Cryptocurrencies Forecasting

Master Degree Program in Data Science and Advanced Analytics

Business Cases with Data Science

Github repository:

https://github.com/andremforte/BC4_GroupV

Group V:

Anis Tmar (m20211157)

André Forte (m20210590)

Opeyemi Mary Akande (m20211320)

Rafael Nunes (m20210832)

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

Index

1. Business Understanding

Cryptocurrencies, like many other financial assets, are extremely volatile in value and are only worth what someone is willing to pay for them. They started to become popular due to their lack of regulations compared to other financial assets.

Investments4Some is a Portuguese hedge fund management organization. They assess the quality of their portfolios using classic statistical approaches and financial indicators. The company is well aware of the common investor's lack of expertise, as well as the potential of Machine Learning approaches to predict market trends and boost predicted returns on investments.

The company knows that, using predictive models, they could anticipate the daily worth of different cryptocurrencies (e.g., predicting trends). For that reason, our goal is to create a forecasting model to predict as accurately as possible future stock prices (in this case, future closing prices). To do this, the company provided us a dataset containing the daily prices of 10 different cryptocurrencies.

2. Data Understanding

We were provided with initial six csv files (close, adj_close, open, high, low, volume), each one of them containing s records of cryptocurrency daily data from exchange platform for the period between April 24th, 2017, and April 25th, 2022. Later on, the dataset was updated to data until May 8th, 2022.

The values of each file were converted in USD, meaning that 1 cryptocurrency corresponds to each specific value per row in different days, depending on open, high, low, close, and adj_close prices.

The different cryptocurrencies presented in each file are Cardano (ADA-USD), Cosmos (ATOM_USD), Avalanche (AVAX-USD), Axis Infinity (AXS-USD), Bitcoin (BTC-USD), Ethereum (ETH-USD), Chainlink (LINK-USD), Terra (LUNA1-USD), Polygon (MATIC-USD) and Solana (SOL-USD).

As mentioned above, the csv files contained data related to the lowest, highest, opening and closing prices. We can also find the adjustment of the closing prices, after all the applicable split and dividend distributions. Besides, we have the amount of the cryptocurrency that changed hands over the course of each day, represented in the file “volume”.

First, we built some visualizations to understand the distribution of the target (in this case, “closing price”) of each cryptocurrency over time. This could be important, since we want to predict its future value, so it could help us to establish some criteria during the next steps. Then, we split the six csv files into 10 different datasets, separating the data by cryptocurrency.

During further familiarization with the data provided, here are some observations collected:

- Except Bitcoin, there are some dates at the upper section of the other cryptocurrencies datasets with null values. This could mean the business only traded Bitcoin and not others during such period.
- There were no duplicated entries.
- The target distribution for all the cryptocurrencies appears to be similar across the years.
- Bitcoin records the most valued cryptocurrency in terms of prices.
- There were no entries with volume below or equal to zero (meaning that there was no lack of liquidity in the market).

3. Data Preparation

We started to prepare the dataset for modeling by deleting the missing entries due to lack of meaning in the analysis. Since we just had six variables, including the target ("close"), it was important to create new features that can be useful for predictions. So, regarding feature engineering, we created some variables based on the original ones and we included external data that might contribute to improve the performance of the models.

The trading indicators are Simple Moving Average (SMA), Exponential Moving Average (EMA), Stochastic Oscillator, Spread, Volatility, Close Off High. These indicators were obtained through calculations involving the original variables. The formulas are shown in Annex 1.

We also imported from yahoo Finance the daily exchange rate of euro to dollar to see if the daily rate of dollar as any influence on the daily trading of cryptocurrencies. An issue was encountered while importing the exchange rate from 'yfinance'. It was observed that there were no weekend data for exchange rate, so we decided to fill in the Saturday and Sunday entries with the exchange rate of the preceding Friday. Finally, and since Bitcoin, being the dominant cryptocurrency in the market, probably had effect on the trading of the other cryptocurrencies, we decided to include the closing price of Bitcoin as additional variable to predict the other ones.

After feature engineering, we normalized each dataset using MinMaxScaler and checked the Pearson's correlation between the variables. As expected, we noticed many variables that were highly correlated, so we decided to drop variables that registered values above 0.85 or below -0.85. With that, we were ensuring that we were selecting the variables that were useful for modeling, avoiding redundancy. The final selected variables for each cryptocurrency can be seen in Annex 2.

4. Modeling

In the Modeling phase, we applied two Machine Learning algorithms: LSTM and XGBoostRegressor.

LSTM (Long Short Memory networks) is a special kind of Recurrent Neural Networks (RNN). RNN are networks with loops in them, allowing information to persist. They can be seen as "multiple copies of the same network, each passing a message to a successor" (blog, 2015). LSTM are designed to avoid the long-term dependency problem. Each layer included in this algorithm takes information from itself, but also from the previous layers, working using different gates and diverse activation functions where the data is passing. (Sagar, n.d.) Extreme Gradient Boosting (XGBoost) is an ensemble machine learning algorithm that can be used directly in regression modelling. (Brownlee, 2021)

To predict each cryptocurrency, we used sliding windows to train and test the models. Since we are creating time series models, using this technique will allow the algorithm to train and test with different sets chronologically (e.g., group of 10 past days predicting the 11th day, then the 9 past days, plus the previous prediction, to forecast the 12th day and so on). In this case, we performed analyses using different length of intervals: 7 and 20 days. We defined 80% to train and 20% to test the models, based on the past and recent data.

We also analysed the performance of the model using different timeframes: using the entire dataset, data from January 1st, 2021, and another one with prices since January 1st, 2022. The choice of this sub datasets was based on the distributions of the closing prices of each cryptocurrency across time made in Data Understanding (e.g., until January 2021, most of the cryptocurrencies registered a constant value, and since we were using far past data to train the model, this may lead us to bad performances of the algorithm).

Regarding LSTM, we created 3 LSTM layers and 1 dense layer, and then we compiled them together. When we were fitting the model, and since we were dealing with time series, we used "shuffle= False", to make sure that the model was not making predictions of past data using future data.

For XGBoost Regressor, we used 1000 estimators, meaning that we were using 1000 of rounds, and eta = 0.2, corresponding to the shrinkage that we were doing at every step we were making.

5. Evaluation

To select the final model of each cryptocurrency, we analysed four metrics: RSME, MAE, MSE and R^2. We decided to analyse these evaluation metrics because we wanted to understand how well the models perform, trying to minimize the error by understanding the difference between predicted and actual values.

We used Mean Squared Error (MSE) because it's useful to understand if the predicted values match the expected values. For example, if the MSE is 0, this represents that the predicted values are the same as the expected values.

We used Mean Absolute Error (MAE) to understand the magnitude of the errors. The absolute difference means that if the result of the expected values minus the predicted values has a negative sign, it is ignored.

The RMSE is similar to MSE, but the root of the value is considered to determine the accuracy of the model, i.e., to assess if the error is significant or not. The lower the value, the better the result. Regarding R-squared, this metric can measure the amount of variance in the predictions explained by the dependent variables' variance. It's important to note that R-squared "does not indicate whether or not the model is capable of making accurate future predictions". A high value represents a strong association between actual and predicted prices.

For that reason, since our goal is to minimize the residuals, meaning that we want to reduce the difference between the predicted values by the model and the actual variables, our focus was on MAE and RMSE. (Analyticsindiamag, 2021)

Based on the results and taking into consideration the overall quality of the created models, we selected the XGBoost Regressor, using a sliding window of 7 days, to perform the predictions for the next two days.

6. Predictions and Business Implications

As mentioned above, the predictions were computed using a new file containing the updated prices of each cryptocurrency until May 8th, 2022. Before predicting the values, we needed to perform the same data preparation process to make sure we can provide the right features to the model.

Since our models are only able to predict the next day, we decided to follow this highlighted reasoning: first, we decided to predict the closing price of Monday (May 9th, 2022), using our model. Then, since we needed inputs of the other variables and we couldn't access it, we fill in the empty values with the prices of the previous day, Sunday (May 8th, 2022). Having this, we predicted the closing price for the next day, May 10th, 2022, using our model.

The predictions for the next two days (May 8th and 10th, 2022) of each cryptocurrency can be seen in Table 1.

		Date	
		09/05/2022	10/05/2022
Cryptocurrency	BTC	33116.26255	36581.19015
	LUNA1	69.62269683	78.04264421
	LINK	15.28577545	15.56168104
	ETH	2677.089072	2749.476956
	SOL	115.2179072	115.6561218
	ADA	1.008748097	0.963340599
	MATIC	1.056502108	1.208942845
	AXS	61.57645498	63.2018159
	AVAX	57.73125212	62.0672619
	ATOM	17.19707254	19.65041371

Table 1. Final Predictions

Forecasting is the practice of using historical and real-time data to estimate future demand. Due to the potential to use and learn from prior mistakes or failures, knowledge of historical patterns might bring insights into trading and so minimize expenses and enhance profit.

The fundamental disadvantage of projections is that they are usually always incorrect, resulting in excess or shortfall of inventories. Also, sufficient and close-to-accurate projection involves time, resources, and money to hire a technical team to deploy superior technologies.

However, the significant volatility of cryptocurrency, particularly during times of crisis, such as the coronavirus, makes forecasting very essential. Although with a fantastic procedure in place and forecasting professionals on the payroll, it is nearly impossible to forecast the future with precision, however understanding what variables drive cryptocurrency demand might potentially help with prediction. Therefore, forecasting may be costly in the short term, but when done perfectly, it may be advantageous in the long term. (planet together, 2020)

For future research, one possible approach could be adding more external data to the dataset in order to get more valuable information that contributes to predictions. As an example, sentiment data from Twitter can be important because it allows us to have access to the general atmosphere of the investors/individuals. With that, we can try to understand the trending expectations of the cryptocurrencies and what might be the possible future evolutions of the market.

References

Analyticsindiamag. (2021, November 1). *A Guide to Different Evaluation Metrics for Time Series Forecasting Models*. Retrieved from https://analyticsindiamag.com/a-guide-to-different-evaluation-metrics-for-time-series-forecasting-models/

blog, c. (2015, August 15). *Understanding LSTM Networks*. Retrieved from https://colah.github.io/posts/2015-08-Understanding-LSTMs/

Brownlee, J. (2021, March 12). *Machine Learning Mastery*. Retrieved from XGBoost for Regression: https://machinelearningmastery.com/xgboost-for-regression/

planet together. (2020, May 13). Retrieved from Advantages and Disadvantages of Forecasting: What's The Best Option for My Supply Chain?: https://www.planettogether.com/blog/advantages-and-disadvantages-of-forecasting-whats-the-best-option-for-my-supply-chain

Sagar, A. (n.d.). *Towards data Science*. Retrieved from Cryptocurrency Price Prediction Using Deep Learning: https://towardsdatascience.com/cryptocurrency-price-prediction-using-deep-learning-70cfca50dd3a

Annexes

Representation	Formula	Meaning
SMA_7	closeprice.rolling(window=7).mean()	Identify the direction of all price trend, without the interference of shorter-term price spikes
EMA_7	closeprice.ewm(span=7, adjust=False).mean()	Identify the direction of a current price trend, without the interference of shorter-term price spikes
14_high	highestprice.rolling(14).max()	Compares highest price traded to a range of its prices over 14 previous trading session
14_low	lowestprice.rolling(14).min()	Compares lowest price traded during the same 14 days period
%K	(closeprice - 14_low)*100/(14-high - 14-low)	The current value of the stochastic indicator sometimes called fast stochastic indicator
%D	%K.rolling(3).mean()	The slow stochastic indicator usually 3-period moving average of %K.
spread	highestprice - lowestprice for the day	Difference in a trading position – the gap between a short position
volatility	spread / openprice for the day	The standard deviation or variance between returns
close_off_high	closeprice- highestprice for the day	Difference in a trading position – the gap between a short position
EUR/USD_close	-	Imported from yfinance
BTC_closeprice	-	Closeprice at the end of the day for bitcoin

Annex 1. New Features

Feature Selection				
ADA	BTC	SOL	LUNA1	MATIC
Close	Close	Close	Close	Close
volume	Volume	Volume	-	Volume
% D	% D	% D	% D	% D
Spread	Spread	-	-	-
Volatility	Volatility	Volatility	Volatility	Volatility
Close_off_high	Close_off_high	Close_off_high	Close_off_high	Close_off_high
EUR/USD_close	EUR/USD_close	EUR/USD_close	EUR/USD_close	EUR/USD_close
-	-	BTC_closeprice	BTC_closeprice	BTC_closeprice

Feature Selection				
ETH	LINK	AVAX	AXS	ATOM
Close	Close	Close	Close	Close
Volume	Volume	volume	volume	Volume
% D	% D	% D	% D	% D
Spread	Spread	Spread	Spread	Spread
Volatility	Volatility	Volatility	Volatility	Volatility
Close_off_high	Close_off_high	Close_off_high	Close_off_high	Close_off_high
EUR/USD_close	EUR/USD_close	EUR/USD_close	EUR/USD_close	EUR/USD_close
-	-	BTC_closeprice	BTC_closeprice	-

Annex 2. Selected Features for Modeling