BC04 – Cryptocurrency prediction model

Group AD

Nick Dafnides 20210648

Leonardo de Figueiredo 20210667

Haitham Abbas 20210625

Roeland Rensink 20200759

[1. Introduction 1](#_Toc826615862)

[2. Data Understanding 2](#_Toc527144568)

[2.1 Describing Data 2](#_Toc885051455)

[2.2 Exploration of data 3](#_Toc1545228868)

[3. Data Preparation 3](#_Toc740646195)

[3.1 Feature Engineering 3](#_Toc2066603252)

[3.1.1 Relative Strength index (RSI) 3](#_Toc480096926)

[3.1.2 Bollinger Bands 4](#_Toc825987639)

[3.1.3 Moving Average Convergence Divergence (MACD) 4](#_Toc2100973222)

[3.1.4 Stochastic Oscillator 4](#_Toc1962243187)

[3.1.5 Trend 4](#_Toc1067164576)

[4. Modeling 5](#_Toc1656011372)

[4. 1 Linear Regression 5](#_Toc75722432)

[4.2 LSTM 5](#_Toc281014165)

[4.3 Custom Time-Based CV split with Randomized Search CV and LGBM Regressor indicator 5](#_Toc1540155920)

[4.4 SARIMAX Model 6](#_Toc208722695)

[Appendix 6](#_Toc1357416956)

[A Total volume traded for each cryptocurrency 7](#_Toc2070961702)

[B Bitcoin price overtime including the volume 7](#_Toc61817248)

[C Bitcoin price forecast using LSTM model 7](#_Toc587308960)

[D MACD example 8](#_Toc1639102426)

# Introduction

In the end of the year 2020 cryptocurrencies got into a new bullish rally and the main cryptocurrency, Bitcoin, came to new all-time highs to approximately 67.000 dollars in the end of 2021. After years of low volume and few attentions to the crypto industry, attention rose again in the end of 2020. When the stock market crashed due to the global pandemic caused by a corona virus in March 2020, retail investors realized they needed to diversify their portfolios better. This triggered a new interest in the crypto industry, because cryptos tend to be uncorrelated to the stock market which is a great way to diversify a portfolio. The combination of new opportunity seeking investors and a great progression in the crypto industry, made a new surge for cryptocurrencies. Even though the prices of most cryptos rose, the application of cryptocurrencies as a stable digital currency in a large society has not been proven yet. This supports the idea that investors are still speculatively investing in cryptocurrencies. Just like any other financial asset, the value of a Bitcoin is worth whatever someone is willing to pay for it. Investing in cryptocurrencies is a great opportunity for risk-loving investors. But due to its volatile characteristic it happens to be that retail investors tend to trade with large emotions.

Our team has been contacted by Investments4Some which is a privately held hedge fund management firm. They use traditional statistical methods and financial indicators to measure the quality of their portfolios. In recent years they have been experiencing with Machine Learning models to start doing data-driven trading in cryptocurrencies. Due to the emotional tendency of retail investors and the large volatility there are great opportunities to build a market price forecasting algorithm. The goal is to build a model that predict the daily value of cryptocurrencies.

We have received datasets containing information about ten different cryptocurrencies. In the second part (Data Understanding) we will describe and explore the data. In the third part (Data Preparation) we will explain what we did to come to our finalized data frame, this also include the explanation of external information used for our model. In the fourth part we will explain three different models that we have chosen and the results that we got with them.

# Data Understanding

## 2.1 Describing Data

We got six csv files delivered to work with. In every csv file there are eleven columns. The first column of all files are the dates and in the other ten columns are the values for ten different cryptocurrencies, namely Cardano, Cosmos, Avalanche, Axie Infinity, Bitcoin, Ethereum, Chainlink, Terra, Polygon and Solana. The values for these columns differed for each file as follow:

* Adjustment close: Closing price after adjustments for all applicable splits and dividend distributions.
* Close: Price at the end of the day
* High: Highest price during the day
* Low: Lowest price during a day
* Open: Price at the start of the day
* Volume: Amount of assets that has been traded over the course of a day

Cryptocurrencies are trading all day long and do not have closing and opening moments like in regular stock markets. But when searching for answer we found that 12:00 AM is the open and 12:00 PM the close of the day.

## 2.2 Exploration of data

When checking for the dates we find that the date range is from 2017-04-26 / 2022-04-25. Furthermore, we see that there are many missing values for some currencies especially in the early stage. This is because some currencies did not yet exist in 2017 or 2018. Therefore, we decided to drop these rows that have missing values. In total we drop 6.442 rows from the total of 18.260. Since there were no values in these rows it will not affect our model. Due to melting the tables all together we will not lose any valuable information. For example, Bitcoin that does have valuable information in the years of 2017 will not be lost.

In appendix A we can see that the crypto’s with the highest volume traded are as expected Bitcoin and Ethereum. This also reflects into the trading price of these currencies. In Appendix B we take the example of how the price of Bitcoin has been developing over time.

# Data Preparation

To create a useful data frame for modeling we had to combine the six csv files together. To do this we first had to change the data type of the first column to datetime. After that we made sure that this column would be our index. For timeseries datasets it is useful to have the date as index. This is needed for the modeling and visualizations parts. In the end we could use a combination of the melt and merge functions to create one big data frame that consisted of 7 features, with the first being the name of the crypto currency and the other features being the six different values as stated earlier in the Describing Data part.

## 3.1 Feature Engineering

To predict a trend or future price there are many strategies to be practiced. Since we try to predict the price for only the next day or next two days we decided to focus on short-term measurements. In this part we will explain the extra features created.

### 3.1.1 Relative Strength index (RSI)

RSI – Relative Strength Index is used to determine whether an asset, in this case cryptocurrency is overbought or oversold. This allows the trader to take advantage before the market corrects itself. The way it works is calculations are preformed to find the range from 0-100. Typically, anything below 30 is considered undervalued, with anything over 70 and the asset is viewed as overbought.

Some interesting analysis we found was that one of the lowest dips of 2021, on July 20th, the closing price for BTC was 29807.35, with an RSI of 34. Now if we look at November 8th when the BTC price hit 67566.83, the RSI hit 68, which would’ve been a good time to sell.

### 3.1.2 Bollinger Bands

Bollinger Bands consist of a centerline, typically a simple moving average price, and two price channels, the lower and upper line. The channels are the standard deviations of the crypto above or below the centerline. These bands might expand or contract when the price becomes volatile. When the price is touching the upper channel (upper Bollinger Band), it indicates an overbought market. When the price is touching the lower channel (lower Bollinger Band), it indicates an oversold market.

### 3.1.3 Moving Average Convergence Divergence (MACD)

MACD – Moving Average Convergence Divergence is a momentum indicator that shows the relationship between two moving averages of a crypto currency. It is calculated by subtracting the 26-period exponential moving average (EMA) from the 12-period EMA. Without going to deep into the meaning of these EMA term we can explain how the MACD works. Traders use the MACD as a technical signal. When the 12-period EMA is above the 26-period EMA it is a signal to buy and vice versa when it is 12-period EMA is below the 26-period EMA it is a signal to sell. In appendix D there is an example where the blue line is the 12-period EMA and the red line is the 26-period EMA.

### 3.1.4 Stochastic Oscillator

The stochastic Oscillator is just like the others a technical indicator for finding signals of overbought or oversold markets. Its values are always between 0 and 100 and so has similar characteristics like the RSI. Values over 80 are corresponding to a market that is overbought and values under 20 are corresponding to a market that is oversold.

### 3.1.5 Trend

We believe that the hype and popularity of the crypto currency might have a positive correlation with its closing price, therefore we searched for a proxy that can explain this phenomenon. After carefully examining several metrics, we decided that the best metric to use is the google trend metric. The metric can have any value between 100 and 0, a value of a 100 will show the time a given term was searched for the most on google, and a value of 0 will show the time when a given term was not searched for on google. We found this metric to be significant and share a correlation of more than 50% with the dependent variable while experimenting with all our currencies, however we opted to use it only on the linear regression model for this phase, in order to further our experimentation to ensure its reliability as an independent feature.

# Modeling

## 4. 1 Linear Regression

Linear regression is one of the basic models, however it proves to have consistent results. Linear regression is an exceptionally good model to use when forecasting time series, to get a prediction, or to compare its results with different models. The degree of reliability of the linear regression model is very dependable on the features used to construct it. To get optimal results we did not use all the features we created while constructing the model, instead we selected features that did not have a remarkably high correlation with the dependent variable, which is the “closing price” variable in our scenario. Using variables that share high collinearity will introduce a problem known as multicollinearity, which will eventually lead to an overfitting model. After dropping the redundant variables, we fitted the model into our training and testing sets that we split carefully so we do not cause data leakage, and later we were able to calculate our predicted values for the target days. Comparing linear regression to the rest of the models, we will notice that the results are somewhat different, that may be because more features were dropped when constructing this model, also a new feature was introduced and used to predict the closing price, which is trend, however the differences were not vast and can be explained the feature choices.

## 4.2 LSTM

After researching the most widely used forecasting techniques, Long Short Term Memory (LSTM) Networks seemed to be quite popular for forecasting the various cryptocurrencies that we were tasked with analyzing. It was quickly realized that each coin needed to have an LSTM model with bespoke parameters as the results were varying greatly when only the use of a single model was implemented. After some inspection we noticed that due to most of the coins having varying amounts of past data, this was one of the parameters that needed to be tweaked to increase our forecasting accuracy (measured with MSE), this was resolved by changing the train test split and allowing models with less historical data to have a slightly larger testing set as this seemed to produce better results.

In terms of preprocessing for this model, the techniques we applied were a standard MinMaxScaler producing values from 0 to 1 (ŷ predictions were then inverted back at the end of the model), removing NaNs from our dataset and finally transforming the data into a numpy array so that it could be fed to the LSTM which was implemented through Keras.

A major issue that we had with applying this type of model to these coins was due to the varying nature of the time series as well as the different amounts of data for each currency, the models did not always converge. In order to combat this we applied a trial and error approach where we would mainly change the activation function as well as the batch size fed into the model per epoch. To a lesser extent the dropout rate was altered on an ad hoc basis to avoid overfitting.

We first carried out predictions over our test set where we compared actual historical prices to our predicted prices for the same past date range, then the models would be tuned until we were happy with them and finally, we would predict a date range that would include and end on the 10th of May (see appendix C).

## 4.3 Custom Time-Based CV split with Randomized Search CV and LGBM Regressor indicator

We knew we wanted to use some sort of timeseries split for modeling, as we can’t introduce random sampling to predict future values. At first, we tried to use a scikit-learn Time Series Split. However, we found that since it only lets us choose the number of splits, and not the set sizes, we would still be off in predictions due to the volatile nature of cryptocurrency.

What we found was a custom time-based cross validation split built by Or Herman-Saffar. Special thanks to her for adapting the time series split concept but adapting it for more real-world situations. For instance, we adapted and customized the model to take small units of training data, in our case 5 days, and have 2 days in our test set. The reason why we chose such small measurements was the ever-changing prices of cryptocurrencies. In addition, we didn’t start actually using the dataset until April 1st 2022. The reasoning again was due to the volatile nature of cryptocurrency.

The actual model used for predicting uses a Randomized Search Cross Validation with a Light GBM estimator. We ended up using all the original features, as well as the features that were created. We’re able to view a printout of the dates used for train and test easily. As well as scoring for the different splits, with additional validators like mean test score and std test score. This had to be adapted for each cryptocurrency. We then take the last split and predict the next 2 days, which in our case was May 9th and 10th. Just looking at the crypto prices for today, the 9th, we’re finding that the predicted values are higher than the actual.

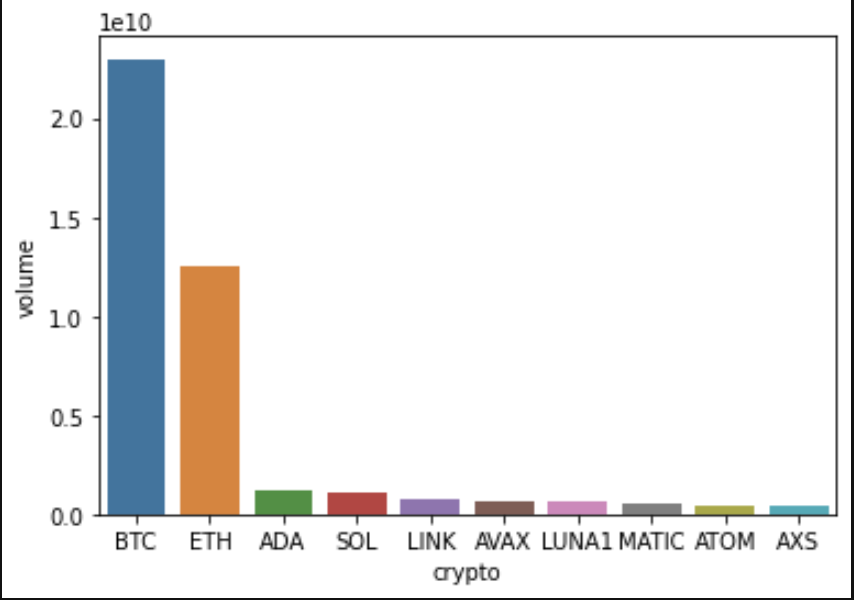
## 4.4 SARIMAX Model

We figured we should at least explore using some of the ARIMA models for this business case. This is due to ARIMA models are typically used for predictors in different types of financial trading. SARIMAX is both an updated ARIMA model as well as a seasonal equivalent model. The difference from a standard SARIMA model, is that SARIMAX also uses exogenous factors, with the autoregressive and moving average component in the model.

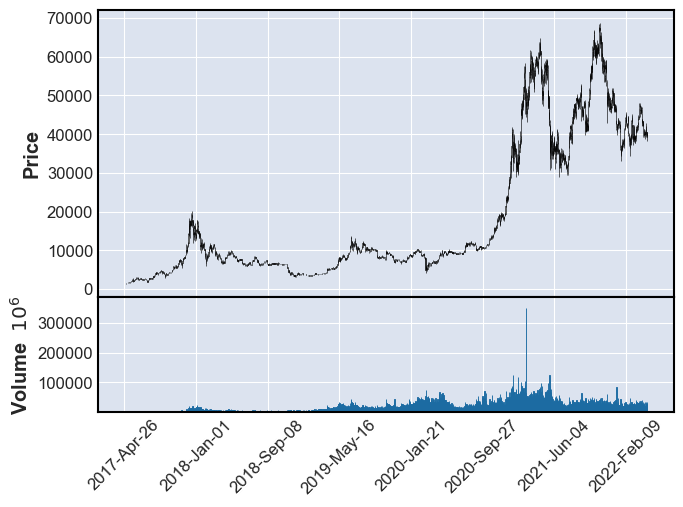
We ran into some problems however in using SARIMAX in that we couldn’t get the multivariate approach to work as we wanted. We ended up applying the model for Bitcoin as a comparison with other models. It compared with our Randomized Search CV model in terms of results, but we had more familiarity with other models, so we decided not to go further with using it for other cryptocurrencies.

# Appendix

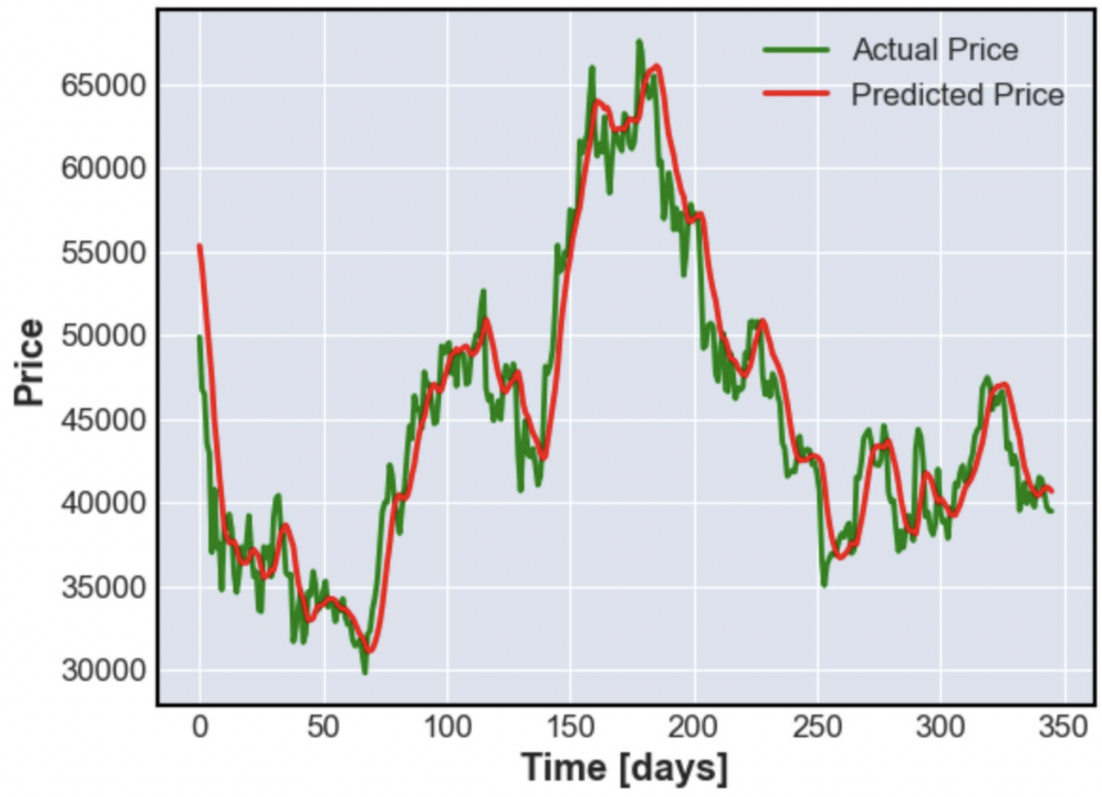
## **A** Total volume traded for each cryptocurrency



## **B** Bitcoin price overtime including the volume



## **C** Bitcoin price forecast using LSTM model



## **D MACD example**

