

**Solent University**

**Faculty of Business, Law and Digital Technologies**

**A Technical Report On The Exploratory Analysis of Machine Learning Models Towards the Development of A Cryptocurrency Trading Platform**

**FOR THE DEGREE OF MSC APPLIED AI & DATA SCIENCE**

**Author: Oriyomi Adepitan**

**Course Title: Applied AI in Business (COM724)**

**Module Leader: Femi Isaiq**

**Date: July 25, 2022**

**Acknowledgement**

I would like to express my deepest appreciation to my supervisor Femi Isaiq, Ph.D. for his guidance and feedback during this project.

A TABLE OF CONTENTS

[**1 Introduction**](#_s8zhimcevipo) **5**

[1.1 Overview](#_9b2qowpqddt7) 5

[1.1.1 Objective](#_xl3wyo99qna0) 5

[1.1.2 Value Proposition](#_k3ri1hfcblul) 5

[1.1.3 Success Metrics](#_qbjcxbaivejb) 6

[1.2 Literature Review](#_slf35blm033l) 6

[1.3 Project Layout](#_3thfcdab30u4) 7

[**2 Data**](#_z7r9rhdqvl4h) **8**

[2.1 Data description and variable selection](#_lsa9tcrl7b1) 9

[**3 Models**](#_carjiqqbmvyd) **10**

[3.1 Long short-term memory (LSTM)](#_fbe3mx22urjn) 10

[3.1.1 Brief overview of the LSTM structure](#_n8xyvfqmhi1z) 11

[3.2 XGBOOST](#_4taql9p3tsst) 11

[**4 Methodology**](#_pir3a6ddgl7e) **12**

[4.1 Data preparation](#_ue0pwxlfzqv7) 12

[4.2 Model Implementation](#_qwktdso7fce3) 13

[4.2.1 LSTM](#_qpztefd2ssqj) 13

[4.2.2 XGBOOST](#_hzhm2p18p76) 14

[4.2.3 Performance evaluation](#_3oxaow1mkxr7) 16

[**5 Results**](#_rrybdzihagi7) **16**

[5.1 LSTM model’s result](#_i84rucbtf0qk) 16

[5.2 XGBoost model’s result](#_6bx4d4600cfq) 17

[5.3 Assessment metrics and Discussion](#_ykey5864bec9) 17

[**6 Conclusion**](#_v5vam8ifrxts) **18**

B References

|  |
| --- |
| List Of Tables   1. **Table 1**: A descriptive summary of the variable in the dataset |

|  |
| --- |
| List Of Images   1. **Figure 1** : Mean Absolute Percentage error 2. **Figure 2**: Project Architectural Layout 3. **Figure 3**: Fetching and cleaning data from source 4. **Figure 4**: A time series plot of closing price from 2018-01-01 to 2022-07-28 5. **Figure 5**: A time series plot of closing price from 2022-02-01 to 2022-08-03 6. **Figure 6**: Sample data showing the variables in the dataset 7. **Figure 7**: LSTM network with the different layers (Mittal, 2019) 8. **Figure 8**:Scaling dataset and splitting into training and testing data. 9. **Figure 9**: Util function to create dependent (y) and Independent (X) datasets 10. **Figure 10**: Training LSTM model , with hyper parameters. 11. **Figure 11**: Initializing space of a required range of values 12. **Figure 12**: Hyperopt Function and Objective Function . 13. **Figure 13**: Hyperparameter Tuning best output. 14. **Figure 14**: LSTM MSE Score. 15. **Figure 15**: XGBoost MSE Score. 16. **Figure 16:** LSTM ‘s predictions against actual prices 17. **Figure 17:**XGBoostModel predictions against actual prices |

# 

# 

# 

# 

**ABSTRACT**

Following the boom of cryptocurrencies prices in recent years, It has increasingly been regarded as an investment asset. Because of its highly volatile nature, there is a need for good predictions on which to base investment decisions. Although existing studies have leveraged machine learning for more reliable and accurate price prediction, few have focused on the feasibility of applying different modeling techniques to samples with different data structures and dimensional features. To predict crypto currency price at different frequencies using machine learning techniques, we first classify each coin price by daily price and high-frequency price. In this paper, we attempt to predict crypto coins price accurately taking into consideration various parameters that affect the value. Using the available data we will predict the sign of the daily price change with highest possible accuracy. We have used LSTM and XGBOOST, also compared with benchmark results for daily price prediction, we achieve a better performance, with the highest accuracies of the statistical methods and machine learning algorithms of 98%. Our investigation of crypto currency price prediction can be considered a pilot study of the importance of the sample dimension in machine learning techniques.

# 1 Introduction

## 1.1 Overview

Crypto Coins and their derivatives life tokens (NFTs) are digital currency, unlike the physical currency which is monitored by a centralized body, usually the Central Bank of the Nation.

These digital currencies ride on a decentralized system, making them highly secure and untraceable. The coins are mined and halved at interval, and their value system varying the prices becomes volatile.

There exists many factors that affect the volatility of the coin price e.g political factors, social influence, value proposition, trust in coins, capital over turn, traders standpoint, and acceptability.

With all these it is important to predict the behavior i.e price / value of these digital currencies to help investors, companies and end users make the best of their investment.

This is why SoliGence has been created , a trading platform that helps those who want to trade and invest forecast the possible behavior of various coins that is likely to yield favorable outcomes / profit.

By predicting closing prices using LSTM and XGBOOST models and thereafter relating the accuracy of the predictions for desired purposes.

### 1.1.1 Objective

Forecasting price of crypto coins that can help users get desired profit for their capital in a set time interval usually daily, weekly, monthly or quarterly.

### 1.1.2 Value Proposition

● Profit Maximization

● Risk Mitigation

● Real Time Coin Prediction

### 1.1.3 Success Metrics

The model success is measured with root mean square error (RMSE) of 15% deviation from the actual value compared to the predicted value is acceptable as a successful forecast. Also Mean Absolute Percentage error can be used.

Success Metric Using Mean Absolute Percentage Error

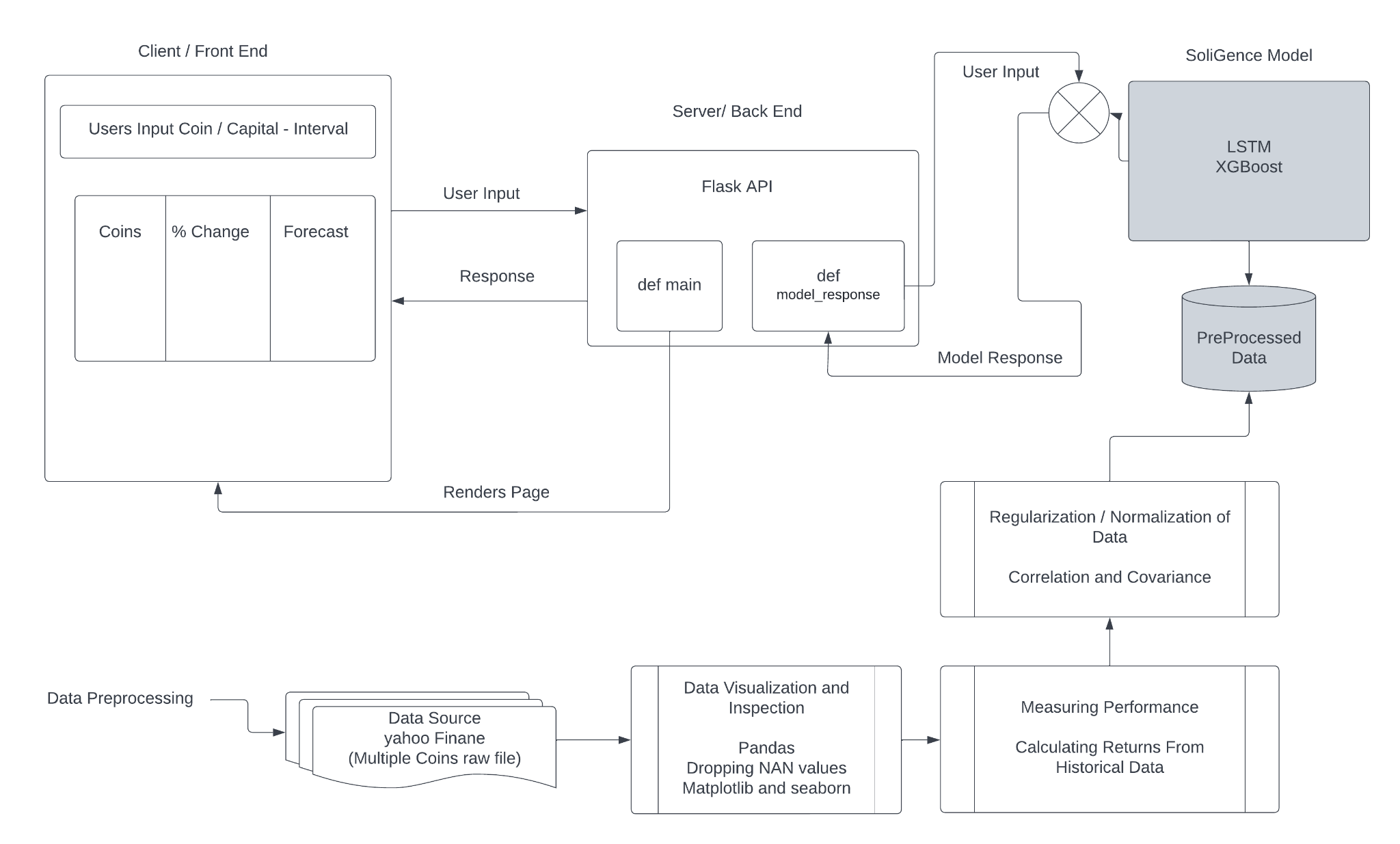
**Figure 1** : Mean Absolute Percentage error

## 1.2 Literature Review

Crypto Currency high volatility along with its popularity has allured researchers and financial analysts to study the behavior and mechanisms driving its price fluctuations. As a result, different models have been experimented with and examined for the purpose of choosing the most suitable one for the task. So far, deep learning models such as LSTM have exhibited better predictive ability compared to the traditional time series models like ARIMA. With an accuracy rate of 52,78% and an RMSE of 6.87%, LSTM empirically proved to yield better predictions regarding the trend of Bitcoin prices, outperforming the ARIMA model whose RMSE was 53.74% (McNally, Roche and Caton, 2018). Similarly, a comparative study on the predictive power of machine learning algorithms further demonstrated LSTM’s ability, scoring 0.02 and 0.99 on MAPE and R 2 , respectively Ashwini, 2020. In contrast, few studies have been conducted on Bitcoin using the SVR model but in those few instances it has been used it featured well recording an R 2 of 78% Alahari, 2020, and a Diebold-Mariano test score of 0.992777 Jana, Ghosh, and Das, 2021.

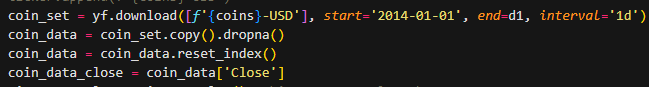
Also, following W. Long, Z. Lu, and L. Cui, deep learning-based engineering for stock price movement prediction, due to noisy and non-stationary characteristics of samples. Feature learning can be performed more efficiently by purposely designed networks. Our model proposes using the multi-filter neural network (MFNN) specially for feature extraction on time series samples and price behavior prediction objective.

## 1.3 Project Layout

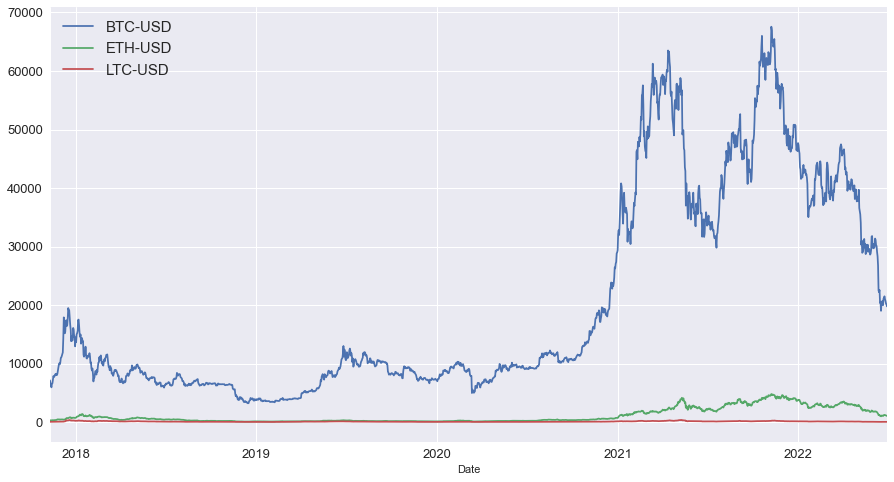
The project is structured in 2 major sections, analysis and visualization of data represented using plots and tables. Also an application GUI using ( Flask and Streamlit) for prediction and forecast, using LSTM and XGBOOST model. Chapter 4 will provide descriptions of the methodology employed during implementation of the models as well as performance evaluation metrics. Finally, the project aims to give conclusions that can be drawn from the results as well as limitations.  
  


**Figure 2**: Project Architectural Layout

# 2 Data

The primary source of data is from yahoo finance and comprises daily crypto currency volume and price data ranging from 2014-01-01 till present day. The curve of the price for various coins seems flat at early stages when there was significantly lower activities in the coins market, as opposed to the times from 2022-03-28 with more dominating market activities.  
  


**Figure 3**: Fetching and cleaning data from source



**Figure 4**: A time series plot of closing price from 2018-01-01 to 2022-07-28



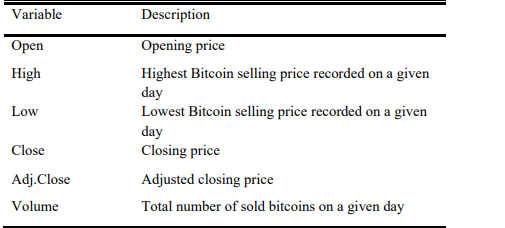
**Figure 5**: A time series plot of closing price from 2022-02-01 to 2022-08-03

## 

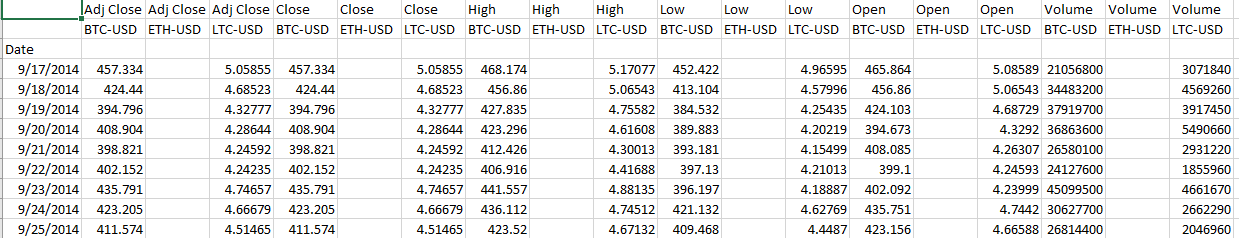
## 2.1 Data description and variable selection

As shown in Table 1 and Figure 5 below, the data in the series have a daily frequency with information about adjusted close, close, high, low, open, with which the coins are traded and also the total number of coins transaction recorded at end of day referred to as the volume.

Relative to stocks having it’s adjusted close and close prices differ due to corporate review like dividends, stocks splits among other variables. The crypto currency market isn’t affected by such amendments leading to same close and adjusted close prices as evident in Figure 5.



**Table 1**: A descriptive summary of the variable in the dataset



**Figure 6**: Sample data showing the variables in the dataset

Going forward the variable ‘Close’ will be our primary data point towards which all analyses are based. It symbolizes the price at which cryptocurrencies coins were selling for a given day.

# 3 Models

## 3.1 Long short-term memory (LSTM)

Developed by Sepp Hochreiter and Jurgen Schmidhuber in the year 1997, LSTM is a remodeled kind of the Recurrent Neural Network (RNN) primarily modeled to mitigate the shortcomings of the basic RNN this includes the long-term dependency and vanishing gradient (this is achieved by employing the tanh activation function). Olah, 2015 RNN struggles at remembering/memorizing information after a long period has elapsed which decreases its predictive power. The ability of LSTM models in memorizing data for longer periods and the existence of the different gates makes the model suitable for volatile time series prediction.

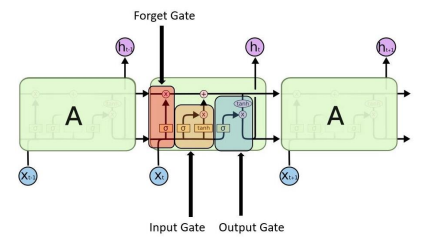
### 3.1.1 Brief overview of the LSTM structure

As shown in figure 3, An LSTM network's structure is made up of the before Input 𝑥𝑡−1 It is sent into the LSTM unit A where the input is subjected to mathematical operations like the sigmoid and tanh to produce an output ℎ𝑡−1.The output is then sent to the following unit via the Cell gate (𝐶𝑡−1) , which is the horizontal line located above the gates and serves as a conduit for the pre-processed data.

The forget gate acts as an evaluation hedge toward the next cell, ensuring that only relevant information is retained. As a result, this serves as the LSTM's memory, storing useful information for use in the prediction process while disregarding/forgetting information deemed irrelevant for making predictions. The input gate functions as a filter, determining the extent to which the information can be rendered useful enough to be stored in the network's long-term memory and thus allowing access to the current unit's cell gate(𝐶𝑡).

Finally, the output gate is responsible for deciding which output ℎ𝑡

is to be given out.



**Figure 7**: LSTM network with the different layers (Mittal, 2019)

## 3.2 XGBOOST

XGBoost is a machine learning boosting algorithm recognized for its high performance based on supervised learning. This algorithm is popular for classification and regression problems. Its main merit is related to its high speed in core computation. The XGBoost working process is based on the following to predict the output. Equation (1) presents the prediction process. E is the dataset, d is the number of features, and i is the number of examples

(1)

To predict the , the process generated from Equation (2). C represents the total records of trees in the model, shows the last tree for this model.

(2)

Finding the best functions requires minimizing the loss and objective regularization by following Equation (3):

(3)

Y is showing the loss function based on the differences of prediction between the output value ˆdx and actual value dx. Ω presents the model complexity and avoids overfitting. This process evaluates by following the Equation (4):

(4)

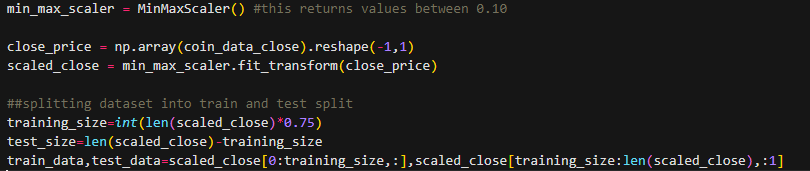
L presents the total records of leaves in the tree, and V shows the weight of every leaf.

# 4 Methodology

## 4.1 Data preparation

After getting source data from yahoo finance, examination of dataset for possible missing values was carried out, Kang, 2013 as their presence could lead to misleading results or even biased predictions. All rows with missing columns were dropped , because the variation in crypto currency values using interpolation or mean value might lead to skewed results. The data was split into training and test data with the ratio of 75/25. 75% of the close price data was used to train both the LSTM and XGBoost models and the 25% was used to test the models accuracy and behavior.

According to Stottner, 2019 scaling our data is said to speed up machine learning models learning process, the scikit-learn package MinMaxScaler() was used to scale both the train and test data from the actual value (i.e $23,000) to a [0,1] range format shown in quate 5



**Figure 8**:Scaling dataset and splitting into training and testing data.

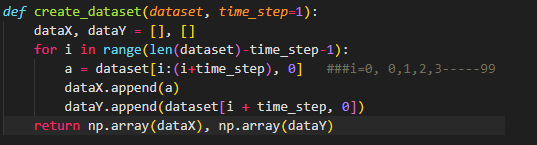
(5)

All scaled values are converted back to their initial form using the scaler.inverse\_transform() after model prediction was completed. This is to aid visual representation of predicted output against actual close price, and to match dataset initial format.

## 4.2 Model Implementation

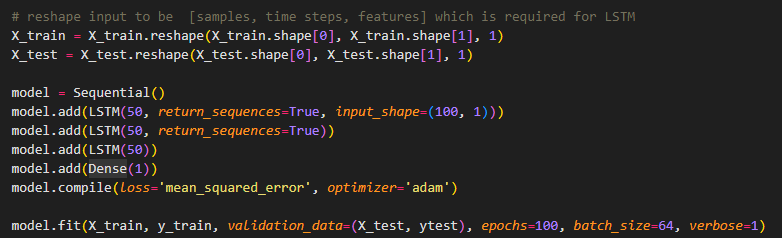
### 4.2.1 LSTM

The LSTM model dataset uses a time stem size for building the X\_train and y\_train from the train data. I used a time\_step of 100, with the formula X = t, t+1, t+2, t+3 and Y= t+4. A util function was written to help with this activity as seen in the figure below.



**Figure 9**: Util function to create dependent (y) and Independent (X) datasets

Next using tensorflow keras model (Sequential), with three hidden layers (Number of neurons), also the ‘adam’ optimizer along with the mean\_squared\_error was used to compile the model. The model was trained using an epoch of 100 and batch\_size of 64.



**Figure 10**: Training LSTM model , with hyper parameters.

This operation is carried out for each crypto coin and saved with the model's path.

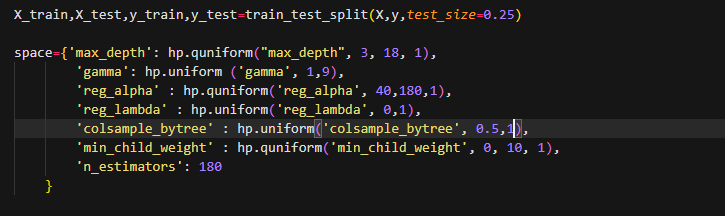
### 4.2.2 XGBOOST

Following a similar approach with the LSTM approach, we use our data source from yahoo finance, but unlike the previous approach where we used just the close prices for model prediction and training, with XGBoost we use the open, high, low, close and volume columns to train and predict behavior of the coins.

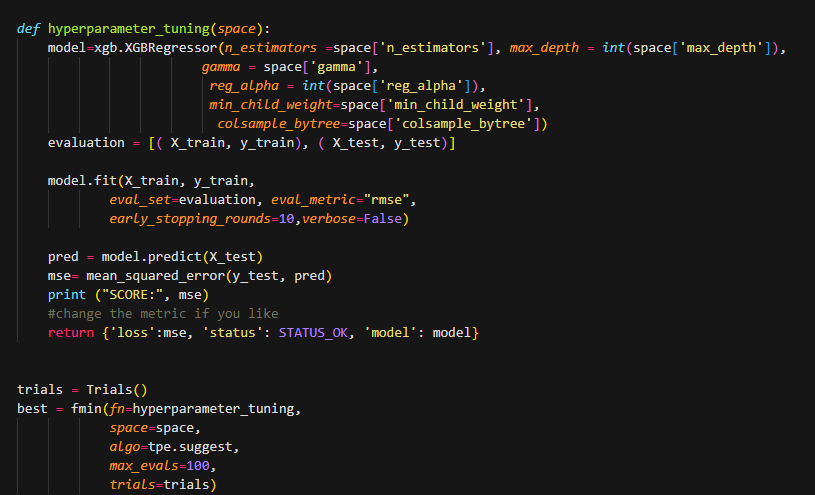
The scikitLearn train\_test\_split method is used to split our dataset into test and training data with a test\_size of 25%. Before training and fitting the model, hyperparameter tuning was carried out using the python hyperopt library which uses the “Hyperopt Bayesian Optimization” for Xgboost.  
The major parameters for the machine learning model were the learning-rate, max-depth, col-samples, weights and gamma.

After completing parameter tuning the best score based on the MSE(mean squared error) was used for our final model fitting and prediction.

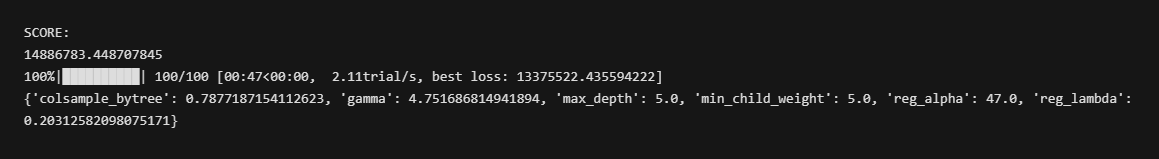
This process was carried out for all columns of our dataset and as such we are able to predict not just the future close prices, but also the possible, high, low , open , close and volume of a particular crypto currency.



**Figure 11**: Initializing space of a required range of values



**Figure 12**: Hyperopt Function and Objective Function .



**Figure 13**: Hyperparameter Tuning best output.

### 4.2.3 Performance evaluation

The LSTM model had a MSE (mean squared error ) of 0.00495 and the XGBoost had a score of 0.0003.



**Figure 14**: LSTM MSE Score.

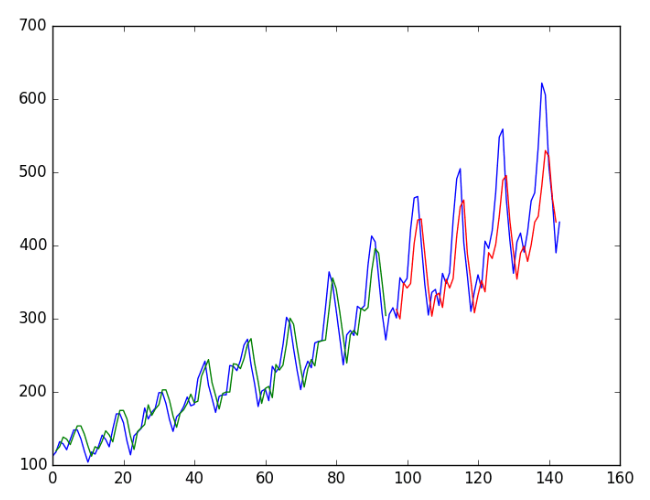


**Figure 15**: XGBoost MSE Score.

# 5 Results

## 5.1 LSTM model’s result

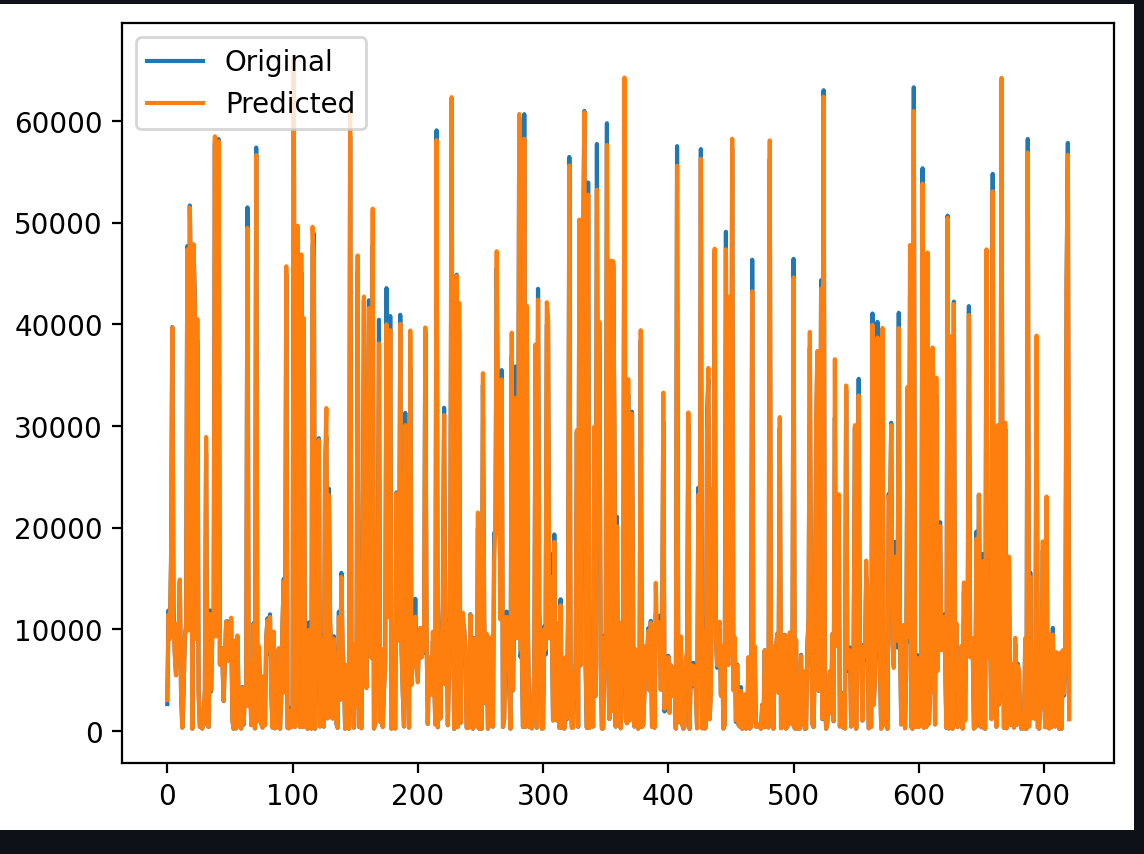
From the LSTM test predict and train predict graph , the model shows a good behavior and not underfit or overfit with. And follows a highly correlated trend line with the actual crypto currency behavior.



**Figure 16:** LSTM ‘s predictions against actual prices

## 5.2 XGBoost model’s result

The XGBoost from the MSE score and its trend line appears to be more accurate with its predictions, and for a shorter period, would be the best option, but with increase in future predicted days the values slowly deviates from the actual value by a higher MSE error.



**Figure 17:**XGBoostModel predictions against actual prices

## 5.3 Assessment metrics and Discussion

The Soligence system is powered by two comparing models LSTM and XGBoost, we predict the future values of user selected coin for LSTM and a more robust dashboard for XGBoost where users can see possible high , low , close, volume pries for selected coin for a selected number of future dates , also univariate analysis and multivariate analysis are provided to further understand trends and possibilities within the system. The XGboost Model is powered by **streamlit.**

# 6 Conclusion

The focus of the project was to build a Crypto currency predictive system, and I selected two outstanding models LSTM and XGBoost, the results have shown that both managed to identify trends and patterns in crypto currency prices. And making it easy for users to make decisions and get live updates based on their selections and inputs.

One major downside with LSTM was the speed to train and predict values and identify trends, and this was improved by pretraining the models and saving for future use, this drastically saved training time , but introduces some staleness in the system , and it would be recommended to use a cron job or build a pipeline to periodically update trained model saved.

Finally the system also provided a daily top feed on crypto currency that helps to identify outliers, politics , crises, lawsuits which could affect the crypto currency trends and overall our model predictions.

# B References

Ashwini, C. (2020). Forecasting cryptocurrency prices using machine learning. Retrieved from:

<https://www.semanticscholar.org/paper/Forecasting-Cryptocurrency-PricesusingMachine-Chaudhari/d7609fb9436d3012b83ddffe0da0480b43224725.>

Alahmari, Saad. (2020). Predicting the Price of Cryptocurrency using Support Vector Regression Methods. Journal of mechanics of continua and mathematical sciences. 15.

10.26782/jmcms.2020.04.00023.

Awad, M and Khanna, R. (2015). Efficient Learning Machines. Theories, Concepts, and

Applications for Engineers and System Designers. pp.68-71. Available at:

<https://link.springer.com/content/pdf/10.1007/978-1-4302-5990-9.pdf>

Bora, N. (2021). Understanding ARIMA models for Machine Learning. Retrieved from:

h[ttps://www.capitalone.com/tech/machine-learning/understanding-arima-models/](http://www.capitalone.com/tech/machine-learning/understanding-arima-models/)

Brownlee, J. (2019). How to use learning curves to diagnose Machine Learning Model

Performance. <https://machinelearningmastery.com/learning-curves-for-diagnosing-machinelearning-model-performance/>

C, Olah (2015). Understanding LSTM networks. Retrieved from:

<https://colah.github.io/posts/2015-08-Understanding-LSTMs/>

Duchi, J., Hazan, E., & Singer, Y. (2011). Adaptive subgradient methods for online learning and

stochastic optimization. Journal of machine learning research, 12, 2145.

<https://www.jmlr.org/papers/volume12/duchi11a/duchi11a.pdf>

Fletcher, T (2008). Support Vector Machines explained. Retrieved from:

<https://www.csd.uwo.ca/~xling/cs860/papers/SVM_Explained.pdf>

Jana, R.K, Ghosh, I, and Das, D. (2021). A differential evolution-based regression framework for

forecasting Bitcoin price. Annals of Operations Research. Retrieved from:

<https://repository.iimb.ac.in/bitstream/2074/17750/1/DAS_AOR_2021.pdf>

JavaTpoin. (n.d). Support Vector Machine Algorithm. Retrieved from:

<https://aus.libguides.com/apa/apa-no-author-date>

Ji S, Kim J, Im H (2019). A Comparative Study of Bitcoin Price Prediction Using Deep

Learning. Mathematics. Retrieved from:<https://doi.org/10.3390/math7100898>.

Kang H. (2013). The prevention and handling of the missing data. Korean journal of

anaesthesiology, 64(5), 402.<https://doi.org/10.4097/kjae.2013.64.5.402>

Kollewe, J. (2021). Bitcoin price surges to record high of more than $68,000. Retrieved from:

<https://www.theguardian.com/technology/2021/nov/09/bitcoin-price-recordhighcryptocurrencies-ethereum>.

McNally, S., Roche, J. and Caton, S. (2018). Predicting the price of bitcoin using machine

Learning. 26th Euromicro International Conference on Parallel, Distributed and

Network-based Processing, pp. 339–343. doi: 10.1109/PDP2018.2018.00060.

Mittal, A. (2019). Understanding RNN and LSTM. Retrieved from: <https://aditimittal.medium.com/understanding-rnn-and-lstm-f7cdf6dfc14e> (Accessed: 30 July 2022)

Nakamoto, S. (2009). Bitcoin price surges to record high of more than $68,000. Retrieved

from: <http://www.bitcoin.org/bitcoin.pdf> (Accessed: 1 july 2022).

Onwutalobi, A. (2011). Overview of cryptography. SSRN Electronic Journal, p. 1. doi:

10.2139/ssrn.2741776.

Tretina, K. and Schmidt, J. (2018). Top 10 cryptocurrencies in March 2022. Retrieved from: <https://www.forbes.com/advisor/investing/top-10-cryptocurrencies/> (Accessed: 7 August2022).

Varsamopoulos, Savvas & Bertels, Koen & Almudever, Carmen. (2018). Designing neural network-based decoders for surface codes. p4. Retrieved from:

<https://www.researchgate.net/publication/329362532> (Accessed: 30 July 2022).