NOVA IMS Information Management School

BC4: Cryptocurrency Value Prediction

MASTER'S DEGREE PROGRAM IN DATA SCIENCE AND ADVANCED ANALYTICS

May 2022

Q Consulting

Laura Isabella Cuna, 20211312

Amelie Florentine Langenstein, 20210637

Tongjiuzhou Liu, 20211012

Nina Urbančič, m20211314

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

Table of Contents

1. Introduction

The idea of cryptocurrencies has existed since the late 1980s, but has gained major attention in the last decade, especially a few years after the introduction of the first decentralized cryptocurrency *Bitcoin* in 2009 (Jones, 2022). Cryptocurrencies allow the transaction of digital currencies without involving financial institutions and more or less have a limited supply, which is capped by mathematical algorithms. Due to this, no government agency or political body can dilute their value through inflation (Reiff, 2021). Also, the lack of regulations simplified international transfer of funds. However, the increased value of cryptocurrencies raised concerns about the speculation and therefore more restrictive regulations were implemented on a global level. This led to a decrease in value of many cryptocurrencies.

Cryptocurrencies still face looser regulations compared to traditional financial assets. They are also much more volatile than other financial assets and are therefore an exciting market for risk-loving investors. But even for experienced investors, it is a tough game as there is no real value behind most Cryptos and the market sentiment has a much higher influence on the value than on traditional assets like stocks.

Data-Driven Trading can help to increase the knowledge of investors by including different external factors and technical aspects into a forecasting model. Cryptocurrencies traders all around the world have been trying to find some patterns in the quickly changing value of cryptocurrencies to correctly estimate how the market will move in the future and generate some income from that. A forecasting model that predicts the values for the next days could help to improve their trading strategy significantly.

Also, the hedge funds management firm Investment4Some has been exploring Machine Learning Models for market price forecasting but were not able to complete or deploy a model yet. Therefore, they have asked us to build a Forecasting Model that helps them to anticipate the market sentiment and prices. In this way, Investments4Some alongside their Partner Warner Buffer can increase their expected returns on investment.

According to the CRISP-DM methodology, we first explored the business needs of Investment4Some and set the respective data mining goal. Then we moved on to exploring the datasets of the different cryptocurrencies provided. After merging the data into appropriate data frames, we conducted some transformation steps and created additional indicators. In the modeling part, we built and assessed different models for each cryptocurrency to be able to forecast the prices of each one for the consecutive days. Finally, we suggest a deployment plan.

2. Business Understanding

The Portuguese company Investments4Some is a privately-held hedge funds management firm that has been operating for many years. After using a more traditional approach to assess their portfolios, they started to explore Machine Learning approaches to forecast future market prices. However, they struggled to bring the models into production. And that’s where we come into play.

Business Objective
Investments4Some wants to stay on top of their game by accurately forecasting the cryptocurrency market trends and increasing the expected returns on their investments.

Data Mining Goal
The aim of this data mining project is to build a forecasting model that predicts the future highest prices of specific cryptocurrency values for the next few days based on historical data.

3. Data Understanding

The data provided consisted of 6 datasets, each representing one feature like *opening price*, *closing price*, *adjusted closing price*, *highest price*, *lowest price,* or the *volume* for the following cryptocurrencies: ADA, ATOM, AVAX, AXS, BTC, ETH, LINK, LUNA, MATIC, and SOL. Each dataset consisted of 1.826 entries represented by 11 columns, the date and 10 values.

The data gave us insight into the values and trading volume of each currency for every day of its lifetime until 25.04.2022. The historical data provided included prices from May 2017 onwards. But this only applies to Bitcoin. All the other cryptocurrencies were introduced later and show missing values until the date of the launch.

To better understand the price changes and trends of the given cryptocurrencies, we plotted a Candlestick graph with a Rangeslider for each of the cryptocurrencies. We also plotted line charts of the daily *highest prices* and *volumes* of all ten cryptocurrencies. Furthermore, we plotted boxplots with the daily difference between the *highest* and the *lowest price* for each currency, along with boxplots of the daily difference between the *opening* and the *closing price* for each currency. With the help of the boxplots, we found that there were no differences between the *closing* and the *adjusted closing prices*. The *adjusted close price* includes adjustments due to splits or the distribution of dividends, which is not applicable to cryptocurrencies.

We also plotted Candlestick graphs with Rangeslider for each currency with the price and the volume.

4. Data Preparation

In all 6 datasets we transformed the date column to datetime format. For our prediction, we decided to only use data since 01.01.2021 because not all cryptocurrencies existed before this date. We also dropped the column *adj_close,* the closing price after adjustments. Additionally, we renamed the columns and merged the data frames of all cryptocurrencies into one data frame.

Moreover we created three indicators: *close_off_high*, *volatility* and *day_diff*. The *close_off_high* represents the gap between the closing price and the highest price of that day. A value of -1 would mean the closing price was equal to the lowest daily price and a value of 1 equal to the highest daily price. The *volatility* represents the difference between the highest and lowest prices divided by the opening price. The *day_diff* represents the daily price change.
Based on the *Date*, we created the following new columns: *day, month* and *weekday.*

5. Modeling

Time Series Splits

We used the *TimeSeriesSplit*, and split data in three different ways:

First, we split the data according to the basic *TimeSeriesSplit* at fixed time intervals in train/test sets, where each split contained higher test indices than before.

Following that, we split the data in equal time intervals using the *Blocked Time Series Split* and used one part of each interval as train and the other part as test dataset. In this case, the data was trained on 89 days, and tested on 7 days.

Furthermore, we repeated the previous split using a *Blocked Time Series Split* for a shorter period but this time splitting it into shorter intervals. In this case, we trained our data on 45 days, and tested it on 3 days.

Build Model

We noticed that there is a correlation between stock market movements and movements in the cryptocurrency market, but decided to not include external data from stock market indices. First of all, the stock market is only open during weekdays and we would not be able to predict values for the cryptocurrencies directly based on daily data. Moreover, to use external data within our final model, we would need to predict the indicators, which would decrease the accuracy and stability of the model. So we decided to not include it in our model.

For modeling, we used the Bitcoin dataset and tried the following time series models on all three splits to predict the highest daily price: Simple Approach (Naive Approach, Simple Average, Moving Average), Exponential Smoothing (Simple Exponential Smoothing, Holt's Linear Trend Method, Holt-Winters Method), and SARIMA (Seasonal ARIMA). Based on the Root-Mean-Squared Error, the Holt's Linear Trend Method gave us the best results.

The supervised model that will predict the future prices of cryptocurrencies will be built on the three predicted indicators *close_off_high*, *volatility*, and *day_diff*. To find the suitable model for predicting each of these indicators, we used the models of Moving Average, Simple Exponential Smoothing and SARIMA.

We used the time series models to assess the indicators and predict them as input for the following supervised models: Linear Regression, Linear Model Bayesian Ridge, Random Forest Regressor and the MLP Regressor.

Assess Model

When building the model we applied different time series splits and machine learning algorithms and compared the performance according the respective RMSE scores.

The *Blocked Time Series Split* showed the best performance and was chosen for the final model. From the time series models, we chose Simple Exponential Smoothing, Holt-Winters Method, SARIMA as benchmark models, because they have a stable average RMSE score across the different Time Series Splits.

We used the updated data to test our final model. We used the three time series models (benchmark models) and MLP regressor to compare their performance and decided to choose MLP regressor as our final model.

6. Evaluation

We used the data of Etherum to test the best time series model (benchmark models) and the best supervised model trained with the Bitcoin data. Comparing the average RMSE scores, the supervised model always performs better than the benchmark model.

There are some limitations in the final model. The parameters could be optimized. We could also increase the number of indicators used to predict the prices of cryptocurrencies. It would be good to include additional information on the different cryptocurrencies in the model.

However, the main objective was to build a forecasting model that gives us prediction results to work on and that can give insights on how to improve the model in the future. In the next steps, we would suggest building individual forecasting models for each cryptocurrency.

7. Deployment

The model we've built can be applied by Investment4Some to help them predict the cryptocurrency values for the next 7 days. The model is applicable for all the 10 different cryptocurrencies. To obtain the most accurate values and results, it is advisable to run the developed model with values close to the day that has to be predicted.

We offer to create a Dashboard that displays the respective cryptocurrencies and their daily predicted values. We could also show the different indicators used for the prediction. The model that will be applied in the background should be automatized by using functions applied to the data from the selected time period. To incorporate new data, a database system could be created that retrieves the cryptocurrency data from Yahoo Finance on a daily basis and feeds the model with new data. The user can set the start and end date for the prediction period. To get an accurate result we recommend to use the model for a maximum of 7 days.

8. References

Jones, E. (2022, January 7). A Brief History of Cryptocurrency [Crypto Vantage]. *CryptoVantage*. https://www.cryptovantage.com/guides/a-brief-history-of-cryptocurrency/
Reiff, N. (2021, August 24). Why Should Anyone Invest in Crypto? *Investopedia*. https://www.investopedia.com/tech/question-why-should-anyone-invest-crypto/