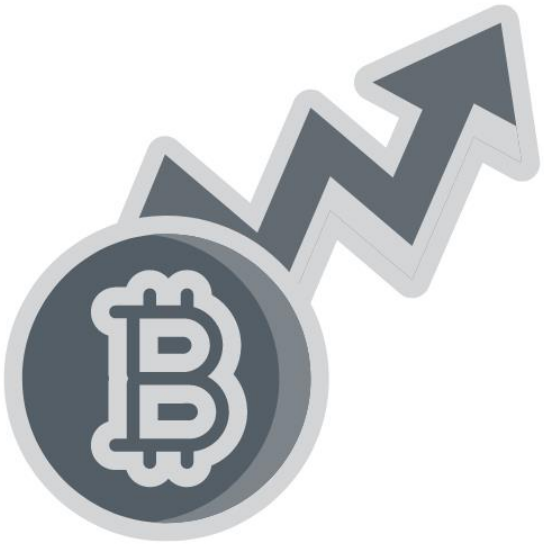
NOVA IMS Information Management School

BUSINESS CASE 4

MASTER'S DEGREE PROGRAM IN DATA SCIENCE AND ADVANCED ANALYTICS

Cryptocurrency Value Prediction



Group PN:

Mafalda Garcia, number: 20210763

Simão Pereira, number: 20210250

Tiago Santos, number: 20210548

Rui Ribeiro, number: 20211017

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

Table of Contents

1. Introduction

To carry out this project we followed a CRISP-DM methodology. Firstly, we start by reconnaissance the business, analyzing its objectives, which materials we had access to, and defining the data mining goals for this project. Next, we will perform recognition of the data itself, seeing its composition in detail and we will start the necessary steps for the data preparation, making all the necessary changes so that it is possible to apply the respective forecasting models to it. A crucial part of this project was done based on the materials available on GitHub and research on the links present in the references.

2. Business Understanding

2.1 Business Objectives

Investments4Some is a long-standing Portuguese, privately-held hedge funds management firm, that uses statistical methods and financial indicators to measure the quality of its portfolios. The company started exploring the implementation of Machine Learning models to apply to market price forecasting however, due to the lack of specialized workforce in the area, the company asked its partners Warner Buffer and Gil Bates for help, who in turn contacted us to build for them a forecasting model able to predict the daily value of cryptocurrencies, anticipate market trends and increase the expected returns of the investments.

2.2 Assess Situation

To build the forecasting model, the company provided us with a dataset containing the daily prices of 10 cryptocurrencies valued in terms of USD. The cryptocurrencies presented in the dataset were: ADA-USD: Cardano; ATOM-USD: Cosmos; AVAX-USD: Avalanche; AXS-USD: Axie Infinity; BTC-USD: Bitcoin; ETH-USD: Ethereum; LINK-USD: Chainlink; LUNA1-USD: Terra; MATIC-USD: Polygon; SOL-USD: Solana.

2.3 Data Mining Goals

The business objectives already mentioned can also be described technically, more precisely, like Data Mining goals. The main goal of this project was to build a forecasting model capable of predicting as accurately as possible future cryptocurrency values. We tested some different forecasting models to find out which model gave the most accurate results for each cryptocurrency.

3. Data Understanding

3.1 Collect Initial Data

The dataset was presented in 6 different CSV files with information about 10 cryptocurrencies and we loaded it into a Jupyter Notebook using the pandas library.

3.2 Description of data

Each CSV file contains information from 26-04-2017 until 25-04-2022 meaning a total of 1826 days.

For each of these cryptocurrencies the following data is present in the different CSV files:

- **Low:** Lowest price during a day

- **High:** Highest price during a day

- **Open:** Price at the start of the day

- **Close:** Price at the end of the day

- **Adj. Close:** Closing price after adjustments for all applicable splits and dividend distributions. The Data is adjusted using appropriate split and dividend multipliers, adhering to Center for Research in Security Prices (CRSP) standards.

- **Volume:** Amount of an asset or security that changes hands over a day

3.3 Exploring Data

When exploring our dataset, we used the .head() method to see a brief overview of the information, .info() to understand the various types of data we had to deal with and the .describe() to understand some statistics of our data.

3.4 Verify data quality

In the verification of our data quality with found some null values which means the absence of information on a certain coin on a certain day. The rest of the data had good quality and we did not find any duplicates or inconsistent values

4. Data Preparation

4.1 Selection and Cleaning of Data

After uploading the data into the Jupyter notebook we noticed that the data was pretty clean, the only problem we found was that some of the rows in the various data frames had Nans because some of the coins were created after the first date of these datasets.

4.2 Formatting the data

When formatting the data, we decided that we should split all the information from each coin into a separate data frame, for this we filtered the information about the coins from the datasets uploaded previously. We ended up having a data frame for each coin with the following columns: “Date”, “Open”, “High”, “Low”, “Close”, “Adj Close” and “Volume”.

In the data frames with nans, we used the .dropna() method to drop the rows with no information because it would not be good to have the nans when modeling.

4.3 Data Analysis

With the data in the shape that we wanted the only thing we noticed was that some of the data frames had really low numbers and others had high numbers, and this could influence our models in some way.

5. Modelling

5.1 Selecting modeling technique

To be able to build a predictive model capable of predicting accurately the future cryptocurrency value we tested 5 different time series forecasting algorithms for each cryptocurrency: Autoregressive (AR), Autoregressive Integrated Moving Average (ARIMA), Exponential Smoothing (ES), Long Short-Term Memory (LSTM), and XGboost.

To decide which was the best forecasting model to apply to each cryptocurrency we use the Root mean Square Error to calculate the forecast accuracy of the algorithm.

5.2 Building the model

We started by building all the 5 time series forecasting algorithms for each currency, however after we analyze the forecast accuracy of each one using the RMSE we ended up realizing the ARIMA, LSTM, and XGboost were the ones that gave the most accurate results so we will proceed to explain with a little more detail the building process behind each one.

Starting with the ARIMA(Autoregressive Integrated Moving Average) model is important to know that it explains the time series using past values (lags and lagged forecast errors) to forecast future values. Before starting the tests with this model, we filtered the data from the dataframes to use only the columns “Date” and “Close” for each coin, with the “Date” column as an index. After that, we verify the

stationarity of the time series (we need it to be stationary to be able to apply the time series forecasting model), and because it wasn't stationary, we ended up doing some transformations using the log and the differencing (to remove the trend and seasonality). After these transformations are completed and the RSS error calculated, we divide the values into train and test to be able to apply the ARIMA model and calculate the existing mean error between predicted and expected values. We repeated all of these procedures for each cryptocurrency.

The LSTM model is a recurrent neural network well-suited to making predictions based on time series data. The main idea behind LSTM cells is to learn the important parts of the sequence seen so far and forget the less important ones. Before starting the tests with this model, we filtered the data from the dataframes to use only the columns "Date" and "Close" for each coin, with the "Date" column as an index. We also transformed the "Date" column to datetime64 and used the MinMaxScaller in some of the dataframes to improve our results. To use this model, we needed to have an input and an output to train and then test our model. For this purpose, we created a list X of 60 days as input and Y was the output, so the model should predict the output, next day price, based on the previous 60 days' values. This process was made for the entire dataset and the final output was a prediction for the day after the last day of the dataframe. To find the best parameters for this model, we tried many different settings.

Finally, the XGBoost (Extreme Gradient Boosting) is an implementation of gradient boosting for classification and regression problems. Gradient boosting is an algorithm that helps in performing regression and classification tasks. Before starting the tests with this model, we filtered the data from the dataframes to use only the column "Close" for each coin. We also created a column "Target" with the next value from the list because similarly to the LSTM model we needed input and output. To find the best parameters, once again, we tried many different settings.

6. Evaluation

6.1 Evaluate results

Given we tried 5 algorithms to predict the cryptocurrencies' values, we need to find a way to compare the performance of said algorithms. For this, we used the Root Mean Squared Error measure to figure out which algorithm is the most appropriate for each cryptocurrency. This is a short-term performance evaluator which performs a term-to-term comparison between estimated and real values. It represents the distance, on average, of a data point from the fitted line.

We chose RMSE over MSE because it's more easily interpreted, given it has the same units as the real values. The table (Table 1) appended displays the RMSE that each algorithm got while predicting the last 50 instances of the database of each currency. Using this we conclude that these are the optimal algorithms to predict the respective cryptocurrencies:

• XGBoost: ETH.

• LSTM: AVAX, BTC, MATIC.

• ARIMA: ADA, ATOM, AXS, LUNA1, SOL, LINK.

6.2 Review process

We believe to have achieved good predictions using the selected algorithms. These are very useful if we consider daily values because we properly predict the variations in each day, and have an idea of the currencies' medium/long-term evolution.

6.3 Determine the next steps

The natural next step would be to continue the research to find a better algorithm. Considering that our algorithms can both predict trends and daily values, the more interesting next steps would be to gather more data about the currencies and find a model that uses that data besides the values.

7. Deployment

With the implementation of our models, Investments4Some will be able to forecast the prices of cryptocurrencies and understand the trends of these coins. To make this a more efficient process, we advise the company to create a robot that automatically updates the database daily so they can have the predictions for the following days.

In terms of investment strategy, we suggest buying/selling the cryptocurrencies in case there is a significant variation defined by the company much like a trading bot.

Additionally, it is important to remind the company to use the specific algorithms we defined for each coin.

8. Conclusion

Throughout this work and while trying many different models we believe that we made some good predictions on the cryptocurrency's daily values.

Although we have good models, we suggest that some more variables are added to the models because the price is only one of the many variables changing the value of the cryptocurrencies.

Given the algorithms that we determined are the most appropriate for each coin, we predicted the following values for these two days (09/05/2022 and 10/05/2022):

	09/05/2022	10/05/2022	Model
ADA	0.760986	0.762376	ARIMA
ATOM	16.64016	16.64782	ARIMA
AVAX	59.138844	58.3935	LSTM
AXS	29.2472	29.5249	ARIMA
BTC	30139.658	28939.113	LSTM
ETH	2705.9790	2658.0698	XGBOOST
LINK	16.5810	16.6297	ARIMA
LUNA1	68.3161	68.5914	ARIMA
MATIC	1.2485583	1.2459166	LSTM
SOL	80.11576	80.59112	ARIMA

Table 2. – Predictions for 09/05/2022 and 10/05/2022

APPENDIX

	AR	ARIMA	ES	LSTM	XGBOOST
ADA	0,14112	**0,04824**	0,15841	0,51090	0,14050
ATOM	6,569	**1,072**	5,426	7,202	2,990
AVAX	15,901	8,927	21,067	**0,192**	7,981
AXS	13,802	**3,278**	15,477	14,437	7,115
BTC	3 325,47	1 136,48	4 052,84	**162,97**	2 524,42
ETH	279,32	414,42	359,76	532,90	**210,80**
LINK	2,527	**2,109**	2,443	2,148	2,574
LUNA1	13,871	**5,209**	15,675	10,613	5,882
MATIC	0,31643	0,05881	0,27730	**0,02300**	0,18630
SOL	19,838	**5,304**	17,143	10,232	12,914

Table 1. – RMSE of each algorithm

References

ARIMA / SARIMA: https://www.analyticsvidhya.com/blog/2021/12/cryptocurrency-price-prediction-using-arima-model/

Autoregression(AR): https://towardsdatascience.com/how-to-use-an-autoregressive-ar-model-for-time-series-analysis-bb12b7831024

ARIMA: https://towardsdatascience.com/bitcoin-price-prediction-using-time-series-forecasting-9f468f7174d3

LSTM: https://machinelearningmastery.com/how-to-develop-lstm-models-for-time-series-forecasting/

XGBoost: https://towardsdatascience.com/xgboost-for-time-series-forecasting-dont-use-it-blindly-9ac24dc5dfa9