Karthikeyan Arumugam
99265685

A study on predicting prices of different Cryptocurrency using Machine Learning Algorithms

Capstone Final Project

Table of Contents

1.Abstract

Nowadays, everyone has a desire to invest and earn an extra income outside of the regular day-to-day work schedule and for a while, investing in government bonds and stocks were the only way to achieve that. One of the most basic lessons in finance is the risk and reward are directly proportional, i.e., the greater the risk, the greater the reward. More recently, the spotlight has shone on cryptocurrency --which is a relatively more lucrative area of FinTech. Cryptocurrency is a new-age digital, decentralized and encrypted medium of exchange that is not governed by any form of institution which makes it more desirable as an asset. While cryptocurrency cannot replace fiat money, it alleviates the barriers of the exchange rates among national currencies. Technological advancement is uniformly influencing all parts of the economy; thereby also influencing the global markets to add cryptocurrency as another industry for people to invest in-- besides company stocks. In this project, I aim to explore the forecasting short-term closing prices of each of the six different cryptocurrency companies and compare the predicted price to the actual price using appropriate machine learning algorithms.

The dataset I chose for this project is obtained from Kaggle Inc. website, which originally contains twenty-three csv files for each of the twenty-three different crypto currencies. However, I chose to save only parts of csv files related to the six crypto currencies I aim to analyse.
The goal of this project is to apply time-series with machine learning aspects- more specifically Long Short-Term Memory (LSTM) model on each of the six datasets and find out which one of them is profitable in the short-term trade.

2.Introduction

As aforementioned, the novelty of cryptocurrency price prediction has contributed to the lack of many research papers on the subject, thereby also encouraging finance and data enthusiasts to experiment on different deep machine learning models that could apply to market data. When traditional time series methods such as AutoRegressive Integrated Moving Average (ARIMA) are applied to crypto currencies, it fails to grasp the limitations due to the non-linear and non-stationary patterns in the cryptocurrency data[3][4].

Despite the newness of the subject, researchers agree that frequent and high volatility of the prices causes an underlying chaos using traditional machine learning methods, thereby forcing Deep Learning Methods to come into play. Most review papers I came across suggested some DL methods are listed in Table 1. However, a few from the list go beyond the scope of this project due time and skill constraints.

3.Literature Review

As aforementioned, the novelty of cryptocurrency price prediction has contributed to the lack of many research papers on the subject, thereby also encouraging finance and data enthusiasts to experiment on different deep machine learning models that could apply to market data. When traditional time series methods such as AutoRegressive Integrated Moving Average (ARIMA) are applied to crypto currencies, it fails to grasp the limitations due to the non-linear and non-stationary patterns in the cryptocurrency data[3][4].

Despite the newness of the subject, researchers agree that frequent and high volatility of the prices causes an underlying chaos using traditional machine learning methods, thereby forcing Deep Learning Methods to come into play. Most review papers I came across suggested some DL methods are listed in Table 1. However, a few from the list go beyond the scope of this project due time and skill constraints.

Methods mentioned in different papers
Long Short-Term Memory (LSTM)
Recurrent Neural Networks(RNN)
Regression-based
Support vector machines (SVM)
Tree-based

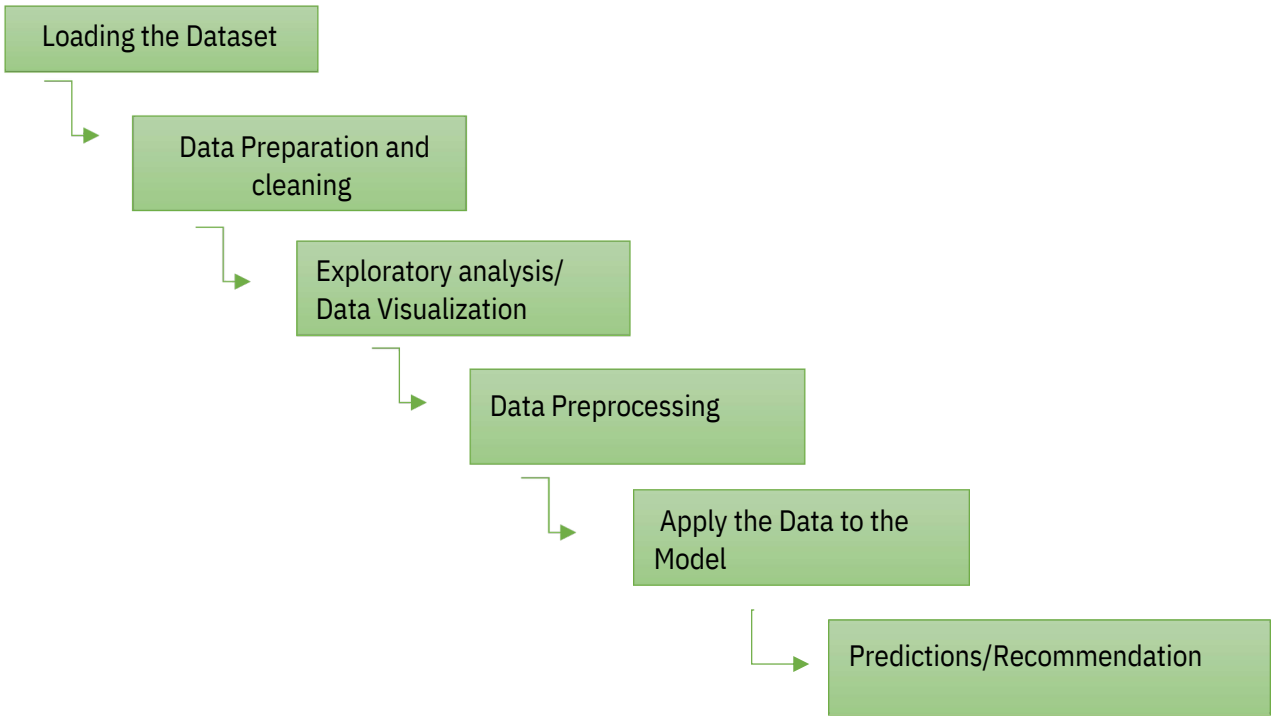
Artificial neural network (ANN)
Stacked artificial neural network (SANN)

The goal of most of the papers was to determine which of the listed DL methods is the most accurate to forecast the price of a cryptocurrency by using classification methods i.e., k-folds, Random Forest Trees to compare each and find out which classifies as the most accurate.

The next similar aspect of recent studies shows the analysis done mostly on BTC, it being the longest active currency to conduct analysis on. As mentioned in the introduction, it is the most popular.

The cited works focus on how each of the methods differ performance levels on the different time intervals of the prices. The papers differ between daily, weekly, and monthly intervals of opening and closing prices. Besides forecasting prices based on historical data, the related works dwelled into researchers also releasing heavily on sentiment analysis. According to a few sources, researchers aimed to find the frequency of “Bitcoin” in tweets and if it related to the predictability of the process of the BTC, probably using Natural Language Processes (NLP) methods which will not be factored into this project.

4.Approach Overview



5. Overview of the Data

The dataset used was obtained from Kaggle Inc and is publicly available under Cryptocurrency Price History [5]. There are 23 csv files that contain historical prices for 23 different cryptocurrencies but the ones this project focuses on are those of Bitcoin (BTC), Ethereum (ETH), Litecoin (LTC), Dogecoin (DOGE), Cardano (ADA), and Binance Coin (BNB). Kaggle makes csv files for each of these cryptocurrencies individually,
however, I merged all the individual files into one csv file saved under crypto_ market.csv.

It was simple to merge all these files into one because each individual file has the same columns: Serial Number, Name, Symbol, Date, High, Low, Open, Close, Volume, Market Capitalization.

The initial dataset makes us focus on the non-uniform dates of each of the currencies. As mentioned, countless times before, BTC was active the longest, therefore pre-dates all the other currencies. As shown in Table 2, the starting trade dates of each currency differ and therefore would affect the predicted price of each.

Table 2

Symbol	Date
BTC	2013-04-29
ETH	2015-08-08
DOGE	2013-12-16
ADA	2017-10-02
BNB	2017-07-26
LTC	2013-04-29

Therefore, during the analysis phases of the project, I will choose to compare all currencies from 2017 to keep the prediction less biased and from a more uniform past data—which I have saved under a different file named 'cryptocurrency_markets.csv'

6. Loading Packages and Data Cleaning & Preparation

When choosing which programming language to use for this project, Python was most appropriate choice since it was most versatile –having a numerous number of packages available for basic statistical analysis and also for building more complex time series models.

Firstly, I loaded packages named Numpy, Pandas, Sklearn, Keras/Tensorflow and Mathplotlib for the purposes of cleaning, preparing, building, and plotting the dataset. The best way to make sense of the dataset is to display the columns, rows and overall structure of the data frame The initial loading and cleaning was done by using Pandas package.

{also mention how many row and colums}

After loading the dataset under the variable name "data", I applied the following:

- data.head(): resulted in the first few columns and rows but gives a good snapshot of the dataset.

	Name	Symbol	Date	High	Low	Open	Close	Volume	Marketcap
0	Bitcoin	BTC	2017-01-01	1003.08	958.70	963.66	998.33	147775008.0	1.605041e+10
1	Bitcoin	BTC	2017-01-02	1031.39	996.70	998.62	1021.75	222184992.0	1.642902e+10
2	Bitcoin	BTC	2017-01-03	1044.08	1021.60	1021.60	1043.84	185168000.0	1.678637e+10
3	Bitcoin	BTC	2017-01-04	1159.42	1044.40	1044.40	1154.73	344945984.0	1.857187e+10
4	Bitcoin	BTC	2017-01-05	1191.10	910.42	1156.73	1013.38	510199008.0	1.630025e+10

- data['Name'].value_counts(): resulted in the counts of how many counts of each cryptocurrency was provided in the entire dataset and I wanted to distinguish by name.

```
Ethereum        1648
Litecoin        1648
Dogecoin        1648
Bitcoin         1648
Binance Coin    1442
Cardano         1374
Name: Name, dtype: int64
```

Lastly, I made sure there were no missing or NA (Not applicable) values in the entire dataset—which luckily, there went any to begin with.

7. Data Exploration & Visualization

a. Market Capitalization

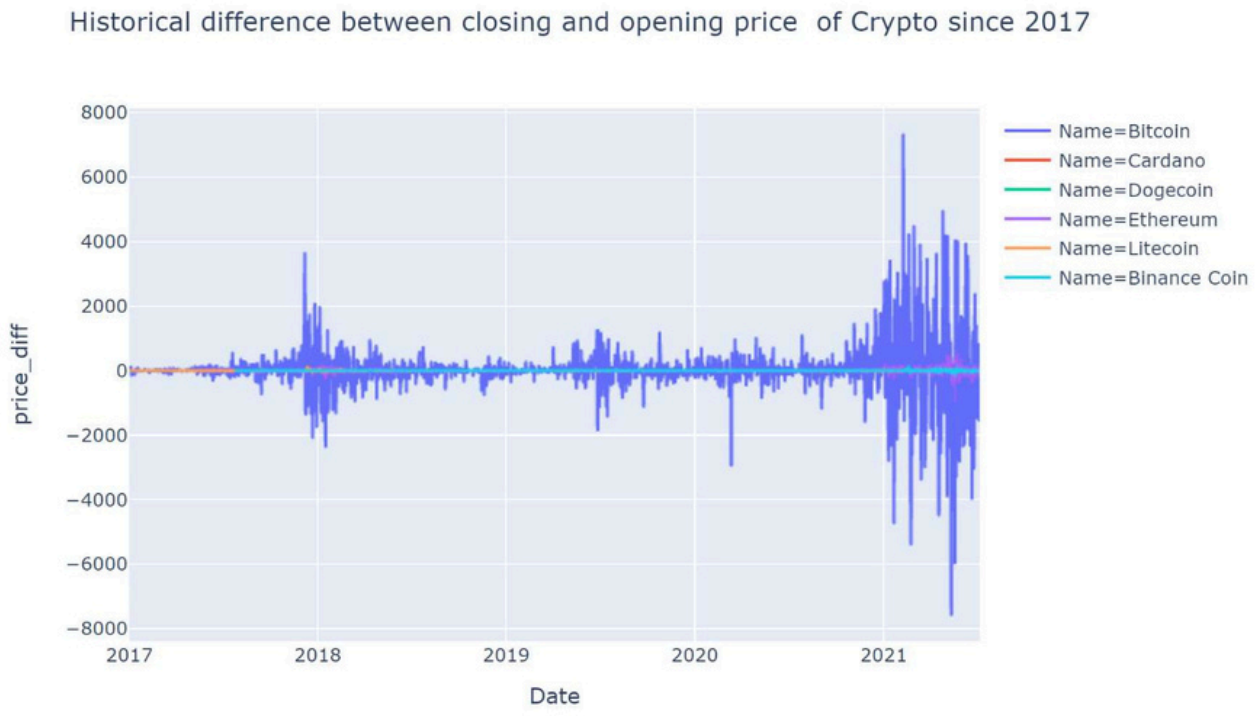
Since the entire dataset contained values for all the currencies in one, the first step was to separate the data to make it easier to visualise the comparisons.

The market capitalization of each currency is to figure out which currency has the most influence in the market. The hypothesis is that Bitcoin would have the most influence due to its longevity and due it being the most popular as the pioneer cryptocurrency project. Prior to 2017, the market capitalization is virtually non-existent. Therefore, I chose to graph the market capitalization starting from the year 2017, from which we can see that Bitcoin is the most traded. We also find that the as Bitcoin prices increased first, so did the other currencies prices increase. Most important, it is evident that Bitcoin

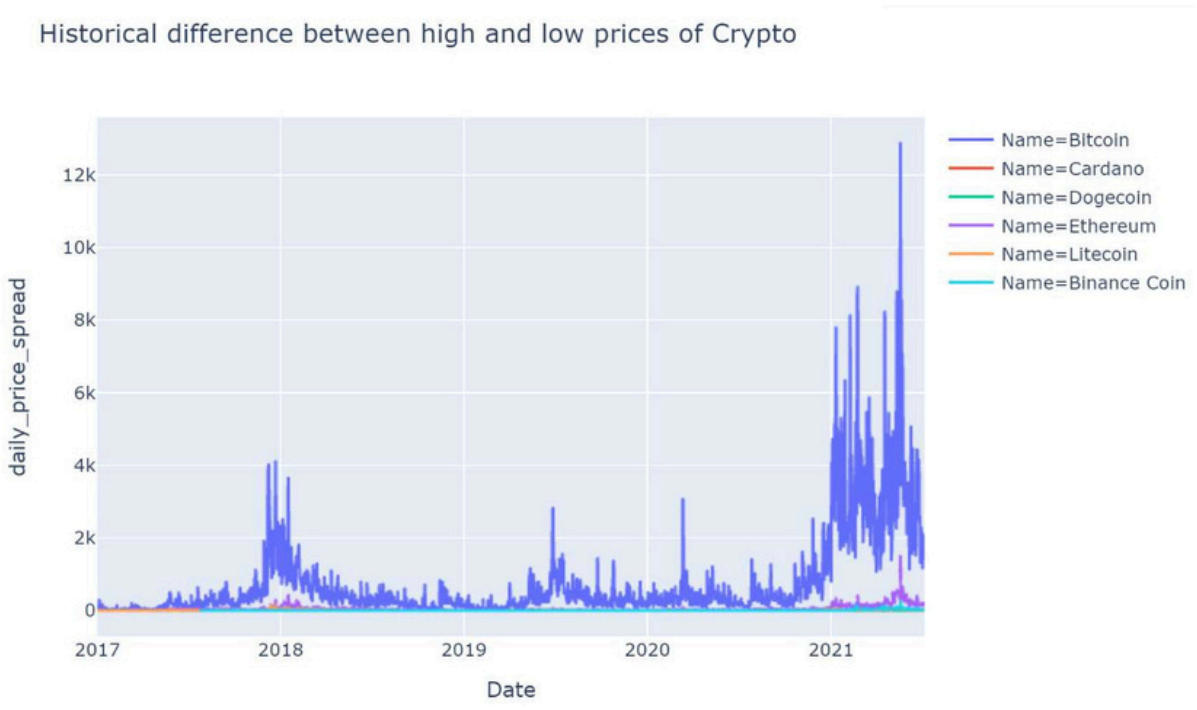
influences the prices increase of the other currencies, as hypothesized in the literature review. The

b. Difference between Closing and Opening Prices

Next, I wanted to explore the difference between the daily closing and daily opening prices for each data point for each cryptocurrency. The graph shows us the Bitcoin clearly has the highest price fluctuations from opening to closing in a day. The oscillation indicted the variation in each currency—besides Bitcoin, the next most varied price is evident in Ethereum as expected since Ethereum was introduced as the second most traded crypto project in the works since 2017. Another aspect that is evident is that since the prices increases, so do the price fluctuations.



c. Difference between High and Low Prices



To reiterate the same theme of increased prices, overall price fluctuations and comparisons between all the cryptocurrencies prices of each data point is meant to showcase daily price variations: Bitcoin and Ethereum again vary the most. Besides those, the other currencies show plateaued trend. We can also 2018 was most active for Bitcoin and post-pandemic has been the most active –the spikes are at all time high since people were probably exploring other means of earning income—due to the increase in unemployment.

8. Data Modelling

As mentioned in the literature review section, there are a different option when it comes to applying machine learning models to forecast prices when it comes to stock price prediction and now even cryptocurrency price prediction. The mathematics boils down to the statistical phenomenon known as Time Series Analysis. However, Time series is applied on stationary data whereas crypto price data is non-stationary.

a. RNN

The most popularly mentioned Machine learning model for forecast prices was the RNN-LSTM model. Recurrent Neural Network (RNN) is an algorithm used for sequential data that remembers its input through its memory. Price prediction is an example of sequential data because we are taking the historical and current information to makes sense of potential future sequences. Price can be considered sequential since price don't show change too much daily but vary over a longer period of time. RNN is known to work well for data such as financial data, speech-text, weather and so on. In layman terms, when an RNN makes a decision, it considers the current input and previously gathered inputs.

b. LSTM

The Long short-term memory network are an extension of the RNN—it just refers to an an increase of memory. As mentioned before, for price prediction, we require analysing information over a longer period of time—therefore, using the LSTM allows us to maintain memory for a longer period of time.

The goal for the project was to build the same RNN-LSTM model on each cryptocurrency type and then compare the results over the same period of time from 2017-2021.

To prepare the data, I first separated the cryptocurrency by type and created separate dataset s for each
For instance: To extract just Bitcoin data, I created a variable “bitcoin” that contained all the columns and rows that factored by desired “Name” column.

```
bitcoin=data[data['Name']=='Bitcoin'].copy()
```

Through trial and error, I realised that it was more robust to separate the training and testing in blocks.

Training and Test Split

The next step was to split the factored dataset into the training and testing dataset. Previous works show various ways of splitting data; in some works, the use of sklearn package was evident—specifically train_test_split(). Instead, I preferred to split the datasets by Date. The training set contained all the information up to January 1st, 2020, and the test set contained information after January 1st 2020. Then, I proceeded to filter out all columns besides Open, Close, Low , High and Volume from the dataset.

Normalization

Normalization is used to scale the numeric values in that data set to a common ground without disrupting ranges of the different columns.

The training data was scaled using the MinMaxScaler() which resulted in an array of scaled numeric.
Since the goal of the project was to predict 60 days into the future, I decided to predict prices based on the past 60 days. Other articles and relatable works suggested to use either a past 60-time step or past 90-time step for the best results.

Therefore, the next step was to create two empty lists: X1_train and Y1_train. X1_train would contain information for the past 60 days and Y1_train would contain price information for 61st day.

Building RNN-LSTM model for the training set and testing set

To build the model, I imported the following libraries needed to create our model:

```
from keras.models import Sequential
from keras.layers import Dense
from keras.layers import LSTM
from keras.layers import Dropout
```

To initiate the RNN model, I created “model1” assigned to Sequential(). As a novice, I followed the steps as stated in relatable works and customized according to the scope of my datasets through trial and error. More accuracy is achieved when more layers are added to the LSTM models. The depth of a neural network indicates its level of accuracy.

Dense refers to adding a densely connected layer to the network and Dropout refers to the preventing overfitting of the training and therefore preventing bias in the performance of the layers of the LSTM.

For the model I added 4 layers for the LSTM with dropout of 0.2 means that 20% of the layers will be dropped. Following the LSTM and Dropout layer, we apply the Dense layer of units= 1, to specify the output of one unit. The output refers to the predicted value.

Next step is to compile these layers that accounts for the loss, the optimizer and validation split.

In machine learning, the loss function accounts for the distance between an actual value and the predicted value. Even without basic statistical knowledge, the obvious move is to reduce that distance between the actual and predicted value. Therefore, the choice for the loss was easily “mean squared error” which is also used in linear regression. An optimizer is used to optimize the minimization of the loss function—therefore it is important to choose one that minimizes the distance the most.

The lack of personal knowledge on the subject forced me to research the reason why majority related works used the “Adam” optimizer. According to an article by Towards Data Science, the Adam optimizer is popular among adaptive optimizers and also in general because it works well with with sparse data.

Lastly the model is fit to run for 20 epochs—epochs being the number of times the model runs throughout the training set. Other works chose to set the epochs at a higher number because the higher the number, the higher number of iterations and therefore

the higher amount of accuracy. It is highly suggested in most articles and comparable works to use 100 epochs; however, I chose to save space and time on the Google Collaboratory platform since I would be building 6 different models, and each would take extremely long if 100 epochs were applied to each.

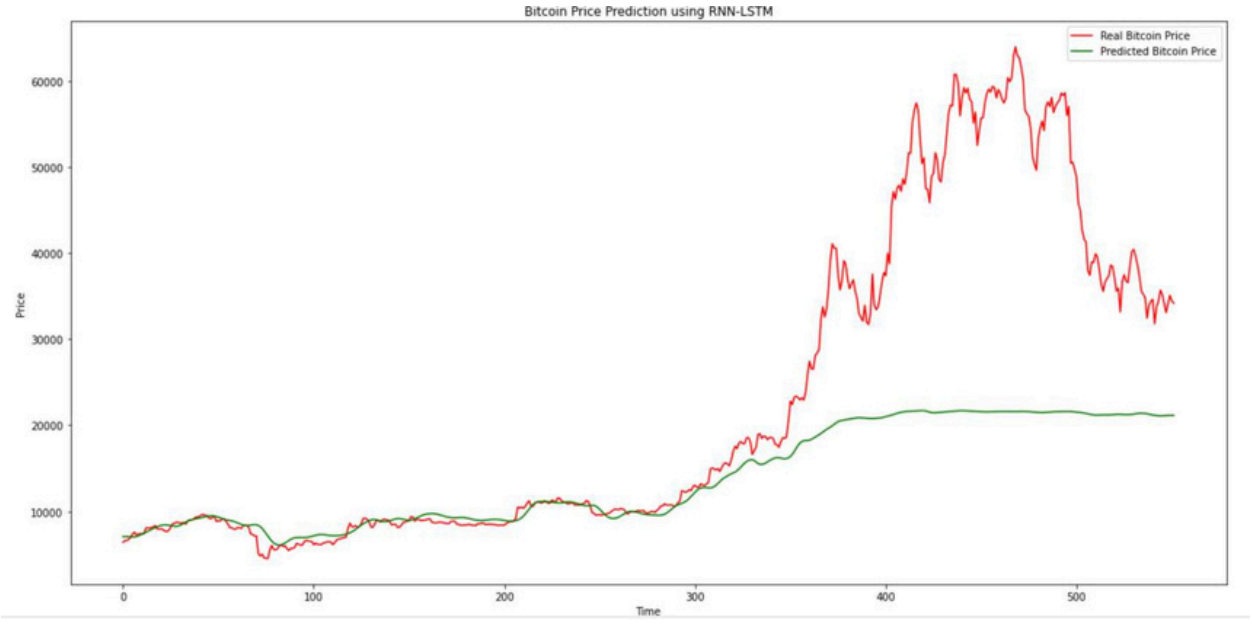
The next step to make use of the testing dataset and all steps that were required the training part are applied to the test set -- excluding the model compiler.

Plotting the Predicted and Actual Value

This stage involves making use of the matplotlib library to graph the predicted and actual pries of each cryptocurrency model. Although my original goal was to forecast prices into the future, the LSTM model used I this project allowed me t visualize and compare the overall trend of Actual Values versus the Predicted Value over the period of 2017-2021.

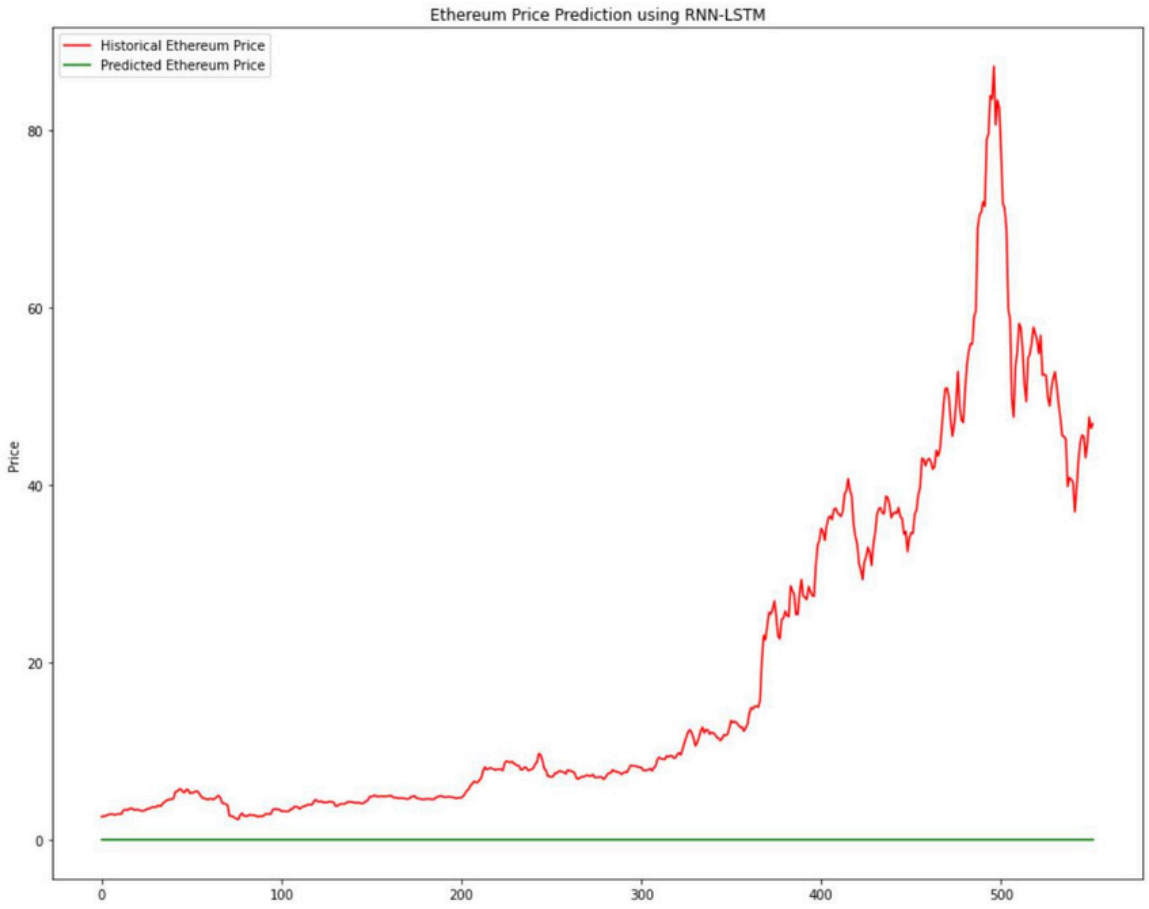
In all the following plot, the green line represents the Predicted Values and red line represents the Actual values.

a. Bitcoin (BTC)



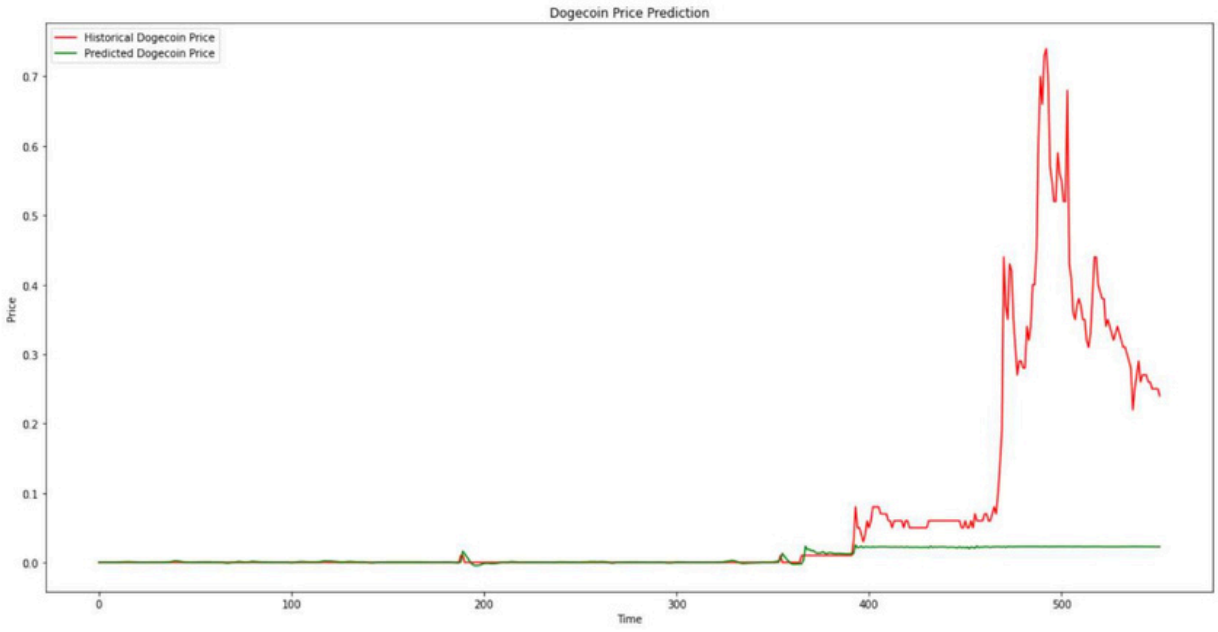
Bitcoin trends match up until the 300th observation where can see that the Actual prices spiked up whereas the predicted price plateaued and remain unchanged. We can assume that spike was a result of the popularity of the Cryptocurrency as a new commodity that people could invest in. Those who invested in Bitcoin between this period gained a lot more than anticipated. The actual prices show a downward trend, however, still having a higher value than the predicted price.

b. Ethereum (ETH)



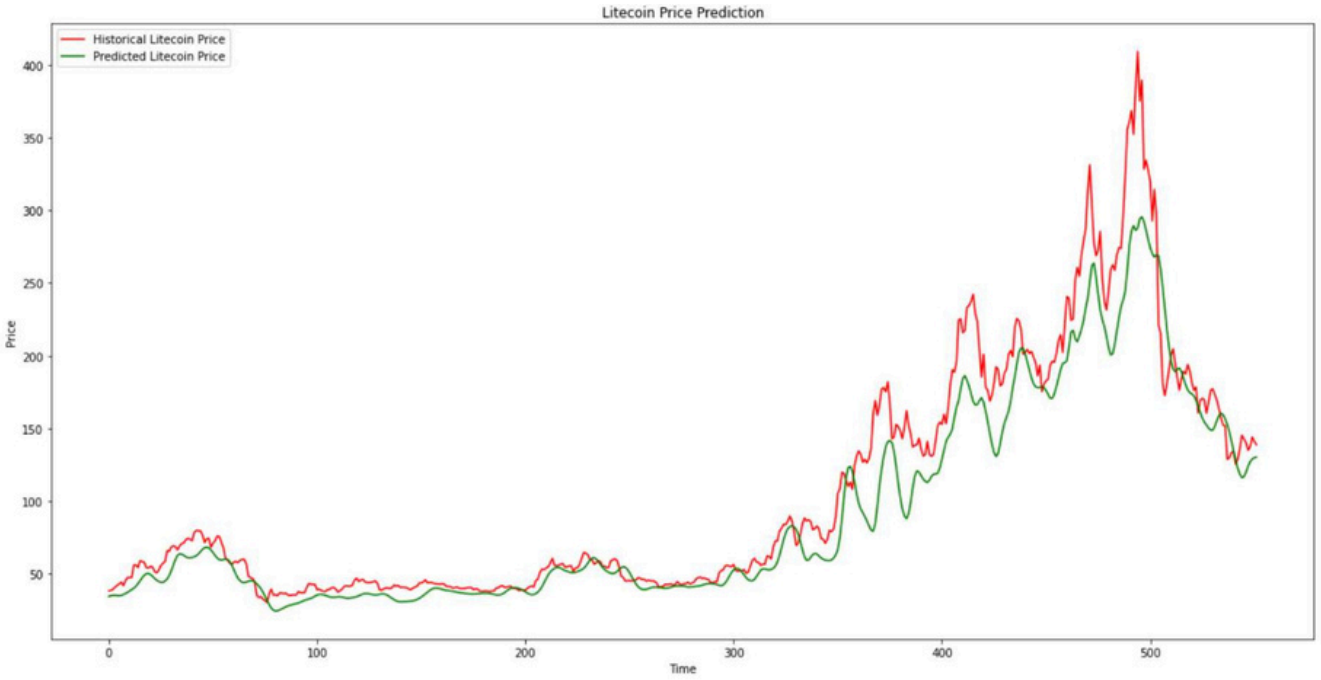
Ethereum's Actual price trend also rises around the 300th observation. We can expect this since it resembles the bullish trend of Bitcoin's Actual Price. Ethereum is the 2nd popular currency in the Crypto world, therefore it makes sense that it will follow Bitcoin's trends.

c. Dogecoin



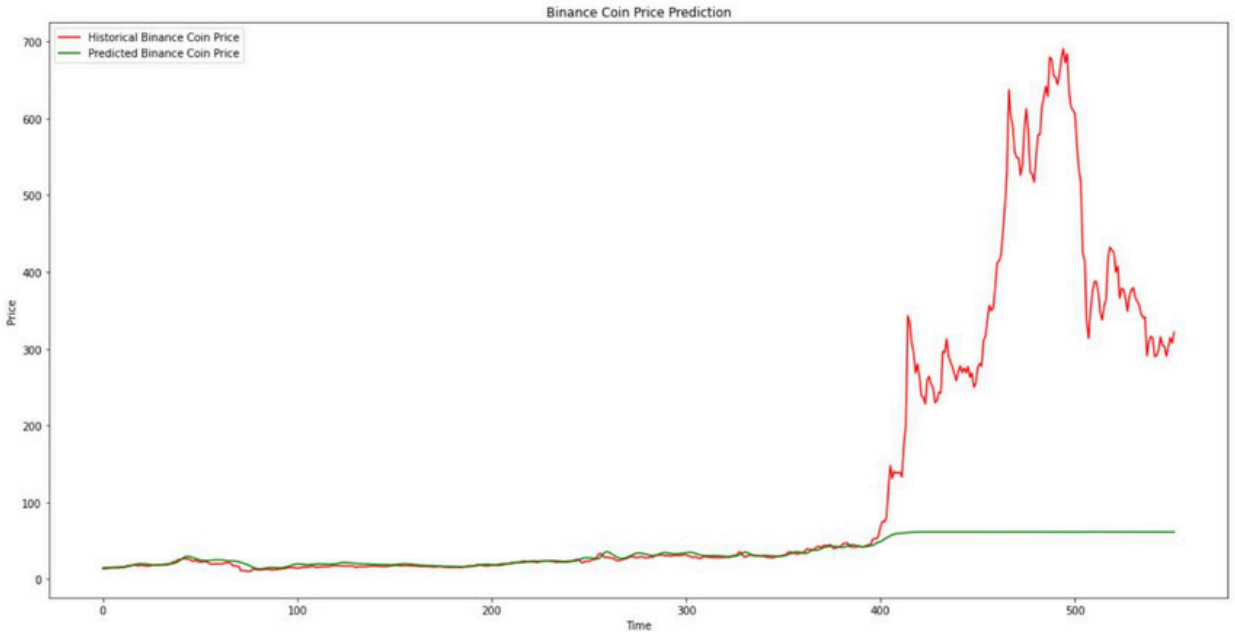
Dogecoin gains momentum later—around the 400th observation---due to it being a latecomer in the crypto markets. It spikes high around the current years –i.e., late 2020 and then shoots back down in early 2021. This sudden bearish trend is a factor of how in early 2021, Elon Musk's remarks about the Dogecoin negatively influenced the price. It is interesting to see that the predicted trend shows not much momentum at all.

d. Litecoin



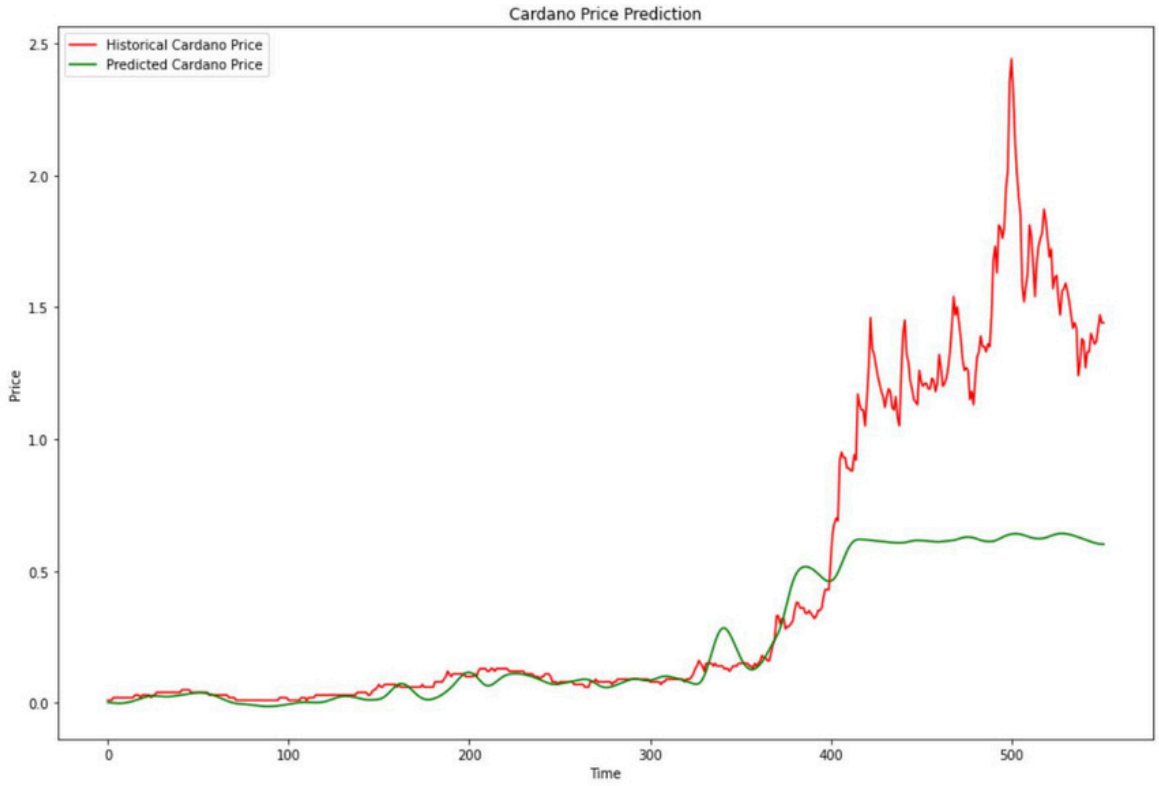
Litecoin's actual and predicted prices both show gradual increases and almost line up with each other. Which means the model accurately represented the trends at which Litecoin was actual going. Due to the accuracy, investors can rely on this model to be the most accurate at representing the future of Litecoin's profits. Additionally, during its bullish trend, fortunately investors most probably gained more than expected as the actual price spike higher than the predicted price.

e. Binance Coin



Once again, we can see that Binance Coin's actual prices show sudden increase or a peak in prices in approximately late 2020 and slight downward trend in 2021.It has a similar trend as to Dogecoin. These two currencies have similar in trends so it would wise for an investor to invest in one or the other, but not both—just so their portfolio is diversified.

f. Cardano



Lastly Cardano's plot shows that the predicted prices did not anticipate the huge increase in the actual price which meant that investors earned a huge return on their investment if they stuck with it during the pandemic. The 400th observation can be assumed as the time around COVID-19 where the return of investment soared. Personally, I invested in Cardano, and it was at an all time high mid 2021 and now it is indicating a bearish trend.

8. Comparable Works

After the literature review, I explored academically inclined websites and blogs that explored similar projects ---most of which were conducted with stock prices data and not Cryptocurrency price data. However, the structure of most project was similar which helped me understand how to apply it to my dataset.

9. Conclusions

There is a reason why investors are warned about the volatility of cryptocurrency price trends. Traditional stock and commodities also have their place in a volatile market; however, the risks are manageable due to diversification --such as products like mutual funds, hedging and government bonds can mitigate fear of intense risk.

On the other hand, cryptocurrencies are relatively new and not much research has been done on how to manage the volatility of these currencies. In comparison to regular stocks, the volatility is also much more erratic and there has been little research done on how to mitigate that.

Due to the erratic behavior of the crypto markets, we see that the predicted trends in the prices is likely to look much different from reality. This is evident in the plots for all six cryptocurrency projects ---but it is most evident in Bitcoin, Ethereum Dogecoin, Binance Coin and even Cardano.

For a stronger example, we can focus on how the predicted prices (green line) of all three, Ethereum, Dogecoin and Binance Coin were showed no growth in the future, however, the actual prices soared. Therefore, it is hard to gauge the behavior of crypto markets.

Self-Reflection

Overall, this assignment helped me explore the use of Machine learning and Deep learning models—if not in depth, then at least superficially. I am also more comfortable with Python and the specified libraries.

Future Changes

I will extend this project to creating a recommender system that would guide a user to decide which cryptocurrency to invest based on the prediction trends. A recommender system is another machine learning algorithm which I would have research separately.

Another extension would be exploring Natural Language Processing algorithm to further explore how crypto market trends are heavily influenced by sentiment analysis—i.e., social media posts, news alerts, etc.

10. References

1: Kharpal, Arjun. “Cryptocurrency Market Value Tops $2 Trillion for the First Time as Ethereum Hits Record High.” CNBC, CNBC, 6 Apr. 2021, htt ps://w ww.cnb c.com /2021/0 4/06/c ryp tocur rency-market-cap- tops -2-trillion -fo r-the-first-time.html.

2: “All Cryptocurrencies.” CoinGecko, https://www.coingecko.com/en/coins/all.

3: Mudassir, Mohammed, et al. “Time-Series Forecasting of Bitcoin Prices Using High-Dimensional Features: A Machine Learning Approach.” Neural Computing and Applications, Springer London, 4 July 2020, https://link.springer.com/article/10.1007/s00521-020-05129-6.

4: Pintelas E., Livieris I.E., Stavroyiannis S., Kotsilieris T., Pintelas P. (2020) Investigating the Problem of Cryptocurrency Price Prediction: A Deep Learning Approach. In: Maglogiannis I., Iliadis L., Pimenidis E. (eds) Artificial Intelligence Applications and Innovations. AIAI 2020. IFIP Advances in Information and Communication Technology, vol 584. Springer, Cham. https://doi.org/10.1007/978-3-030-49186-4_9

5: Srk. “Cryptocurrency Historical Prices.” Kaggle, 7 July 2021, https://www.kaggle.com/sudalairajkumar/cryptocurrencypricehistory.

6: Adcock, Robert, and Nikola Gradojevic. “Non-Fundamental, Non-Parametric Bitcoin Forecasting.” Physica A: Statistical Mechanics and Its Applications, North-Holland, 12 June 2019, https://www.sciencedirect.com/science/article/pii/S0378437119309859.

7: Yogeshwaran, S. & Kaur, Maninder & Maheshwari, Piyush. (2019). Project Based Learning: Predicting Bitcoin Prices using Deep Learning. 1449-1454. 10.1109/EDUCON.2019.8725091.

8: Simonetti, Juan. “Short Term Bitcoin Price Prediction with Deep Learning.” Medium, Geek Culture, 25 Aug. 2021, https://medium.com/geekculture/short-term-bitcoin-price-prediction-with-deep-learning-ab4386e84b5.