Text

Description automatically generated

**BITCOIN PRICE PREDICTION AND TREND ANALYSIS**

INFO 6105

Prof. Handan Liu

Team 21:

Adina Nibijiang - 001549008

Kartheek Dabbiru - 001575541

12/08/2021

**Introduction:**

Bitcoin is a digital cryptocurrency, rather than paper money like US dollars that can be physically seen or touched. It is generated and stored in a computer through specific mathematical calculations. Bitcoin is called "virtual currency" because it is a digital encrypted currency with peer-to-peer (P2P) transmission. To put it simply, it can be understood that Bitcoin is a string of numbers with a cash value. Bitcoin is the world's first and the most successful blockchain application to date. Blockchain technology was born because of the birth of Bitcoin. With the development of blockchain technology, Bitcoin has become a subset of the blockchain, and both are in rapid development.

**Goals:**

As a team, we are looking forward to

* Predict and analyze the price and prediction of Bitcoin
* Provide some investment advice to those who need it
* Understand the future trends of the cryptocurrency Bitcoin

**Implementation Methodology**:

1. Preparation of Data
2. Exploratory Data Analysis
3. Feature Engineering
4. Model Selection, Training & Evaluation

**Preparation of Data:**

Before moving further into data analysis, the importance of data preparation is always the initial process - it leads to accurate insights (Chen, M. 2020). The following screenshots within Data Preparation are the steps we have taken to clean the dataset.

Import Data -

First, we import the dataset and check if the dataset is valid and is imported correctly.

Table

Description automatically generated

Check Data Records -

After importing the dataset, we then check the shape to see how many records this dataset consists of.

Text

Description automatically generated

This dataset consists of 4847377 rows and 8 columns

Check Columns Datatype -

It is always important to understand the type of data we are working with. Our dataset has records of the following data types

Table

Description automatically generated

Check the Summary of the Dataset -

Graphical user interface, text, application, table

Description automatically generated

According to the screenshot above, individuals can check the basic summary of each variable like Maximum value, Minimum value, and Mean value within the dataset across all the columns

Add New Column ‘Dates’ -

Since the Timestamp is hard to measure the correct dates, so we are converting Timestamp value to Date-Time format and adding the new column 'Dates'

Table

Description automatically generated

From the screenshots, we can see that the last date update of this dataset was done in 2021-3-31.

Checking for records with NA Values in the Dataset –

Text, letter

Description automatically generated

Instead of filling those records, we decided to remove the missing data from our datasets to make our datasets more accurate.

**Exploratory Data Analysis:**

It is an open-ended process where we develop statistics and figures to find a trend or relationship with the data. So basically, EDA helps us to know more about our data and that what can we infer and understand from it. These findings, information can help us to know what features we can choose in our model and find a scope of improvement in our way of feature selection.

Graphical user interface, chart, application, line chart

Description automatically generated

In the above plotted graphs, we can observe that price (both highs and lows) increase in the series starts from Dec 2017 to Dec 2018 and another price increase trend starts from Dec 2020 to Dec 2021, and this time it is a steeper slope. Before Dec 2016 there are no major changes in the value of Bitcoin.

Chart, histogram

Description automatically generated

I

n the above plot of BTC price over time, we can see that there is a top point in 2018, and another top point in 2021

Chart, line chart

Description automatically generated

In the above plot we can see that BTC Price started to increase from 2017.

Chart, line chart

Description automatically generated

In the above plot we can see that BTC Volume is gradually lowering since 2017 because many people started being aware of Bitcoin and also because price value of Bitcoin is getting increased.

**Feature Engineering:**

It is the process of taking raw data and choosing or extracting the most relevant features. It helps us to remove the features from the model that are not required, this helps us to create a better and more interpretable model. A machine learning model learns from the data we provide, so the data must contain every relevant information so that the output that is being predicted by the model is very accurate.

We use the DataFrame corr() method to show what features are correlated with each other and to what extent.

Table

Description automatically generated

Since the number above it's not as easy to view or interpret as a plot,

Here is the Correlations Matrix heatmap:

Chart, calendar

Description automatically generated

We then select ‘Open, High, Low, Volume\_(BTC), Volume\_(Currency), Weighted\_Price’ columns to fit our model as features, and we will also select the ‘Close’ as our target variable because we want to make a prediction on BTC price.

Then we split the data into two datasets as training and test sets



**Model Selection & Training:**

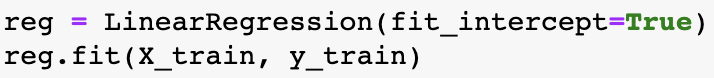
Our main goal is to train the best performing model possible, using the pre-processed data.

For the above dataset where we are going to predict the prices. This, clearly, is a problem that can be solved using regression.

So, we are going to use Linear Regression Model and Random Forest Regressor to train our data, predict the trends of BTC prices over time and then conclude which model is an optimum solution for training the data for the chosen dataset.

**Linear Regression Model**

First, we train the model by linear regression and predict the price for 300 days.



We then calculate the score of the linear regression model to see how the model performs on the test data.

Text

Description automatically generated with low confidence

It shows we have score > 0.99, which indicates that this is a good model to predict price of Bitcoin.

Here are the coefficients for our model from the above results. There are coefficients of each feature and the corresponding the feature names.

Table

Description automatically generated with medium confidence

In order to gain an understanding of how our model is performing, we scored the model against 3 metrics: R squared, mean squared error and root mean squared error. Following are the score results for the metrics mentioned above

Text

Description automatically generated

Since this dataset’s end record date is in March 2021, so we decide to predict price by shifting original dataset for 300 days.

Here is the plot we gain showing the previous Bitcoin prices and our predicted prices, in blue and pink. Blue representing the price from the existing data and Pink representing the predicted price.

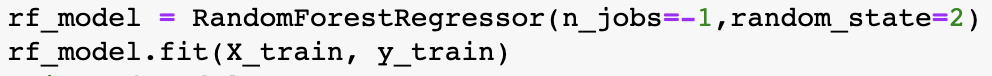
**Chart, line chart

Description automatically generated**

**Random Forest Regressor**

Our approach with Random Forest Regressor is going to be similar and is as follows.

We train the model by linear regression and predict the price for 300 days.

****

We then calculate the score of the Random Forest Regressor model to see how the model performs on the test data.

Graphical user interface

Description automatically generated with medium confidence

It shows we have score > 0.99, which indicates that this is a good model to predict price of Bitcoin.

In order to gain an understanding of how our model is performing, we scored the model against 3 metrics: R squared, mean squared error and root mean squared error. Following are the score results for the metrics mentioned above

**Text

Description automatically generated**

Here is the plot we gain showing the previous Bitcoin prices and our predicted prices, in blue and pink. Blue representing the price from the existing data and Pink representing the predicted price.

**Chart, line chart

Description automatically generated**

From the analysis and predictions from both the models, we can observe that the accuracy and predictions are strikingly similar since both the models are performing better in our project.

**Description of Dataset**:

In this project, we are using dataset extracted from Kaggle. This dataset is historical data of cryptocurrency transactions from period 2012-01 to 2021-03. We are using this data to predict cryptocurrency prices and perform trend analysis.

**Our data set consists of 8 attributes and they are as follows**:

Timestamp - Date (in Epoch Unix format) of data collection; It will later be

transformed into a "human" date for better understanding, Intervals of

approximately 1 in 1 minute

Open - Initial currency trading value in that measurement range.

High - Highest value reached by the asset during that measurement interval.

Low - Lowest value reached by the asset during that measurement interval.

Close - Value of the asset at the time of closing the measurement range.

Volume\_ (BTC). - Volume, in BTC, traded on Bitstamp during a given measurement

interval

Volume\_ (Currency) - Volume, in USD, traded on Bitstamp during a given measurement

interval

Weighted\_Price - Average asset price in that range; Calculated based on traded volumes.

It will be considered as the average price for analytical issues.

**Data Source:**

<https://www.kaggle.com/mczielinski/bitcoin-historical-data>

**Results and Analysis**:

In machine learning, different models have different performance levels. Choosing the right model to train our dataset can be very complex to interpret. To understand and analyze the models, we can find out the scores of the models using the test data and then come to a conclusion if that model is an optimum solution or not.

In predicting the prices and trend analysis of Bitcoin crypto currency, we have used Linear Regression model and Random Forest regressor model to train our data by fitting the models to our dataset. Let’s compare the scores of these two models that are obtained against the test data.

Score obtained from Linear regression model:

Text

Description automatically generated with low confidence

Score obtained from Random Forest regressor:

Graphical user interface

Description automatically generated with medium confidence

We can clearly notice that the scores obtained from both the regression models give a striking similar value which is greater than 0.99.

We can infer that both the models are good fit for predicting the price of Bitcoin.

To further differentiate between these two model approaches, let’s compare the scores of these two models against the metrics: R squared, Mean Squared error and Root Mean Squared error.

Metric scores for Linear Regression model:

Text

Description automatically generated

Metric scores for Random Forest Regressor:

**Text

Description automatically generated**

Again, we can clearly notice that the R squared scores for both the models are very close.  
Only comparing against the Mean squared error score and the Root mean squared error can we infer that one of the models perform better than the other.

Since the Mean squared error score and the Root mean squared error of Linear regression model are lower, we can conclude that the scope of error while using linear regression model is low in predicting the price Bitcoin. Hence Linear regression model can be considered as a better model in this case than Random Forest regressor by a slight margin.

**Conclusion**:

To summarize and conclude this project report, a data set containing various quantitative features of bitcoin has been taken from Kaggle dataset pool to perform machine learning algorithms and predict Bitcoin price and trend analysis.

Data has first been cleaned and engineered to make it easy to work with. Then the feature set and target sets have been initiated and the data has been trained using linear regression model and Random Forest regressor models. The resulting predictions (data) has been visualized by plotting graphs with features and clearly labeled legend.

After comparing the scores of two models, a conclusion has been made that Linear Regression model performs slightly better than Random Forest regressor by considering the error metric scores.

**References:**

Chen, M. June 30, 2020. What is Data Preparation and Why is it Important? Retrieved from https://blogs.oracle.com/analytics/post/what-is-data-preparation-and-why-is-it-important