Forecasting Crypto

Business Case 4

Marjorie Kinney	*m20210647*
Bruno Mendes	*m20210627*
Lucas Neves	*m20211020*
Farina Pontejos	*m20210649*

Business Cases for Data Science

NOVA Information Management School

May 2022

Introduction

Forecasting techniques utilize historical data to make informed estimates and predict the determining direction of future trends. Moreover, forecasting is a sizable challenge as the success of the predictions can heavily rely on the amount of available data, time to analyze it, and the complexity of external events that influence the analyzed trends.

In this sense, our main goal will be to leverage these techniques to help Investments4Some predict cryptocurrency prices which, like many other financial assets, are particularly volatile. For that purpose, we'll employ a variety of machine learning algorithms and measure their prediction performance – all the while aiding them with technical analysis metrics.

Business Understanding

Investments4Some is a long-standing Portuguese, privately-held hedge funds management firm that uses traditional statistical methods and financial indicators to manage their portfolios of assets. Although, the company is aware of the potential of machine learning methods to anticipate market trends and increase the expected return of their investments.

Investing in cryptocurrencies can be exciting for many risk-loving investors. Cryptocurrencies are essentially digital currencies that allow individuals to transact online without the use of traditional financial institutions. This type of investing comes with increased risks associated with unregulated markets (i.e. high speculation) that can present a challenge even to the most experienced investors. Although, during the past year, cryptocurrencies lost a significant portion of their value as more regulations were enacted globally.

Apex Pattern Deployers has been tasked with creating models for predicting the price of ten cryptocurrencies to assist Investments4Some advise their clients.

Data Understanding

The source data consists of multiple csv files containing $USD values of the daily opening price, closing price, adjusted closing price, highest price, lowest price and volume for 10 different cryptocurrencies between April 26th 2017 and April 25th 2022. These cryptocurrencies are: ADA, ATOM, AVAX, AXS, BTC, ETH, LINK, LUNA1, MATIC, SOL.

There are 1826 rows per file, each representing a day as portrayed in the Date column which is common to all of them. The dates are in YYYY-MM-DD format.

In general, the currencies did not show much relative change in value until 2021, although Bitcoin, Cardano, and Ethereum had some movement in 2017. A time-series analysis indicates that all of the coins are non-stationary, and follow an ARIMA (1,0,0) model. Bitcoin and Etherium are the most highly priced coins by far, with Cardano being the coin with the lowest value.

In addition to giving us more insights about the data, the technical analysis also provides additional inputs for us to use in our machine learning models. See the Appendix for the results of the technical analysis.

Data Preparation

The data for adjusted closing price was not imported, since none of the typical stock price adjustments such as stock splits and dividends are needed for cryptocurrency.

A technical analysis was completed on the rest of the provided data. For each coin, measures of volume, volatility, trend, and momentum were calculated and graphed. Measures to include as exogenous variables for the algorithms were selected based on a review of the available literature. The following measures were selected: On-balance volume (OBV), average true range (ATR), moving average convergence divergence (MACD), and relative strength index (RSI).

Volume indicates interest in a given cryptocurrency. OBV uses volume to predict changes in price, using the logic that if the volume increases sharply without a significant change in price, the price will eventually jump upward or fall downward as a consequence. Volatility indicates the risk of a given cryptocurrency, based on the rate at which its value increases or decreases over time. ATR shows market volatility by using a 14-day moving average, with high values indicating more volatility and low values indicating less volatility. Trends indicate if the value of a given cryptocurrency is moving up or down. MACD deals with trends and can signal when it is best to buy or sell based on how far above or below baseline the exponential (recent) moving average is. Momentum indicates the strength or weakness in a cryptocurrency's price. RSI measures the magnitude of recent price changes, with 70 or above indicating potential overvaluation which may lead to a trend reversal, and 30 or below indicating potential undervaluation. The information for these indices for each coin were exported as a dataset to be used in the subsequent analysis.

In addition, several external indices were added to help train the models. These indices were selected based on the perceived usefulness for predicting cryptocurrency, according to information obtained through background research. The S&P 500, the Gold Futures, and the Cboe Volatility Index were selected.

We merged the data provided for each coin, the data from the technical analysis, and the data from the external indices into a single dataset, with the Date column as index. The rest of the columns were named as follows: "value_coin" (e.g. "btc_close"), or "coin_variabletype_variable" (e.g. "btc_momentum_rsi"). In this data-set we removed everything pre-2021 as that ensured that all the coins existed for easier comparison, and a lot of the coins had minimal movement until 2021. Due to the nature of dealing with time series data, we do not consider more records to be necessarily better for training the models in this case. This is because the data is heavily influenced by exogenous trends and events that impact the values and severely change the behavior of the variables.

Modeling

In order to ensure no data leakage, we were careful to avoid using a traditional cross-validation strategy. Instead, a moving window validation method was employed. After experimentation with several different values, a training size of 90 days was selected for training, a window size of 7 days was selected for the time series split, and a test size of 1 day was selected for the test. Only data from January 1, 2022 onward was used for the train/test split, as better results were obtained by limiting the data than by using all the available data. Data from April 2022 was reserved for prediction.

A baseline model was created in order to have a standard against which the other models should be measured. This model simply uses the previous day's price as a prediction for the current day's price. After creating this model, we tested several industry-standard machine learning regressors that are fit for forecasting situation problems such as the present case. They were:

- Decision Tree Regressor
- Multi-Layer Perceptron Regressor
- K Neighbors Regressor
- Gradient Boosting Regressor
- Extreme Gradient Boosting Regressor
- Linear Regressor
- Random Forest Regressor

A pipeline was created in which the data from each time series split was scaled, the best features and parameters were selected, the model's performance was scored, then the model was fit to the data.

In addition to using the models above, we also implemented SKforecast with two algorithms: the Decision Tree Regressor and the Extreme Gradient Boosting Regressor. In these cases, the results were not better than machine learning models on their own.

Evaluation

In order to properly evaluate our models while also being able to compare them, we chose the mean absolute percentage error (MAPE) as a metric. This is a relative measure commonly used to verify accuracy in forecasting by expressing the errors as a percentage of the actual data. In this sense, it easily portrays the extent or importance of errors in our models as our main goal is to get the closest predictions possible in each cryptocurrency.
Therefore, this metric is calculated as the average absolute percent error for each time period minus actual values divided by actual values.

In addition to calculating each model's performance, we also compared it to the baseline model described above, as well as to the models used with SKforecast. From all the models evaluated, the following were the best performing for each coin:

COIN	Model	MAPE
ADA: Cardano	GradientBoostRegressor	0.03222
ATOM: Cosmos	RandomForestRegressor	0.01944
AVAX: Avalanche	DecisionTreeRegressor	0.02494
AXS: Axie Infinity	GradientBoostRegressor	0.04438
ETH: Ethereum	MLPRegressor	0.02149
LINK: Chainlink	DecisionTreeRegressor	0.02693
LUNA1: Terra	GradientBoostRegressor	0.03155
MATIC: Polygon	MLPRegressor	0.02525
SOL: Solana	XGBRegressor	0.04079
BTC: Bitcoin	GradientBoostRegressor	0.01834

Table 1: Winning model MAPE for each cryptocurrency

Deployment

The following closing price predictions were made for May 9th-10th, 2022:

Coin	Closing Price in USD
ADA	0.88
ATOM	26.02
AVAX	65.15
AXS	50.76
BTC	47128.00

ETH	2833.01
LINK	14.35
LUNA1	58.52
MATIC	1.36
SOL	81.09

Table 2: Predictions

Please note that these predictions are based on models that were only trained on data that existed prior to April 26, 2022. If these models were retrained on more recent data, the predictions would likely be more accurate.

For the deployment of our proposed solution, the proposed best models for each coin can be updated on a daily basis by incorporating new data. This data could be web scraped after a predefined close time of the coin prices. Furthermore, for our solution to be scalable we also propose only keeping a predefined number of days prior to the desired prediction time in the models, in order to account for the high variability of the cryptocurrency landscape and improve the predictions.

Lastly, the deployment would be done through the use of a dashboard that can show real time key analytics of the referred coins for a specified time period, such as the current prices of the coins, the prices predicted by our models for the upcoming days and technical analysis key indicators.

Conclusion

Our models are able to predict the price of the ten cryptocurrencies with an average error of 2.85%. Since the predictions are for the price of the cryptocurrencies at close, it may be useful to have more granular data (for example, hourly data, rather than two data points per day). Since there is no close in this market, it may be more useful to predict the high or low instead of the price at a given hour (close).

In developing a trading strategy, it is also important to define a time horizon. Prediction windows should be adjusted based on whether the intent is to trade coins daily or with a longer view.

Appendix

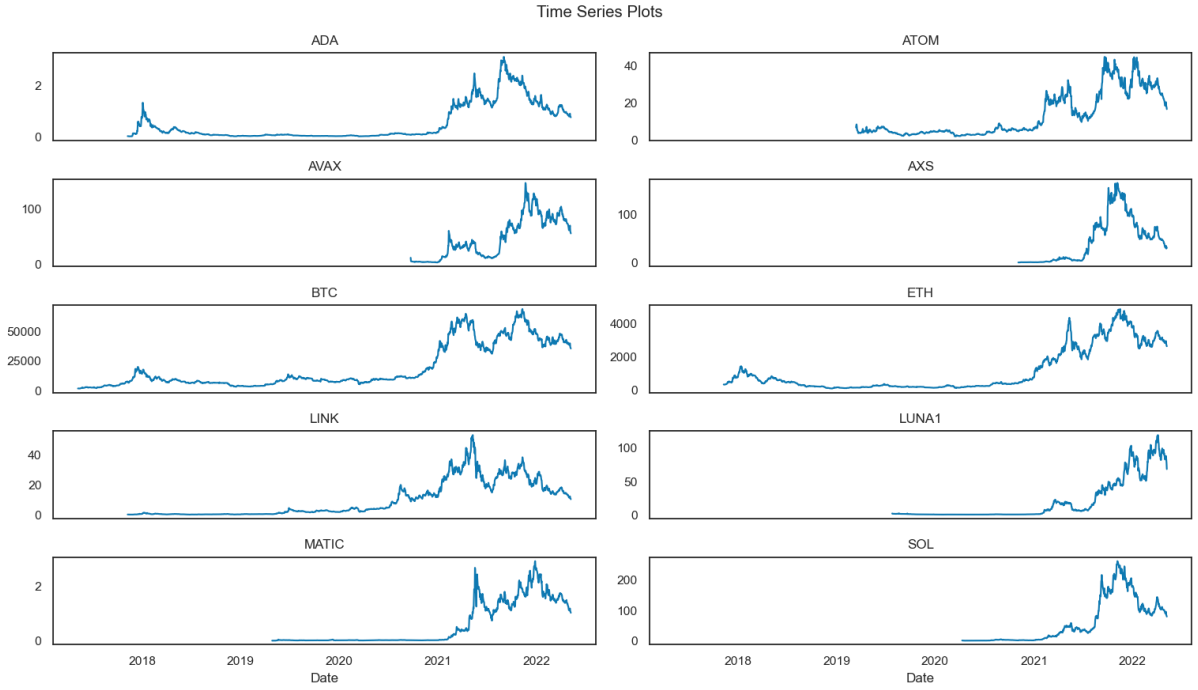

Figure 1: Time Series for Each Currency

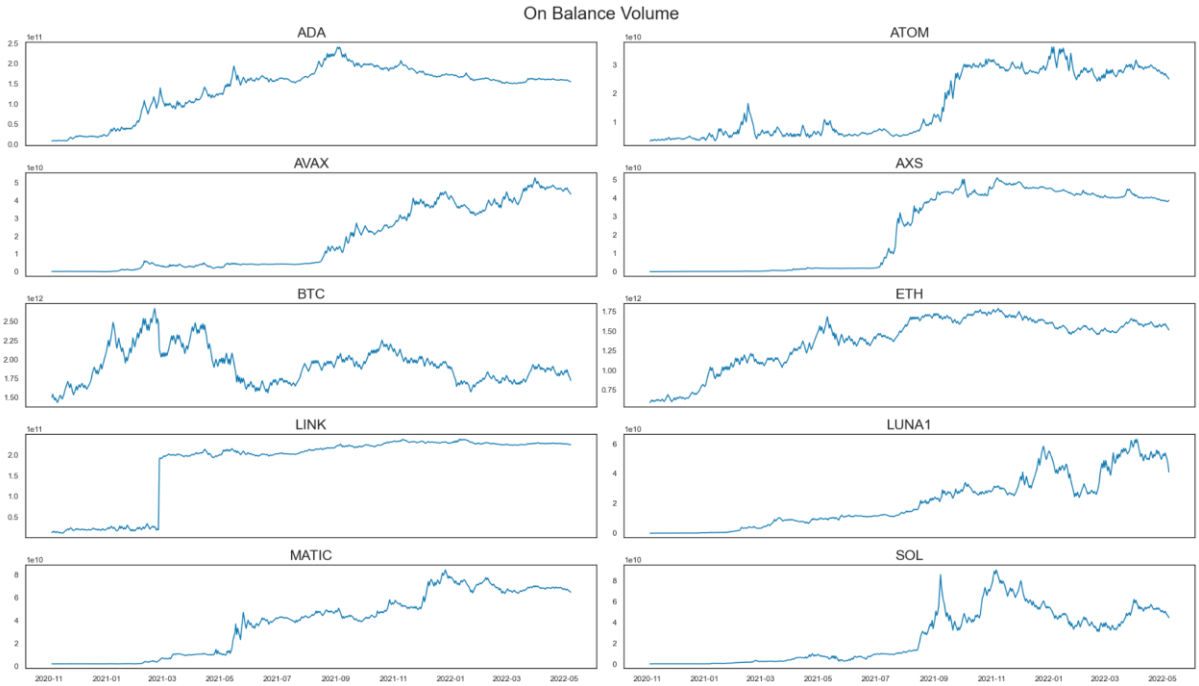

Figure2: Technical Analysis- Volume - On Balance Volume (OBV)

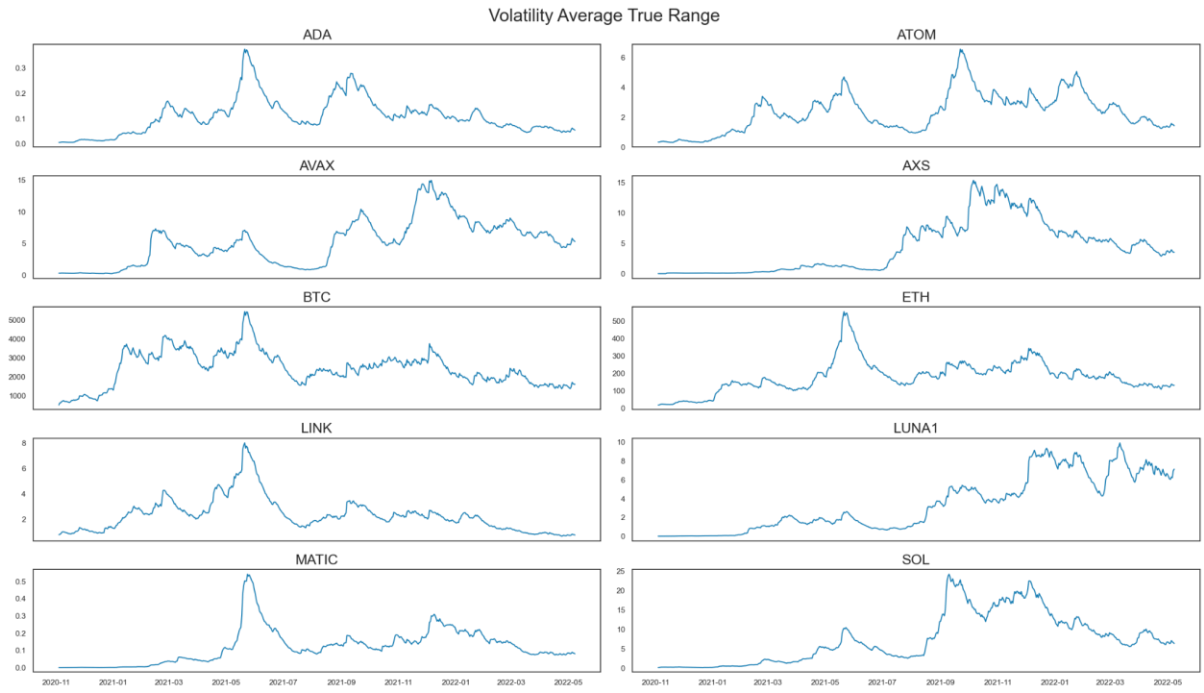

Figure 3: Technical Analysis- Volatility - Average True Range (ATR)

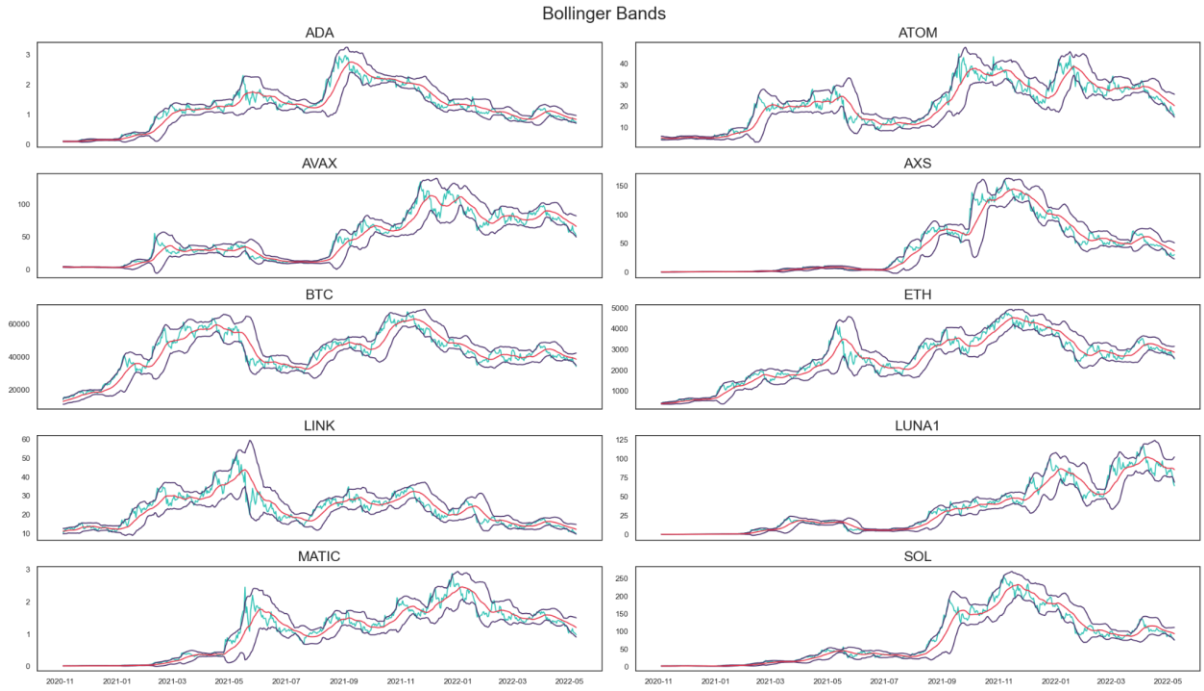

Figure 4: Technical Analysis - Volatility - Bollinger Bands (BB)

Figure 5: Technical Analysis - Trend - Moving Average Convergence Divergence (MACD)

Figure 6: Technical Analysis - Momentum - Relative Strength Index (RSI)