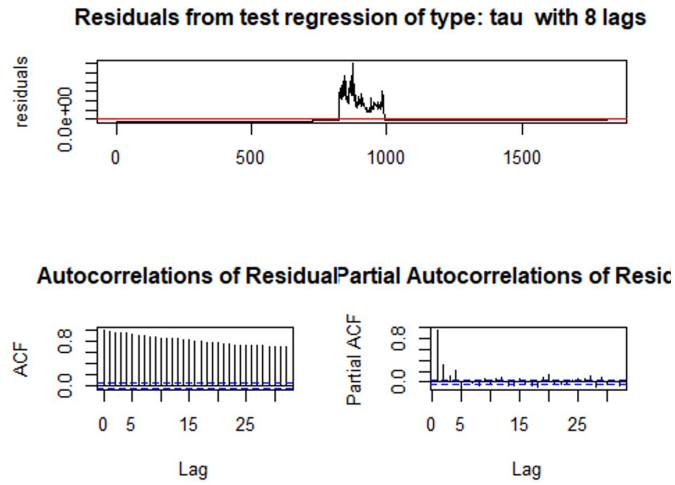


Figure 25: ACF and PACF for BTC Closing price

Link:

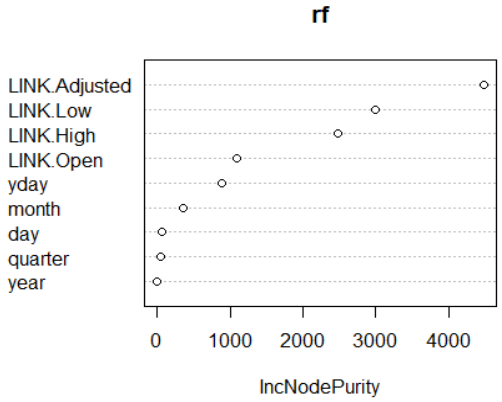


Figure 26: Importance Of variables by node Purity

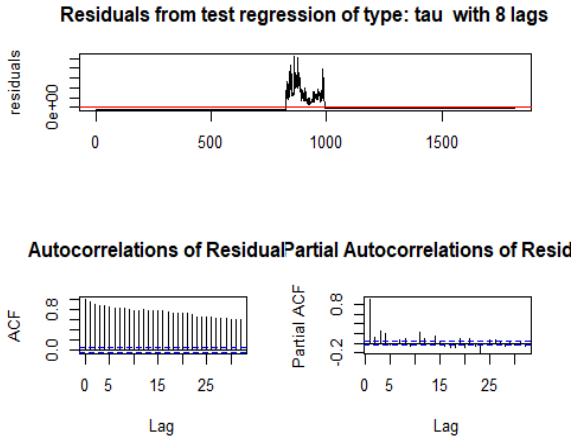


Figure 27: ACF and PACF for LINK.

Dot:

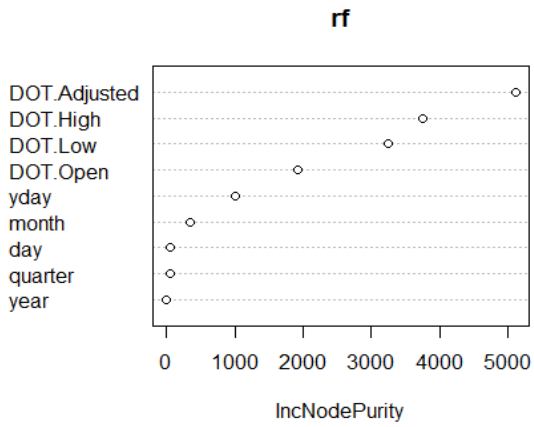


Figure 28: Importance Of variables by node Purity

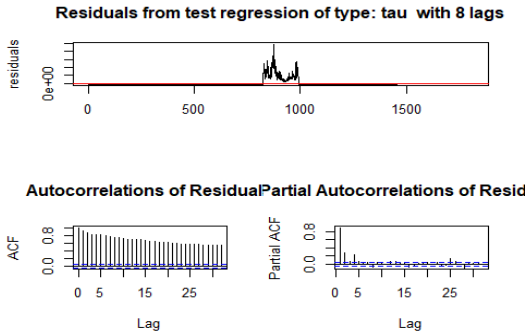


Figure 29: Figure 27: ACF and PACF for DOT.

Solana:

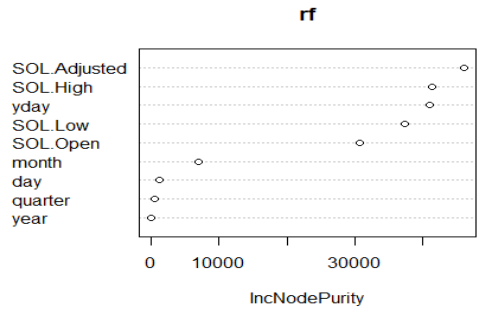


Figure 30: Importance Of variables by node Purity

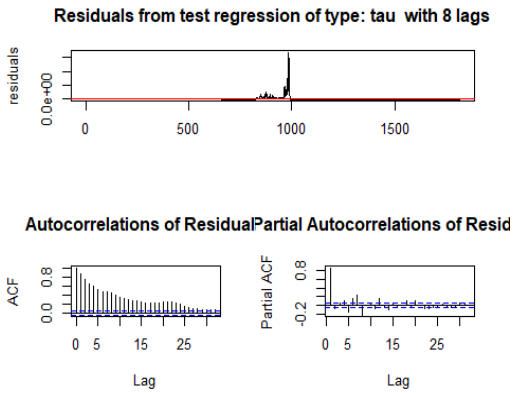


Figure 31: ACF and PACF for SOL.