Bitcoin Price Prediction and Time Series Data Mining

Farazuddin Mohammad(016176836)*, Pranav Chandra Kallepali (016176836)*, and Prerna Garsole (016176836)*

*San Jose State University

***Abstract*— The aim of this project is to build, test and analyze the statistical and machine learning models to predict the closing price of bitcoin price.We can gain the necessary knowledge we need to understand the future of cryptocurrency using machine learning models.The dataset is downloaded from the Yahoo Finance website. We have used two strategies to build the model for bitcoin price prediction i.e first strategy is using the previous data of only bitcoin and the second strategy is to use the previous data to predict the price using the data from other existing cryptocurrencies like Tether, Binance, Ethereum.**

- **Github Link for All Source Codes** `https://github.com/MDfarazuddin99/BTC_Price_Prediction_and_Data_Mining`
- **Data Set Source:** `https://finance.yahoo.com/quote/BTC-USD/`
- **Google Colab for Farazuddin's Contribution** `https://colab.research.google.com/drive/1qcpG0PVVpt3SAjnBwyXy7pNhn4MO52ez`
- **Google Colab for Prerna's Contribution** Prerna Colab Notebook
- **Google Colab for Pranav's Contribution** Pranav's Colab Notebook

I. Introduction

The first decentralized digital money in the world, Bitcoin is the most secure since it uses a block chain instead of a middleman like a bank. As a type of digital gold, the price of a single bitcoin has been sharply rising since 2010. As a result, bitcoin is extremely risky for investors as its price varies often. The peer-to-peer electronic payment system known as Bitcoin was created in 2008 to address the inherent flaw in the trust-based model of transactions. Since then, it has evolved into an asset or commodity-like product exchanged in more than 16,000 exchanges worldwide. 1 Although supporters claim that one of Bitcoin's key uses is to replace fiat currency, the mystery surrounding its exact nature persists. Investors view Bitcoin as a speculative investment comparable to the Internet stocks of the previous century rather than a currency according to the standards employed by economists.

Prior to disrupting established payment and monetary systems, policymakers as well as the general public were interested in Bitcoin's trading and rising popularity. At its height in 2017, Bitcoin's market value was 300 billion US dollars, almost as much as Amazon in 2016.

After the recent ups and downs in cryptocurrency values, Bitcoin is now more frequently seen as a valuable investment. Due to its extreme volatility, accurate forecasts are necessary to guide investment choices. Few studies have focused on the viability of employing various modeling techniques to samples with various data structures and dimensional properties, even if current studies have utilized machine learning for more precise Bitcoin price prediction. For daily price prediction of Bitcoin, a collection of high-dimension features, including property and network, trading and market, attention, and gold spot price, are employed, but for 5-minute interval price prediction, basic trading features obtained from a cryptocurrency exchange are used. In comparison to more complex machine learning algorithms, statistical techniques like Logistic Regression and Linear Discriminant Analysis for daily Bitcoin price prediction using high-dimensional features reach an accuracy of 66%. more straightforward machine learning methods, outperforming them. We outperform benchmark results for daily price prediction, with the statistical methods' and machine learning algorithms' maximum accuracy values of 66% and 65.3%, respectively. For the 5-minute interval price forecast of Bitcoin, machine learning models with an accuracy of 67.2%, such as Random Forest, XGBoost, Quadratic Discriminant Analysis, Support Vector Machine, and Long Short-term Memory, outperform statistical approaches. Considered a pilot study of the significance of the sample dimension in machine learning methods, our examination into Bitcoin price prediction.

A. Work Contribution

The project's activities have been distributed equally among the team members, and a tabular listing of each team member's specific activities has been created.

Team Member	Contribution
Prerna Garsole	Dataset Preperation, Data Cleaning and Preprocessing and Exploratory Data Analysis
Farazuddin Mohammad	Statistical Models for Time Series Prediction, KNN, Random Forest Model and XGBoost Model
Pranav Chandra Kallepali	LSTM Model Implementation and Model Explainability using SHAP

II. Data

We have used the Yahoo Finance's yfinance python API to downlaod the past data of the BTC prices and it's corresponding feature variables.

A. Dataset Explanation

We have used two approaches to build models to predict the price of Bitcoin.

- Using the Bitcoin related Price Variables
- Using Bitcoin and Other Crypto-currency Coins Price Variables.

	low	high	open	close	volume
count	3.113276e+06	3.113276e+06	3.113276e+06	3.113276e+06	3.113276e+06
mean	1.152825e+04	1.154230e+04	1.153537e+04	1.153541e+04	9.689568e+00
std	1.490292e+04	1.492446e+04	1.491374e+04	1.491379e+04	2.239875e+01
min	6.000000e-02	6.000000e-02	6.000000e-02	6.000000e-02	1.000000e-08
25%	1.849557e+03	1.850000e+03	1.849925e+03	1.849980e+03	1.331197e+00
50%	7.099990e+03	7.105000e+03	7.102045e+03	7.102060e+03	3.707298e+00
75%	1.067558e+04	1.068550e+04	1.068027e+04	1.068040e+04	9.761575e+00
max	6.690000e+04	6.699900e+04	6.694122e+04	6.694122e+04	1.549613e+03

Fig. 1. Data Description

1) Dataset - 1: Write About data set 1 insert the same shit in table format

2) Dataset - 2: We noticed high correlation between the closing prices of BTC and that of other crytpocurrencies such as :

- **Ethereum (ETH) :** Blockchain technology underpins the decentralized, international software platform known as Ethereum. Most people are familiar with it because of its native cryptocurrency, ether (ETH).
  Anyone can use Ethereum to develop any secure digital technology. It has a token created to compensate users for work done in favor of the blockchain, but if accepted, users may also use it to pay for material products and services.
  Scalable, programmable, secure, and decentralized are all features of Ethereum. It is the blockchain of choice for programmers and businesses building technology atop it to transform numerous sectors and how we go about our daily lives
- **Binance (BNB) :** The cryptocurrency known as Binance Coin, which trades under the sign BNB, was created by the Binance exchange. Although BNB is currently the native currency of Binance's own blockchain, the Binance chain, it was originally built on the Ethereum network. In order to permanently destroy the Binance coins it has in its treasury, or "burn," one-fifth of quarterly revenues are used by Binance to repurchase them. In 2017, Binance was developed as a utility token for reduced trading fees, but since then, its usage have grown to include paying transaction fees (on the Binance Chain), booking trips, enjoying entertainment, using internet services, and accessing financial services. Only Bitcoin, Ethereum, and USD Tether have larger market caps than Binance, which had a market worth of more than $56 billion at the time of writing.
- **Tether (USDT) :** According to its website, Tether (USDT) is a stablecoin that is backed "100% by Tether's reserves" and is pegged to the U.S. dollar. The Hong Kong-registered corporation iFinex, which also operates the cryptocurrency exchange BitFinex, is the owner of Tether. When Tether first debuted in July 2014 as RealCoin, it underwent a rebranding in November of the same year. It first went public in February 2015. Tether, which was first built on the Bitcoin blockchain, currently supports the Omni and Liquid protocols for Bitcoin as well as the blockchains of Ethereum, TRON, EOS, Algorand, Solana, OMG Network, and Bitcoin Cash (SLP).

Date	Close (BTC)	Volume (BTC)	Close (ETH)	Volume (ETH)	Close (USDT)	Volume (USDT)	Close (BNB)	Volume (BNB)
2022-12-11 00:00:00+00:00	17104.193359	14122486832	1263.868530	3362005848	1.000078	17036610394	284.390411	496833486
2022-12-12 00:00:00+00:00	17206.437500	19617581341	1274.619019	5151109364	1.000107	25872687355	276.278198	974762447
2022-12-13 00:00:00+00:00	17781.318359	26634741631	1320.549194	8812883119	1.000017	37452330494	271.951691	1925504273
2022-12-14 00:00:00+00:00	17815.650391	25534481470	1309.328735	7830915428	1.000105	31427740393	268.231537	1026628081
2022-12-15 00:00:00+00:00	17718.691406	26388742144	1291.887329	8318310912	1.000113	32406151168	265.234283	1053307328

Fig. 2. Other Coin Dataset Plot

III. Exploratory Data Analysis

Image result for exploratory data analysis Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns, spot anomalies, test hypotheses, and check assumptions with the help of summary statistics and graphical representations.

S No.	Column Name	Type
1	BTC Closing Price	Target Variable
2	BTC Volume	Feature Variable
3	ETH Closing Price	Feature Variable
4	ETH Volume	Feature Variable
5	USDT Closing Price	Feature Variable
6	USDT Volume	Feature Variable
7	BNB Closing Price	Feature Variable
8	BNB Volume	Feature Variable

A. BTC close Plot

B. 2019 bitcoin prices

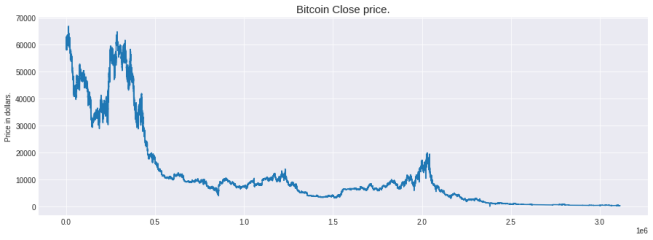

Fig. 3. close Plot

C. Different coin prices

Please refer to 5

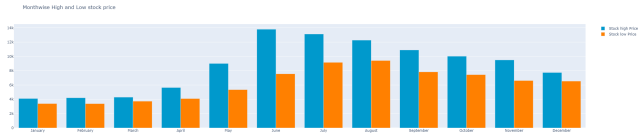

Fig. 4. 2019

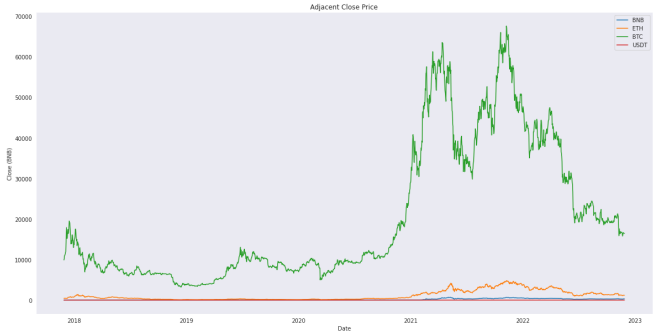

Fig. 5. four crypto currencies

D. Histogram Plot

Please refer to Fig. 6

E. Lag Plot

Please refer to 7

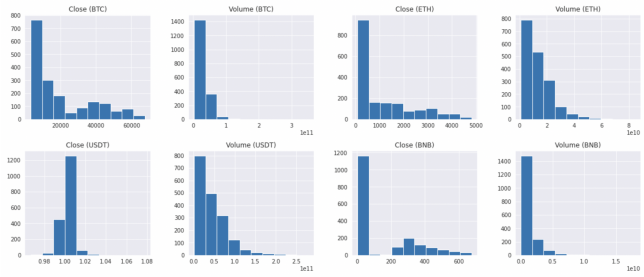

Fig. 6. BTC Histogram Plot

F. Correlation Plot

Please refer to Fig. 8

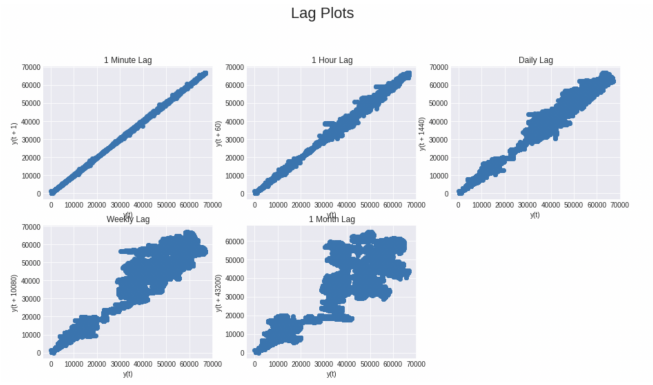

Fig. 7. BTC Lag Plot

G. Correlation Plot of 4 different coin data

Please refer to 9

H. Pair Plot

Please refer to 10

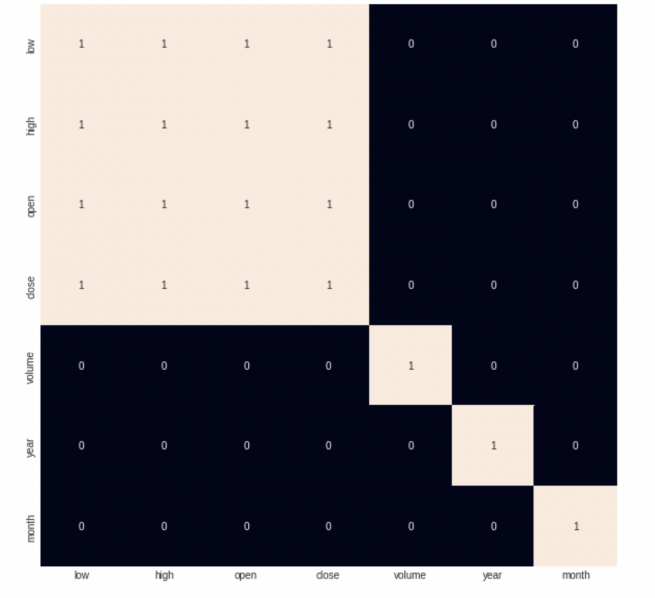

Fig. 8. BTC Correlation Plot

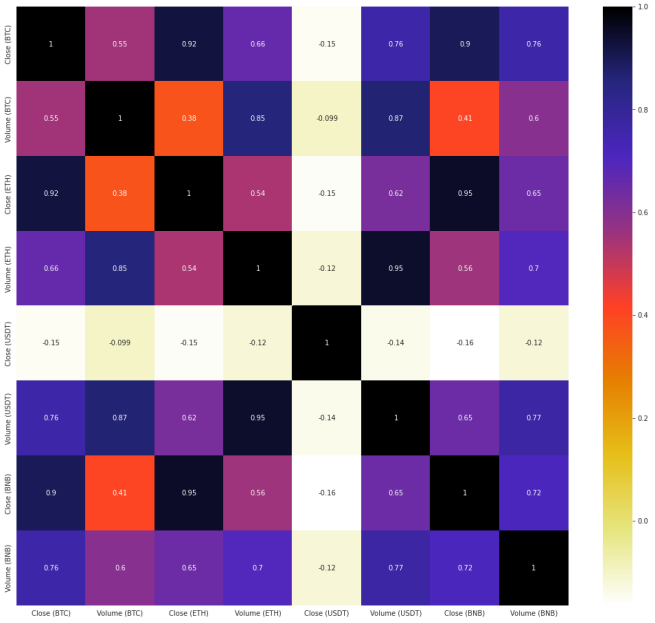

Fig. 9. Bitcoin Correlation Plot

I. Density Plot

Please see Fig. 11

J. Shap Feature importance

Please refer to 12

K. Shap model output

Please refer to 16

IV. Experimental Results

We have implemented the following Algorithms/models for the BTC Price Prediction.

- Autoregressive Model

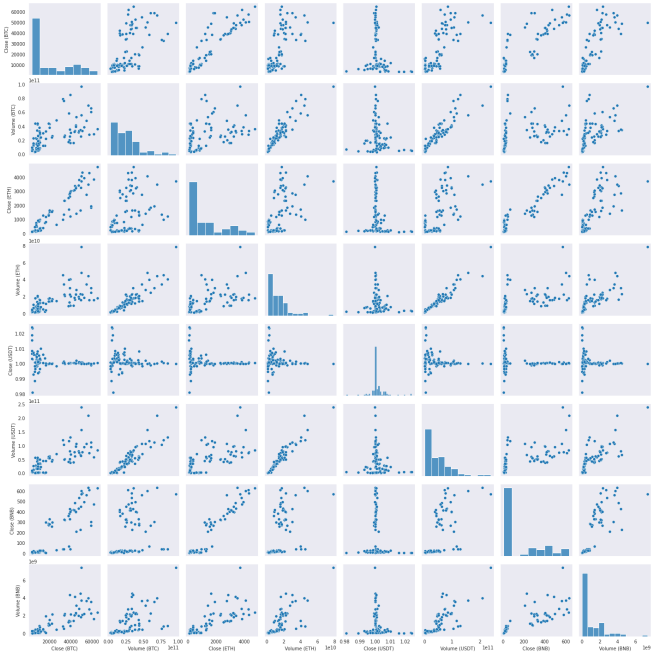

Fig. 10. Pair wise comparison

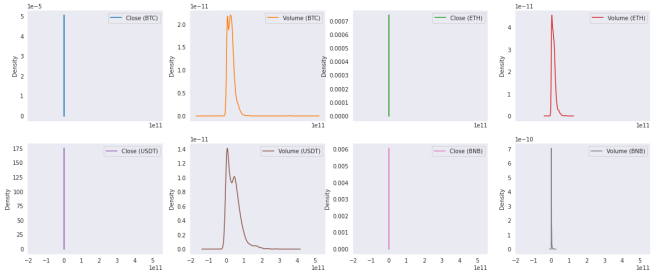

Fig. 11. Density Plot

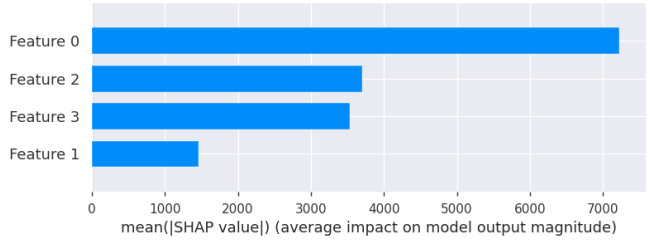

Fig. 12. Feature importance

- Moving Average Model
- Integrated Autoregressive Moving Average Model
- KNN Regression Model
- Decision Tree Regression Model
- Random Forest Regression Model
- Gradient Boosting Model

A. Auto Regressive Model

Time series forecasting is a distinct field of study since simple machine learning algorithms cannot be applied. In this article, the price of bitcoin is predicted using time series

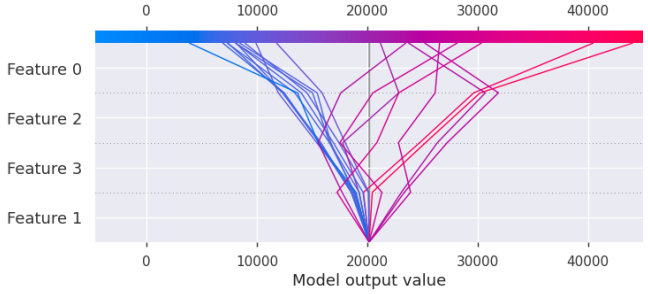

Fig. 13. Shap Output

models such as AR (Auto Regressive model), MA (Moving Average model), and ARIMA (Autoregressive Integrated Moving Average model).

B. Moving Average Model

Time series forecasting is a distinct field of study since simple machine learning algorithms cannot be applied. In this article, the price of bitcoin is predicted using time series models such as AR (Auto Regressive model), MA (Moving Average model), and ARIMA (Autoregressive Integrated Moving Average model).

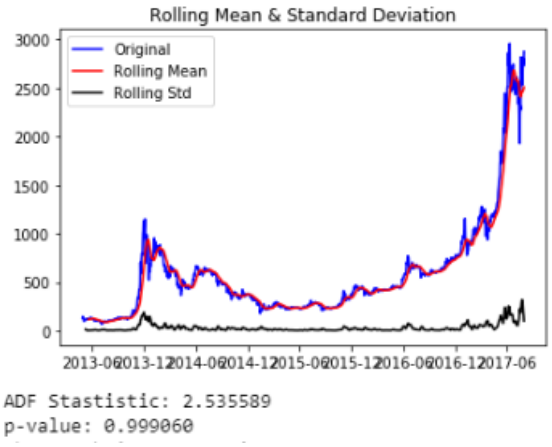

Fig. 14. Autoregressive Model

C. Integrated Autoregressive Moving Average Model

D. KNN Regression Model

It might very easily be applied to problems with both order and relapse. However, it is more frequently used in characterization-related business situations [10]. K nearest neighbors is a simple calculation that keeps all of the cases that are readily available and arranges new cases based on a majority vote of its k neighbors. The K nearest neighbors of the case being assigned to the class generally consider it to be normal. These capacities for separation may be of the Euclidean, Manhattan, Minkowski, or Hamming types. The first three capacities are used for constant capacity, and the fourth capacity (Hamming) is used for definite factors. In the event when K = 1, the situation is essentially consigned to

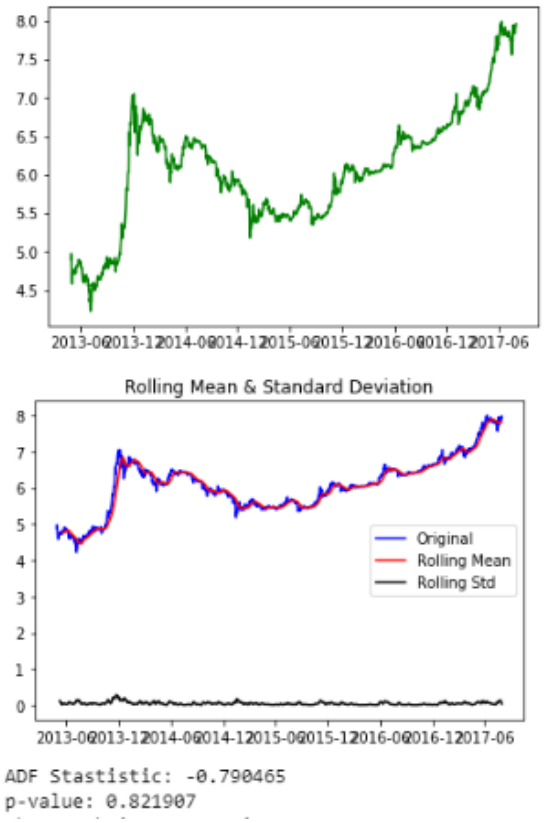

Fig. 15. Moving Average

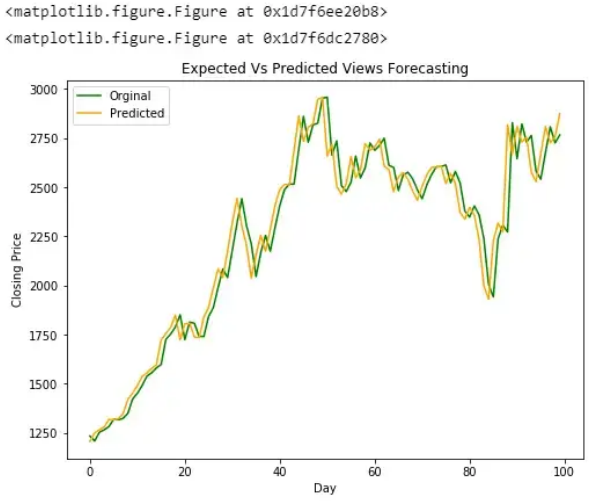

Fig. 16. ARIMA Model Prediction

```
rf = RandomForestRegressor(**rf_random.best_params_)
rf.fit(X_train, Y_train)

Y_pred_rf = rf.predict(X_test)
r2rf = metrics.r2_score(Y_test, Y_pred_rf)

print("-"*30)
print("Accuracy: ", r2rf)
print("-"*30)
```

```
------------------------------
Accuracy:  0.9682968644957287
------------------------------
```

Fig. 17. Random Forest Performance

Model	Temporal Length	Sensitivity	Specificity	Precision	Accuracy	RMSE
LSTM	100	37%	61.30%	35.50%	52.78%	6.87%
RNN	20	40.40%	56.65%	39.08%	50.25%	5.45%
ARIMA	170	14.7%	100%	100%	50.05%	53.74%

Fig. 18. All Model Comparison

Model	Epochs	CPU	GPU
RNN	50	56.71s	33.15s
LSTM	50	59.71s	38.85s
RNN	500	462.31s	258.1s
LSTM	500	1505s	888.34s
RNN	1000	918.03s	613.21s
LSTM	1000	3001.69s	1746.87s

Fig. 19. RNN Vs LSTM Comparison

the class of its closest neighbor. Now and again, picking K ends up being a test while performing kNN displaying.

E. Decision Tree Regression Model

It is a type of directed learning computation that is mostly used to solve order problems. Surprisingly, it is effective for both blatant and reliable ward variables. We divided the population into at least two homogeneous groupings for this calculation. This is done dependent on most huge properties/autonomous factors to make as particular gatherings as could reasonably be expected.

F. Random Forest Regression Model

V. Evaluation

Table shows that the RNN had the lowest RMSE and the LSTM had the highest accuracy. This may be caused in part by the class imbalance in the ARIMA forecast's predictive component (the price tends to always increase). The high specificity and precision (specificity, precision = 100 was influenced by this. This indicates that it does a respectable job of spotting price direction shift but does not necessarily imply strong overall performance.

Below table contrasts various models to assess how well the models were trained. It used an Intel Core i7 running at 2.6GHz. An NVIDIA GeForce 940M 2GB GPU was used. Both were using an SSD with Ubuntu 14.04 LTS loaded. For ease of comparison, the RNN and LSTM were given an identical batch size and temporal length of 50.

VI. Conclusion

In this project we have performed EDA to Observe Trends in features used int BTC Price Prediction. Built and Tested the following models Statistical Models: Auto Regression, Moving Average and ARIMA, KNN, Gradient Boosting Regression,Decision Tree Regressor, Random Forest Regression, LSTM. We have also Added Model Explainability using

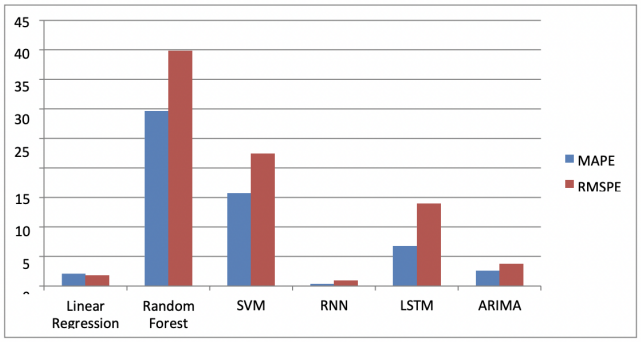

Fig. 20. RNN Vs LSTM Comparison

Shap API.

VII. Future Work

1. Doing Hyperparameter tuning on the ensemble models to get better performance.

2. Adding confidence data based on the expert suggestion to tweak the model in live conditions can be useful.

3. Adding Acceleration component for faster training and performance Boost.