NOVA IMS
Information Management School

D A T A L I N
Bringing Intelligence

BUSINESS CASES FOR DATA SCIENCE

BC4: Cryptocurrency Value Prediction

MAY 2022

ALICE VALE R20181074
EVA FERRER R20181110
RAFAEL SEQUEIRA R20181128
RAQUEL SOUSA R20181102

Business Understanding

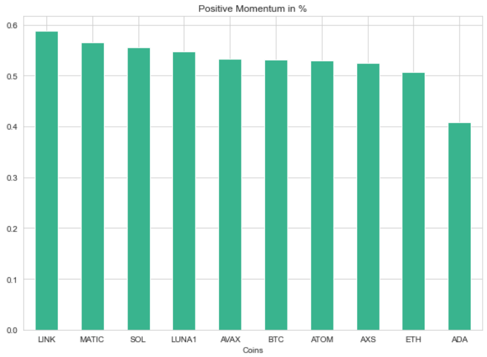
Investments4Some is a hedge fund management enterprise based in Portugal that has recently shown an interest in *Machine Learning* algorithms as a path to enhance their business. However, the initial employment of this solution was fruitless due to the complexity of the task. For that reason, *Datalin* was invited to develop a daily forecasting solution for cryptocurrencies that will allow the company to be one step ahead of the market pricing trends and, consequentially, witness an enhancement in their investments.

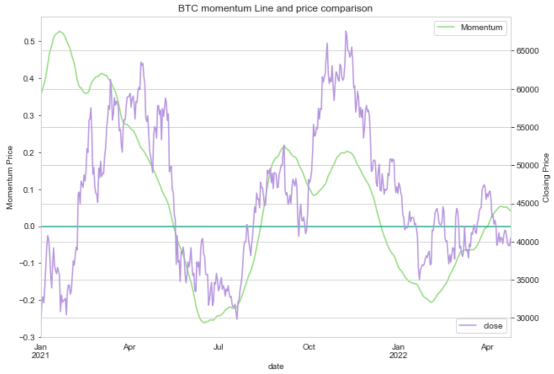
Data Understanding

Daily data on 10 distinct cryptocurrencies was provided by the company to develop the desired predictive model, namely the lowest and highest price, the opening and closing price, the adjusted closing price, and the volume. It is relevant to mention that the prices in the dataset are listed in American Dollars (*USD*). The data corresponds to the timeframe from April 26th, 2017, to April 25th, 2022.

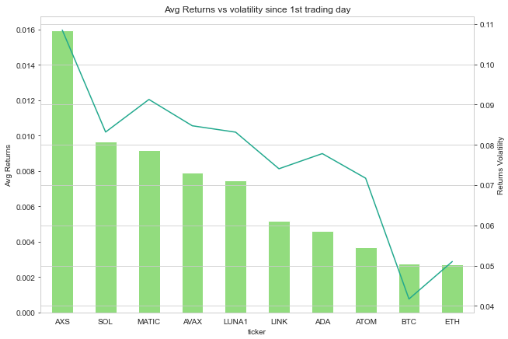
Exploratory Data Analysis

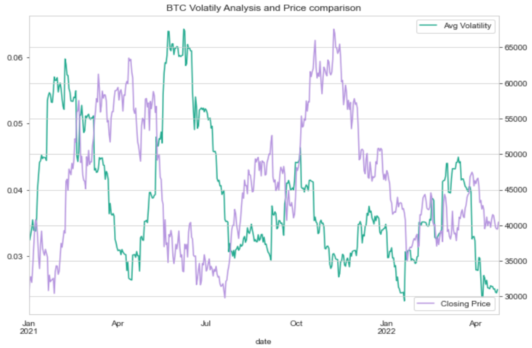
To better understand the data we had in hands, a *Ticker Signature* was created. The mean of the *Moving Average (MA)* for each coin was considered, namely by diving a low range MA by a high one. This is a great indicator since if the low range MA is higher than the high range represents that the price of the coin is, in average, higher than usual. With this information, we created the Positive Momentum in percentage, to compare results between tickers, reaching to the conclusion that *LINK* shows the highest results. Additionally, the Comparison between the *BTC Momentum Line* for all the available years, and the *Closing Price* was draw.



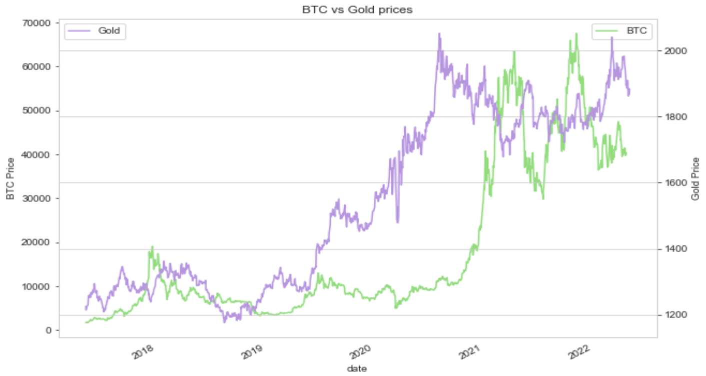


The closing prices were also studied, by comparing for each coin the average and the standard deviation of the returns, the latter being an indicator of the volatility of prices. Similarly to what was shown above, the *Volatility* was graphically compared to the Bitcoin Price, and we can see that the volatility "predicts" the degree of change of the Price.





To add more diversity of comparing power into our project, we performed visualization analysis with oil, gold and S&P Index prices. In fact, in 2021, the Bitcoin prices increased exponentially surpassing the gold ones, and since then the difference between them has been somewhat stable, with cyclical patterns of: when the price of one them increases, the other one decreases, showing how negatively correlated they are. This is aligned with the well-known S&P versus Gold prices trend, that assume the same graphical symmetry.



Data Preparation

In the initial preparation stage, data cleaning processes were applied to guarantee the greatest data quality possible for the construction of the predictive model. When inspecting for missing values it was noticed that each variable had *6442* empty cells, possibly due to the differences in the launching dates of the currencies. These observations were removed from the dataset. Outliers and incoherencies were not explored since the data available showcases a time series scenario. In other words, it was considered that all instances of the variables were legitimate and represent the true reality of cryptocurrency trading, where prices can display a large variance without it indicating errors or illogical values.

The variable creation process was focused on gathering financial indicators that could translate existing pricing patterns. The markers chosen to be included fall under three different categories of indicators, namely *Overlap Study, Momentum* and *Volatility,* that were selected to maximize the diversity of data available for model training. It is essential to mention that some of the indicators chosen were calculated for *lags* of 2, 7 and 20 days so to get a more complete overlook of their evolution.

Having in mind the first category, *the primary purpose of an overlap study indicator is to objectively identify the direction of a trend by smoothing out the volatile nature of the price action.* Under this group the indicators chosen were the *Exponential Moving Average (EMA)* and the *Simple Moving Average (SMA)*. Regarding the second category*, momentum indicators are useful for determining how fast the price of the underlying security changes. These indicators plot the rate of price change rather than the price change itself.* For example, the *Relative Strength Index* (*RSI), Rate of Change (ROC)* and *Momentum (MOM)* were selected as part of this group. Finally, Volatility indicators *attempt to measure the volatility of a security's price action.* [1]This means that as volatility increases, so do the chances of willing or losing money. To represent this type of information, indicators such as *Average True Range (ATR)* and *Standard Deviation (STD)* were selected.

[1] Jing-Zhi Huang, William Huang, Jun Ni, *Predicting bitcoin returns using high-dimensional technical indicators*, The Journal of Finance and Data Science, Volume 5, Issue 3, 2019.

Modelling

Since we are dealing with data with a strong *Time* influence, the initial goal was to split the dataset into train and test so that the most recent observations would be used in the test set. On top of that, there is data regarding multiple types of digital coins, so the split could not be done using the indexes, but rather the dates, because the same date will correspond to many observations. Nonetheless, because the ultimate goal was to predict the behavior of prices for the upcoming week, it did not make sense to not use the most recent observations to train our model, we could be losing crucial insights that would most certainly negatively impact the predictions.

On the other hand, we wanted to apply a time-based cross validation technique, to do that so we needed non-missing data for all the 10 available coins, considering the lags of 20, that were related to the same period, in order to define a training set. Therefore, the train set corresponds to data from 2021, and then it was possible to define the target variable as the *close* price for the following day. However, to avoid issues with volatility and fluctuation of prices, a new target was created that accounts for the ratio change of todays and tomorrows' price, having today as baseline. Finally, *Min Max* scaling was applied to the train set.

To better explain why no test set was defined, it is important to briefly discuss what the *Time-Based Cross Validation*[2] approach is. With a sliding window training approach, you define a timestamp for training and a significantly shorter one to test, using several splitting points in time to create more robust models. In this case, we have defined training for every 200 days and testing the subsequent day, therefore validation is applied to every single day for as long as there is data available, except for the first 200 days, so starting at July 20th, 2021. A total of 277 cross-validations were applied, meaning those were the number of models developed for each one of the 10 tickers.

To perform *Feature Selection*, the correlations between the target and the independent variables were assessed, showing that *ATR* and *STD* with 2 lags are the most correlated features. For the correlations among features, a threshold of absolute value of 0.85 was defined to avoid multicollinearity issues. Additionally, a RFE and Lasso selection were computed, and their results were intertwined, resulting in 10 variables for the final subset: *ATR_2, RSI_2, ATR_20, Avg_Price_Disparity, AROON_down_2, AROON_UP_2, ROC_2, OBV, RSI_7 & ROC_20*.

For this implementation, we decided to use *Recurrent Neural Networks*[3], that are ideal to predict sequential data using backpropagation. They use memory blocks which consist of one cell and several gates, the latter allowing the network to keep the inputs' influence for more sequences, thus for more time. The *Long Short-Term Memory (LSTM)* uses gates to learn long-term dependencies, namely what are the sequences that matter keeping or can be forgotten. This model is particularly relevant because it deals with a major drawback of these type of models: gradient vanishing or short-term memory. The second model applied was the *Gated Recurrent Unit (GRU)* has two gates, one for resetting and the other for updating, which transfers information.

[2] Information consulted on *Towards Data Science*, on the article *Time Based Cross Validation*

[3] *Jaquart P., Dann D., Weinhardt C, Short-term bitcoin market prediction via machine learning, The Journal of Finance and Data Science 7 45-66, March 2021*

To implement these models to the specific context of cryptocurrency coins, we have developed functions that filter the training set per categories, when there are many observations within the same timestamp, so that the model is trained one time for each type of coin, storing all the relevant assessment information in dictionaries where the keys are the categories, also known as the tickers or coins.

Evaluation

For the assessment of the forecasting models built, two approaches were followed. First, values for both *Root-Mean Squared Error* and *Mean Absolute Error* were analyzed. The lower the values the better for these indicators since both measure the differences between predictions and real value, in the scale of the target.

Secondly, the nature of the predictions was categorized based on the direction of the result. Since the target applied here is the percentage of change of the coin price, this value can either be positive or negative, representing respectively an expected increase or decrease on the price, in comparison with the day before. For example, if the prediction was a positive percentage and the value did indeed increase between those days, then the model was accurate at predicting the behavior of the market. This was considered a classification problem, so the f1-score, precision and recall metrics were evaluated at this stage. An error analysis along time was plotted, and the improving of the forecasts was clear with time, leading to some degree of stabilization in the optimal models constructed in the more recent months, overlooking the exogenous shocks registered in specific moments.

The approach was to start with a simpler model as benchmark, namely a *Linear Regression*, and then gradually increase complexity whilst evaluating if the change was indeed justifiable. All were object of fine-tuning of the respective parameters, through *Grid Search* if applicable, to reach the best optimization possible. Five other models were deployed and evaluated: *Random Forest, Sklearn Neural Network, Light Gradient Boosting, LSTM & GRU*. The Multi-Layer perceptron stood out for its poor performance in this situation, being disregarded completely. As for the remaining, the metric values achieved for the *RMSE* and *MAE* were quite similar and reasonably acceptable, not ever surpassing 10% for any coin.

When it comes to the movement analysis, it was clear that the most damaging target labels would be the false positives, meaning the model predicted that the coin value would increase, but it went down, since it can lead to higher losses. Therefore, we focused our analysis in getting stable overall f1-scores and higher precision, so to guarantee the model was expertly avoiding false positives.

By joining these two approaches, the final model chosen was the LSTM Network *(epoch=1, batch_size=50, units = 32, optimizer=RMSprop, activation= relu, lr=0.1)* because of the minimum values for the *RMSE* and *MAE* registered, both for the percentage target as well as the monetary one, and for the great balance it showed in the classification analysis, with precision, recall and *f1-score* around 0.6. To note that the activation was set to *relu* so to overcome the vanishing gradient problem explained.

Deployment

The *Long Short-Term Memory Network* was the best model encountered for the prediction of the next day closing price of coins. However, this model can be considered quite computationally expensive, which explains our approach of constructing it with just one epoch, even though it could be further optimized. Therefore, if the technical requirements for its correct usage are not met, it is not recommended to be deployed as a daily analysis tool.

In that situation, our suggestion would be to deploy the *Light Gradient Boosting* model constructed since it is not only less complex, with very similar evaluation metrics to the *LSTM*, but can also provide other relevant features such as the ability of retrieving the most important features to explain the price for each specific coin.

The predictions for May 10th delivered were created with the use of LSTM and with the new dataset provided, with the most recent information dated of May 8th. Because of the lack of data regarding the day before, a new function was created where the predictions are made based on two days before the targeted day, predicting firstly the previous day and then the targeted day. The forecasted values can be consulted in the table below.

Predictions	May 9th	May 10th
ADA	0.7486	0.7465
ATOM	15.8187	15.8337
AVAX	51.9529	52.2031
AXS	30.6326	30.8327
BTC	34289.0230	34339.0550
ETH	2540.0745	2539.6770
LINK	10.0603	10.0270
LUNA1	65.1772	64.7064
MATIC	0.9915	0.9874
SOL	75.7911	76.1362

At this point, the function created runs 10 models, one for each specific coin and so, if necessary, it is extremely adaptable to the inclusion of more coins. New indicators or other relevant features can also be included, cautiously so to not fall into the curse of dimensionality. The addition of new data, as new prices are released, will work smoothly since the model will just add those new days instead of undergoing a complete new training. This way, it is possible to validate the model step by step, meaning, the prediction of yesterday can be validated with the closing price of today, when released, to get a sense of the immediate performance of the model, and the necessity for updates if applicable.

These predictions can be enhanced through the addition of new types of information. The crypto market is extremely influenceable to external behaviors, and it is shown that *there is a significant correlation between Google trends, tweet sentiment, and tweet volume* [4] and the price oscillations registered. By understanding how impactful these features would be to the predictions, they should be considered to this forecasting. Moreover, getting data with higher granularity could also increase the performance, capturing specifical behaviors within a day or even an hour.

[4] Jay, P., Kalariya, V., Parmar, P., Tanwar, S., Kumar, N., & Alazab, *M. Stochastic Neural Networks for Cryptocurrency Price Prediction*, 2020

Annexes

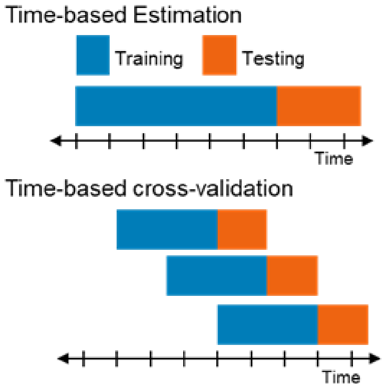
Data Understanding – Cryptocurrency Coins

Currency	*Alias*	*Launch*
Cardano	*ADA*	2017
Cosmos	*ATOM*	2016
Avalanche	*AVAX*	2020
Axie Infinity	*AXS*	2020
Bitcoin	*BTC*	2009
Ethereum	*ETH*	2015
Chainlink	*LINK*	2017
Terra	*LUNA1*	2019
Polygon	*MATIC*	2019
Solana	*SOL*	2017

Data Preparation – Feature Engineering

Variable	***Description***
EMA	Exponential Moving Average
ATR	Average True Range
RSI	Relative Strength Index
MOM	Momentum
STD	Standard Deviation of the daily percentual change
Stoch	Stochastic Oscillator; No lags were specified
KAMA	Kaufman's Adaptive Moving Average
OBV	On-balance Volume; No lags were specified
ROC	Rate of Change or Percentage of Change
AROON	Aroon Oscillator based on the Aroon up and Aroon down lines for up and downtrend, respectively
avg_price	Compares the currency's current price and its average price
SMA	Simple Moving Average
Gold_STD	Standard Deviation of the price of gold
Gold_STD_adj	Adjusted Standard Deviation of the price of gold
Gold_AvgP_adj	Average adjusted gold price

Modelling – Imagine for reference on Time-Series Cross-Validation



Evaluation – Full Metric Assessment

		ADA	*ATOM*	*AVAX*	*AXS*	*BTC*	*ETH*	*LINK*	*LUNA1*	*MATIC*	*SOL*
LR	Test MAE	0.06418	1.568523	3.909681	4.343531	1235.123	109.5622	0.982618	3.296486	0.082326	6.329235
	Test RMSE	0.091008	2.267622	5.284094	6.254887	1608.02	141.7868	1.323058	4.440709	0.110094	8.61671
	Test % MAE	0.040788	0.053784	0.057935	0.060442	0.026636	0.032984	0.043941	0.063119	0.052397	0.049662
	Test % RMSE	0.053684	0.075233	0.078864	0.096743	0.034841	0.042659	0.05594	0.084993	0.071376	0.064836
LGB	Test MAE	0.058534	1.462643	3.754233	3.812558	1189.473	105.4085	0.94503	3.14819	0.075513	5.901014
	Test RMSE	0.084191	2.1428	5.078159	5.667234	1594.183	137.6686	1.291296	4.319195	0.10411	8.129761
	Test % MAE	0.03676	0.050051	0.053293	0.0526	0.025521	0.031656	0.042158	0.059361	0.046334	0.045762
	Test % RMSE	0.048677	0.070846	0.071606	0.087919	0.034247	0.041245	0.054265	0.079391	0.063711	0.060662
RF	Test MAE	0.061572	1.510099	3.764419	3.851161	1275.191	108.4804	0.945377	3.166835	0.076781	6.082742
	Test RMSE	0.088193	2.194545	5.110329	5.686097	1692.698	142.3865	1.314152	4.335814	0.105068	8.34263
	Test % MAE	0.03839	0.051226	0.052737	0.052932	0.027101	0.032227	0.042092	0.059157	0.046659	0.04664
	Test % RMSE	0.050452	0.071393	0.071464	0.087975	0.03573	0.041977	0.055149	0.078338	0.063321	0.061146
MLP	Test MAE	19.53589	24.91217	76.86921	71.18243	3597668	185597.7	294.0028	48.07609	1.252418	133.2159
	Test RMSE	23.92941	37.23086	106.9198	105.2285	3809615	195611.9	350.2613	69.22188	1.600589	207.6628
	Test % MAE	12.17948	0.858436	1.122378	0.943781	74.95085	54.56808	13.08618	0.849104	0.782885	0.948194
	Test % RMSE	13.93434	1.179483	1.521511	1.586416	76.75668	56.11872	15.16959	1.095022	1.00257	1.361361
GRU	Test MAE	0.060815	1.497564	3.779897	3.692568	1285.747	106.4551	0.926587	3.048098	0.071904	5.96861
	Test RMSE	0.086313	2.260513	5.238473	5.579948	1724.83	140.5408	1.246365	4.142524	0.098845	8.149208
	Test % MAE	0.038647	0.051311	0.053538	0.050184	0.02748	0.031939	0.041249	0.05791	0.043869	0.047938
	Test % RMSE	0.052145	0.073848	0.071416	0.088743	0.035879	0.041707	0.053415	0.076734	0.059236	0.064617
LSTM	Test MAE	0.05537	1.470471	3.765561	3.77889	1134.254	99.51852	0.894096	2.952665	0.067245	5.831681
	Test RMSE	0.084026	2.206829	5.232949	5.46188	1533.874	128.7724	1.219065	4.015627	0.093059	7.960158
	Test % MAE	0.034562	0.050131	0.051532	0.051227	0.024127	0.029654	0.039971	0.055815	0.040543	0.045392
	Test % RMSE	0.048689	0.071455	0.068693	0.084695	0.031771	0.038205	0.051871	0.072358	0.054525	0.060054

Evaluation – Difference in RMSE for dollars and percentage predictions

AVERAGE RMSE					
%	***GRU***	***LGB***	***LR***	***LSTM***	***RF***
ADA	0.052	0.049	0.054	0.049	0.05
ATOM	0.074	0.071	0.075	0.071	0.071
AVAX	0.071	0.072	0.079	0.069	0.071
AXS	0.089	0.088	0.097	0.085	0.088
BTC	0.036	0.034	0.035	0.032	0.036
ETH	0.042	0.041	0.043	0.038	0.042
LINK	0.053	0.054	0.056	0.052	0.055
LUNA1	0.077	0.079	0.085	0.072	0.078
MATIC	0.059	0.064	0.071	0.055	0.063
SOL	0.065	0.061	0.065	0.06	0.061
AVERAGE RMSE					
$	***GRU***	***LGB***	***LR***	***LSTM***	***RF***
ADA	0.08631	0.08419	0.09101	0.08403	0.08819
ATOM	2.26051	2.1428	2.26762	2.20683	2.19455
AVAX	5.23847	5.07816	5.28409	5.23295	5.11033
AXS	5.57995	5.66723	6.25489	5.46188	5.6861
BTC	1724.83	1594.183	1608.02	1533.874	1692.698
ETH	140.5408	137.6686	141.7868	128.7724	142.3865
LINK	1.24636	1.2913	1.32306	1.21907	1.31415
LUNA1	4.14252	4.31919	4.44071	4.01563	4.33581
MATIC	0.09884	0.10411	0.11009	0.09306	0.10507
SOL	8.14921	8.12976	8.61671	7.96016	8.34263