Bitcoin Price Prediction

Surabhi Lambe
Atharv Patwardhan
Nachiket Pradeep Surti
Vijayalaxmi Salikar
Unnati Bhaskar Kini

Date: 15-07-2022

Feynn Labs

Bitcoin is amazingly transformative because it's the first time in the entire history of the world in which anybody can now send or receive any amount of money, with anyone else, anywhere on the planet, without having to ask permission from any bank or government.

-Roger Ver

Abstract
Cryptocurrency is a new sort of digital asset that has evolved as a result of advances in financial technology, and it has opened up substantial research opportunities. In this paper, LSTM deep learning model, time series ARIMA model, regression models such as linear regression, support vector machine, and decision trees have been used to predict bitcoin prices. Hyperparameter optimization and which model performs better predictions have been discussed. The importance of having these models is that they can have significant economic ramifications by helping investors and traders to pinpoint crypto currency sales and purchasing.
As a plan for future work, a recommendation is made to investigate other factors that might affect the prices of the crypto currency market such as social media, tweets, and trading volume. The future price of bitcoin depends on whether digital currencies can serve as useful financial assets. Major companies that already accept bitcoin include Microsoft, AT&T, Starbucks, PayPal, and Amazon.

Prototype selection
Bitcoin is a decentralized digital currency that can be transferred on the peer-to-peer bitcoin network. Bitcoin transactions are verified by the network nodes through cryptography and recorded in a public distributed ledger called a blockchain. The cryptocurrency was invented in 2008 by an unknown person or group of people using the name Satoshi Nakamoto. The currency began to use in 2009 when its implementation was released as open-source software.

The Bitcoin scheme carries attributes of a payment system in that it facilitates the transfer of value between parties. Unlike traditional payment systems, which typically involve the transfer of value denominated in a sovereign currency such as the US dollar, Bitcoin has its own metric for a value called a bitcoin.

History
The domain name bitcoin.org was registered on 18 August 2008. On 31 October 2008, a link to a paper authored by Satoshi Nakamoto titled Bitcoin: A Peer-to-Peer Electronic Cash System was posted to a cryptography mailing list. Nakamoto implemented the bitcoin software as open-source code and released it in January 2009. As mentioned before, Nakamoto's identity remains unknown.

Why is Bitcoin Trending?
Bitcoin is one of the world’s best-known digital currencies and has enjoyed huge success over recent years. This is a currency that has increased in value as well as in usage, and experts believe that this is for a number of key reasons. Bitcoin is classed as a safe, appreciating, and decentralized digital asset, which has helped to aid its popularity as well as its value. The no-VAT ruling in Europe has also helped to enhance the popularity and value of this digital currency.

One of the key driving forces behind the increased value of Bitcoin is its widespread usage across the globe, with many now opting to make payments via Bitcoin rather than using other methods. For consumers, being able to use Bitcoin provides increased ease and convenience when it comes to online transactions with the added bonus of security and peace of mind – something that you do not always get when you use payment methods such as credit and debit cards due to the higher risk of fraud.

Prototype Development

We have used 5 different models on the data set. From the very simplest Linear Regression to the most trending LSTM model.

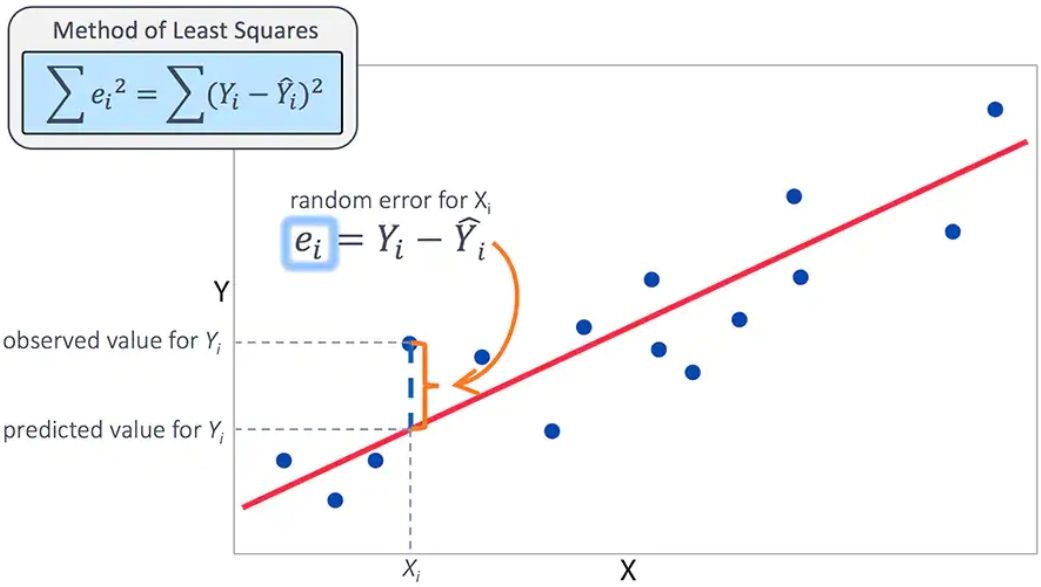
Linear Regression

Linear regression: One of the most commonly used methods of regression.If X1, X2, X3…. Xn is the independent variable and Y is the target variable. We need to predict Y given X's. The method of linear regression is a statistical model that fits a linear relation of Y with X's. The relation is linear in the coefficient of X's.In a 2-Dimension, the equation is a line. In 3-Dimension, the equation is a plane and in higher dimensions, it is called a hyperplane.

Why Linear Regression?

A regression-based model can be used as a simple baseline model that can be built relatively easily.

Interpretation of the model is often quite straightforward as compared to other more powerful models which tend to be black boxes



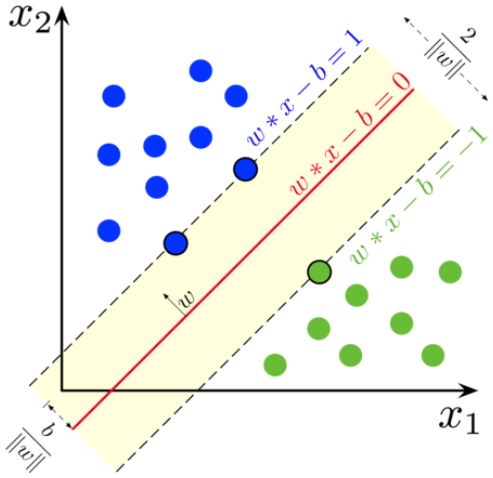
Support Vector Machine

SVM lies in the class of supervised algorithms used for classification. Let's start with a 2 class example:

Given classes X1 and X2, we want to find the decision boundary which separates the 2 classes the best i.e. with minimum error.

SVM does this with a 'Hyperplane'. Now this hyperplane can be a single line in case of 2-dimensional data and can be a plane in 3-dimensional one.

Support Vector Machines uses the concept of 'Support Vectors', which are the closest points to the hyperplane. The Margin is defined with the help of the Support Vectors (hence the name). In our example, Yellow stars and Yellow circles are the Support Vectors defining the Margin. The better the gap, the better the classifier works. Hence support vectors play an important role in developing the classifier. Every new data point in test data will be classified according to this Margin. If it lies on the right side of it, it'll be classified as a Red circle otherwise as a Blue star. The best part is, that SVM can also classify non-linear data.

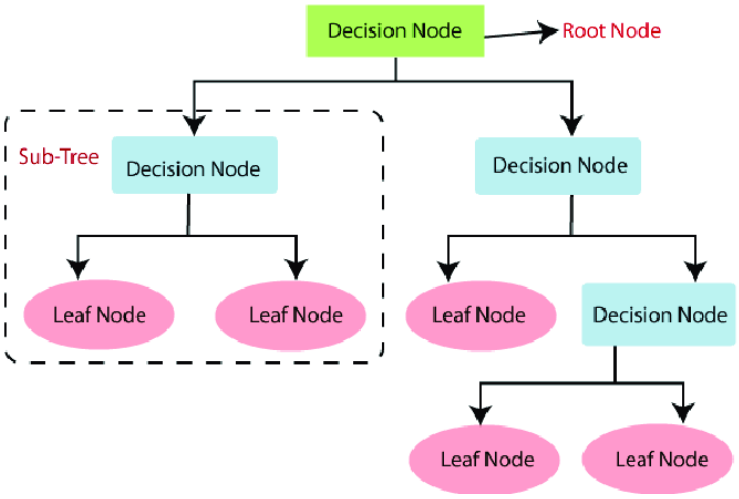


Decision Tree Model

A decision tree is the most powerful and popular tool for classification and prediction. A Decision tree is a flowchart-like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label.

Construction of Decision Tree :

A tree can be "learned" by splitting the source set into subsets based on an attribute value test. This process is repeated on each derived subset in a recursive manner called recursive partitioning. The recursion is completed when the subset at a node all has the same value of the target variable, or when splitting no longer adds value to the predictions. The construction of a decision tree classifier does not require any domain knowledge or parameter setting, and therefore is appropriate for exploratory knowledge discovery. Decision trees can handle high-dimensional data. In general decision tree classifier has good accuracy. Decision tree induction is a typical inductive approach to learning knowledge on classification.

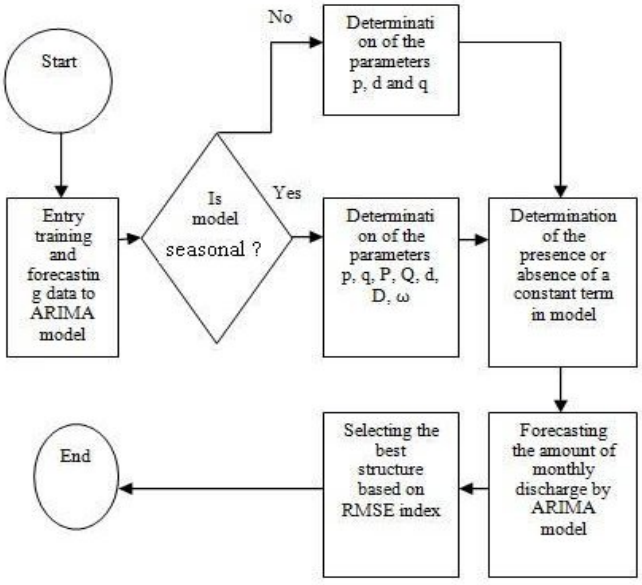


Arima Model

The ARIMA model has been used for analyzing time series data since the 1970s, and there are good reasons that it has stuck around; it is simple and powerful. In this blog post, my goal is to give you a solid foundation for understanding this model and hopefully encourage you to use it for analyzing time series data.

The ARIMA model (an acronym for Auto-Regressive Integrated Moving Average), essentially creates a linear equation that describes and forecasts your time series data. This equation is generated through three separate parts which can be described as:

AR — auto-regression: equation terms created based on past data points
I — integration or differencing: accounting for the overall "trend" in the data
MA — moving average: equation terms of error or noise based on past data points
Together, these three parts make up the AR-I-MA model.

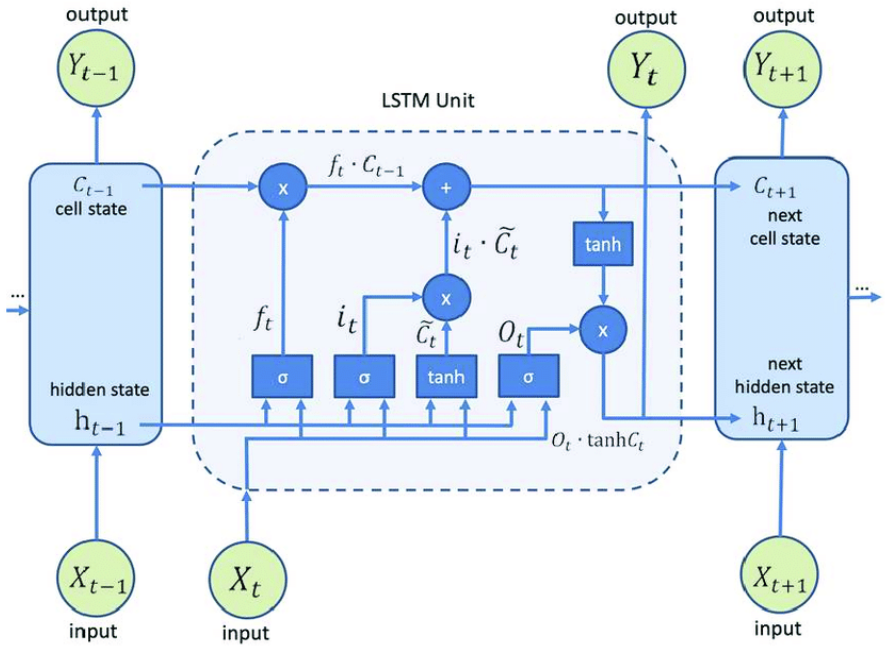


LSTM Model

The LSTM Network model stands for Long Short Term Memory networks. These are a special kind of Neural Networks which are generally capable of understanding long-term dependencies. LSTM model was generally designed to prevent the problems of long-term dependencies which they generally do in a very good manner.

The LSTM Network models generally have the potential to remove or add data carefully which is regulated by a special structure known as gates. The first step in processing LSTMs is to determine what information we need to throw from the cell.

The next step is deciding what new information we should store in the cell. At last, we decide what we want as an output. The output is generally based on the state of cells.



Dataset

The dataset chosen in this paper is from the year 2021 (BTC-2021min).

Link to the dataset:

https://www.kaggle.com/datasets/prasoonkottarathil/btcinusd?select=BTC-2021min.csv

Libraries used

Pandas, numpy, matplotlib, datetime, sklearn, statsmodels, keras.

Data Analysis

The dataset contains 610782 observations and 9 columns.

The columns are unix, date, symbol, open, high, low, close, Volume BTC, and Volume USD.

No NULL values are present in the dataset.

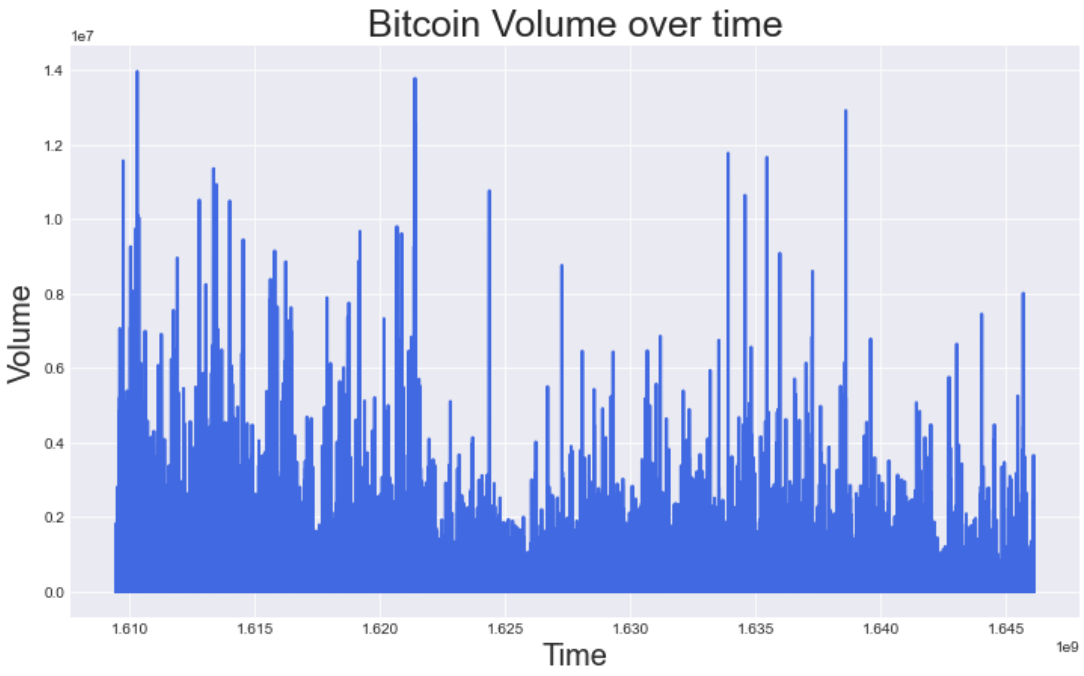
The first 5 rows of the dataset:

	unix	date	symbol	open	high	low	close	Volume BTC	Volume USD
0	1646106180	2022-03-01 03:43:00	BTC/USD	43046.58	43046.58	43046.58	43046.58	0.000000	0.000000
1	1646106060	2022-03-01 03:41:00	BTC/USD	43018.23	43046.59	43018.23	43046.58	0.142977	6154.673021
2	1646106000	2022-03-01 03:40:00	BTC/USD	43022.24	43022.24	43016.03	43016.03	0.009230	397.037957
3	1646105940	2022-03-01 03:39:00	BTC/USD	43035.16	43035.16	42999.44	42999.44	0.820950	35300.390268
4	1646105880	2022-03-01 03:38:00	BTC/USD	43077.82	43077.82	43049.46	43049.46	0.022210	956.143143

```
df.describe()
```

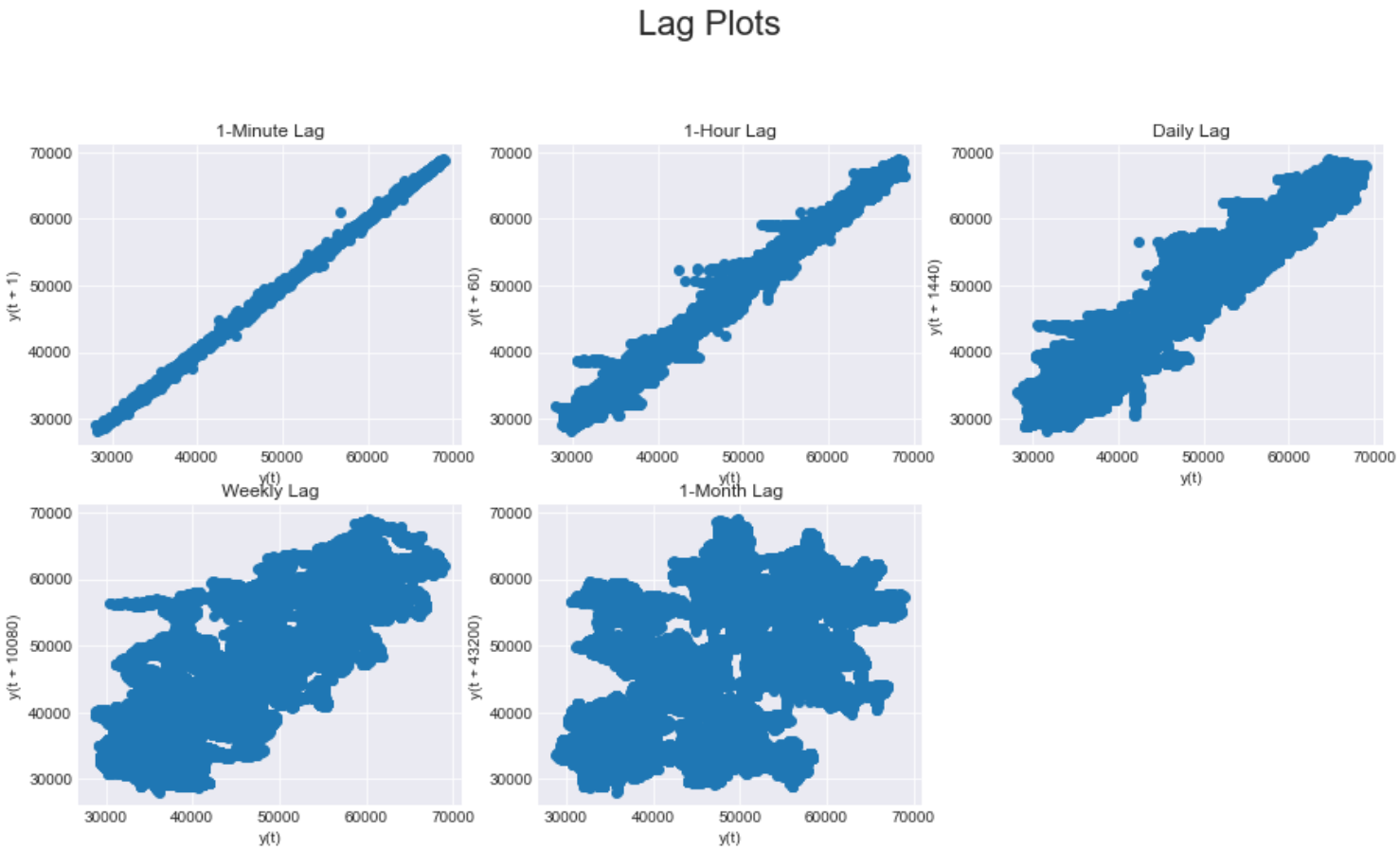
	unix	open	high	low	close	Volume BTC	Volume USD
count	6.107820e+05	610782.000000	610782.000000	610782.000000	610782.000000	610782.000000	6.107820e+05
mean	1.627783e+09	46482.265381	46509.783597	46454.085712	46482.567211	3.165254	1.389474e+05
std	1.057906e+07	9453.990285	9454.798088	9453.135638	9454.107853	7.893321	3.298068e+05
min	1.609459e+09	28093.810000	28397.520000	27734.000000	28073.030000	0.000000	0.000000e+00
25%	1.618621e+09	38129.052500	38157.882500	38099.437500	38129.137500	0.193192	8.859586e+03
50%	1.627783e+09	46666.960000	46697.745000	46636.710000	46667.295000	0.924971	4.190121e+04
75%	1.636944e+09	54907.797500	54938.655000	54877.405000	54907.412500	2.989888	1.349341e+05
max	1.646106e+09	69000.000000	69000.000000	68799.990000	69000.000000	398.565948	1.396792e+07

Bitcoin Volume over Time plot



Bitcoin Weighted Price Plot

Lag Plot



LSTM Model implementation

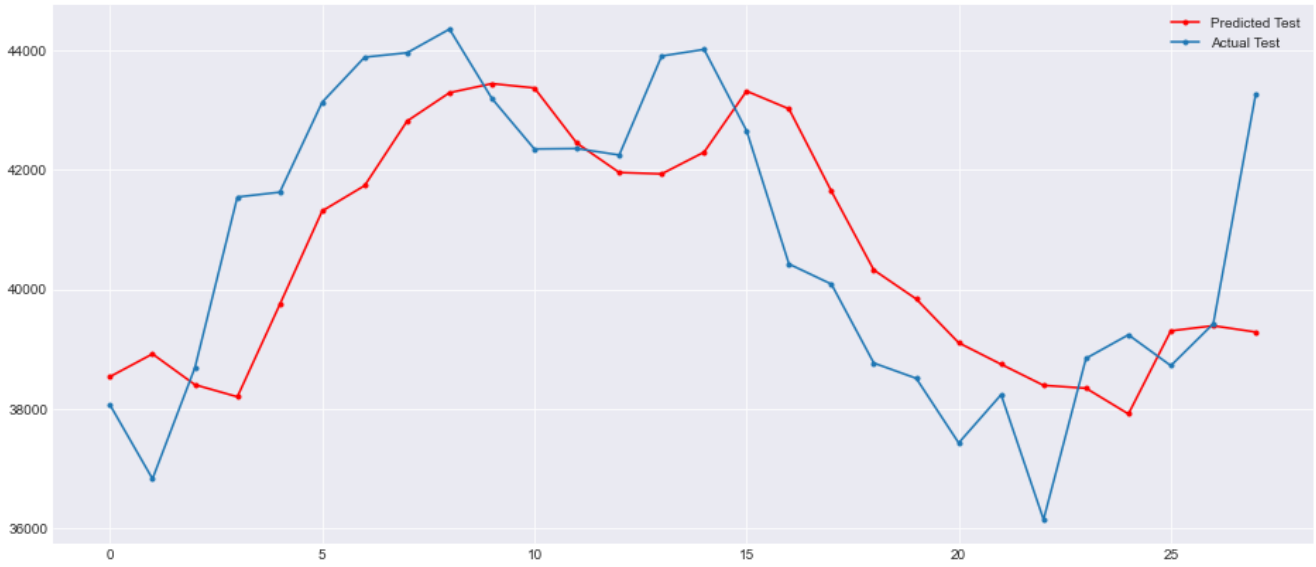
Dataset is first pre-processed and made ready to use for the LSTM model. The detailed and commented code is given in the GitHub code file.

The data is grouped according to dates and then split into train and test data. Min Max scaling is performed post train test split. After this, the input is reshaped into a 3-D tensor of [samples, time steps, features]. The LSTM model is trained using TensorFlow Keras.

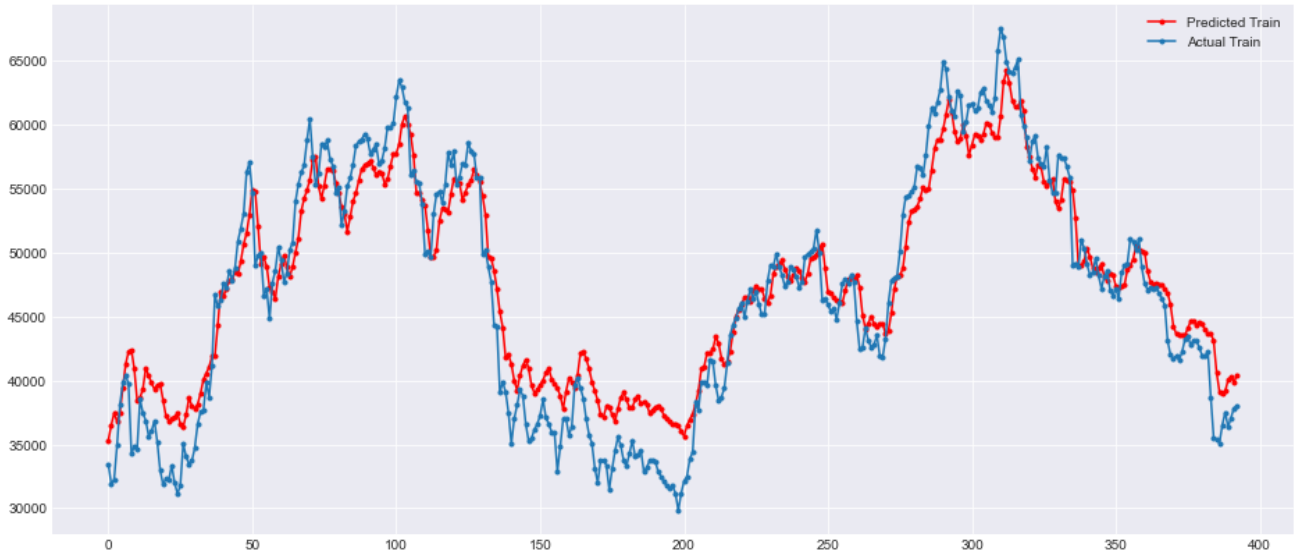
Minimum loss in the predicted data was obtained by using the tanh activation function, setting look_back to 2, and adding 2 layers. ReLu activation function also showed promising results with a loss of approximately 4%. The plots and main conclusions are as follows:

```
# Activation function = relu
# with 5-> Epoch 00012: val_loss did not improve from 0.07791 Test RMSE: 2293.292 Train RMSE: 4423.008      2 layers
# with 4-> Epoch 00015: val_loss did not improve from 0.05242 Test RMSE: 1881.073 Train RMSE: 3640.659      2 layers
# with 3-> Epoch 00015: val_loss did not improve from 0.05613 Test RMSE: 1946.466 Train RMSE: 4309.407      3 layers
# with 2-> Epoch 00017: val_loss did not improve from 0.04293 Test RMSE: 1702.252 Train RMSE: 3628.300      3 layers
# with 2-> Epoch 00018: val_loss did not improve from 0.04136 Test RMSE: 1670.942 Train RMSE: 2998.144      2 layers


#look_back = 2 and layers = 2-> Epoch 00017: val_loss did not improve from 0.04140


# Activation function = tanh
# look_back = 2 and layers = 2->Epoch 00016: val_loss did not improve from 0.04113
```

Predicted test vs Actual test



Predicted train vs actual train



RMSE (Root Mean Square Error)

```
rmse_lstm_test = math.sqrt(mean_squared_error(test_actual, predicted_btc_price_test_data))
print('Test RMSE: %.3f' % rmse_lstm_test)
```

```
Test RMSE: 1669.531
```

```
rmse_lstm_train = math.sqrt(mean_squared_error(train_actual, predicted_btc_price_train_data))
print('Train RMSE: %.3f' % rmse_lstm_train)
```

```
Train RMSE: 2988.623
```

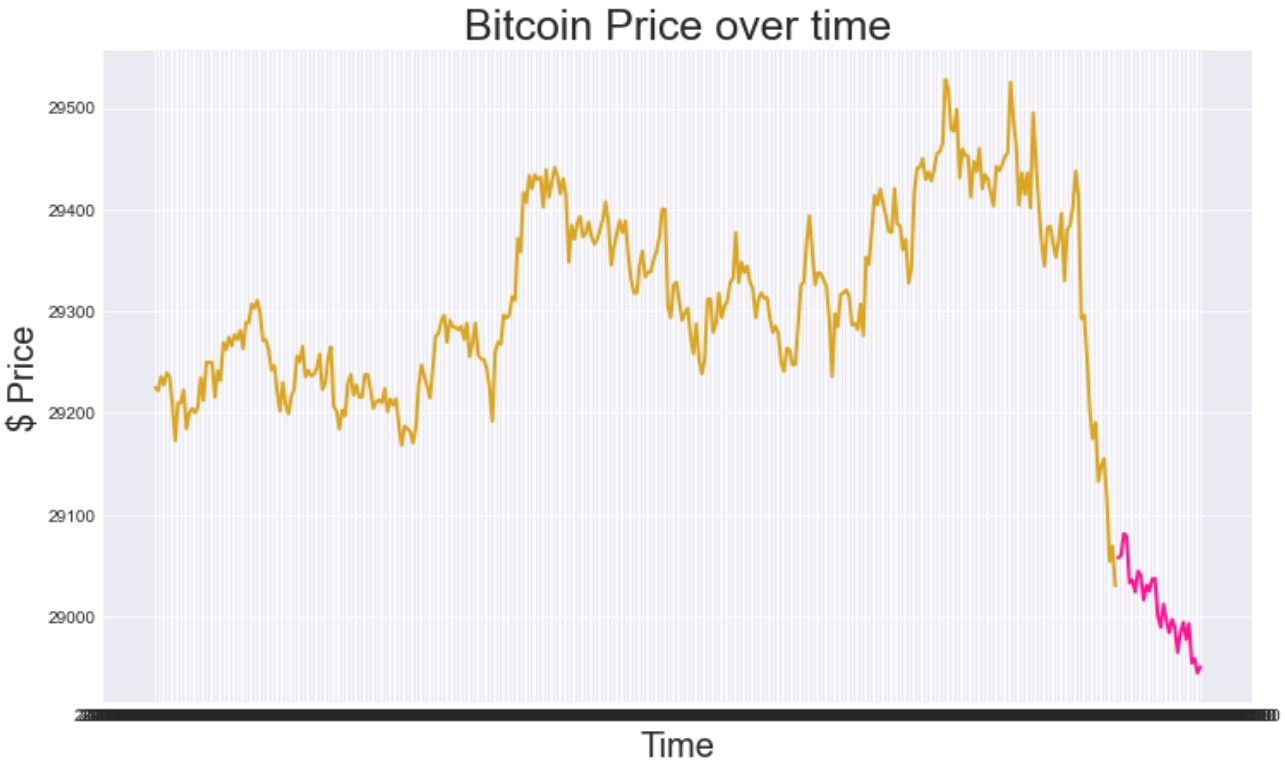
RMSE tells you how concentrated the data is around the line of best fit.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}$$

It indicates the absolute fit of the model to the data and provides the average model prediction error in units of the variable of interest. They are negatively-oriented scores, which means lower values are better.

Linear Regression model implementation

Bitcoin 'Price over Time'



Deep pink color values are the predictions. The test score obtained is 0.99.

Support Vector Machine model implementation

Predictions were made for the last 30 days for this model, but the loss obtained was more compared to models like linear regression.

```
model = SVR(kernel='rbf', C=1e3, gamma=0.00001)
model.fit(xtrain, ytrain)
```

Code

Github link: https://github.com/srl2002/Bitcoin_Price_Prediction.git

Business and Financial Modeling

As of 2022, over 15,000 multinational businesses are accepting payment in bitcoin worldwide. In general, people like the idea of a decentralized currency. Experts say that cryptocurrency payment methods will be universally adopted within the next 10 years. Bitcoin is also a good indicator of the crypto market in general because it's the largest cryptocurrency by market cap and the rest of the market tends to follow its trends. Bitcoin's price had a wild ride in 2021, and in November set another new all-time high price when it went over $68,000.

Looking at all these factors, many people have started investing in crypto and especially in bitcoin. Hence there is a need to have a handy market analyzer that can predict if bitcoin prices will rise or fall. This is where the Bitcoin Price Predictor comes in. It can be given in a subscription-based model to the customer. The customer will be charged on a monthly or weekly basis. The software can also be incorporated into big crypto trading platforms like coinbase, wazirx, etc. as an add-on service.

Conclusions

For the Bitcoin Price Prediction, we used the following models: LSTM, Linear Regression, SVM, Decision Tree, and Arima Model. The most suitable model for the prediction was Linear regression, with an accuracy of 99%. The other models cannot be used because, in LSTM, there was a 4% loss, and the SVM had a slow execution time due to the large data set compared to the other models and had an accuracy of 14%. All these models are prediction models, so the investment must be made at your own risk since the predictions are based on collected data and might not be accurate.

The disadvantage of LSTM-based RNNs is that they are challenging to understand and hard to predict how they behave. Rigorous hyperparameter tuning is also necessary to get acceptable results. Additionally, patterns change quickly and cryptocurrencies do not precisely follow historical data, which lowers the accuracy of predictions. Forecasting cryptocurrency prices is challenging because of price volatility and dynamism. The ARIMA model is unable to handle seasonal data due to the data's dynamic nature and strong seasonal influence.

The cryptocurrency was originally referred to as "rat poison squared" by Warren Buffett, whose Berkshire Hathaway company has provided investors with incredible profits over many years. Buffett vowed never to touch it. But Buffet has indirectly supported bitcoin by buying NuBank, a virtual "neobank" active in the cryptocurrency market.
Former Goldman Sachs chairman Lloyd Blankfein has declared that his views on bitcoin and other cryptocurrencies are "changing." The creator of Twitter, Jack Dorsey, left his position as CEO of that organization to manage Block, a pioneer in payment processing that is currently creating new uses for digital currency.

By the end of 2022, adoption by these and other corporate titans might help stop bitcoin's wild fall and increase its worth.

References:

1. Fitting the Multiple Linear Regression Model | Introduction to Statistics | JMP
2. Support-vector machine - Wikipedia
3. Structure of the decision tree (DT) algorithm. | Download Scientific Diagram
4. ARIMA (p,d,q) flowchart | Download Scientific Diagram
5. Basic structure of the LSTM model. | Download Scientific Diagram
6. https://www.kaggle.com/general/215997
7. Rajat Kumar Rathore, Deepti Mishra, Pawan Singh Mehra, Om Pal, AHMAD SOBRI HASHIM, Azrulhizam Shapi'i, T. Ciano, Meshal Shutaywi, Real-world model for bitcoin price prediction, Information Processing & Management, Volume 59, Issue 4, 2022, 102968, ISSN 0306-4573, https://doi.org/10.1016/j.ipm.2022.102968
8. Bitcoin Price Prediction 2022: Will It Recover? | GOBankingRates