XRP Price Prediction

Author: Francesco Piccoli

Abstract

Is it possible to correlate XRP's price to some of the features extracted from Ripple's transaction history? Volume is often used in regression models for price forecasting, but it's difficult to effectively estimate how much of that volume is actually real. One big player could easily move money back and forth from two different accounts, generating a high amount of what is referred to as "fake volume", which might not be significant to forecast price movements.

Is there any other data source we can look at to predict XRP's price, such as Google Trends data or the number of Reddit subscribers in Ripple's subreddit? Our hypothesis is that these features are a signal of people's engagement in the XRP Ledger, and may contain willingness-to-buy information.

For this reason we collected data from several different public sources to try to predict price movements, and built different classification models to test our hypothesis. A test set accuracy of 60.74% in predicting whether XRP will appreciate or depreciate with respect to USD was reached. The prediction is made on XRP's price on day T+1, using information from day T. An investment strategy was then built using our best model's prediction, and it yielded a final portfolio value higher than the chosen benchmark.

Data collection

We collected both hourly and daily data of the following variables:

Hourly

- **Google trends** data of the keywords XRP, Binance, and Coinbase. This data may provide information on how much interest is generated around XRP. Binance and Coinbase are popular exchanges where it's possible to buy and sell XRPs. Obtained using the Pytrends API from General Mills.
- **Volume** of XRP transacted, a typical variable used in price forecasting. Obtained using Cryptocompare's API.
- **Exchange rate** between XRP and US dollar, which is our dependent variable. This data is an agglomerate of exchange rates that gives the best price estimate across 70 different exchanges. Obtained using Cryptocompare's API.

- **Momentum** factor, which is a dummy variable indicating whether the XRP price in the previous hour went up or down.

Daily

- **Google trends** data of the previously mentioned keywords., obtained by computing the daily mean from the hourly data.
- **Reddit** data, obtained using the Pushshift API. In particular, similar to what has been done in [1], we use the number of new daily subscribers in Ripple and Coinbase subreddits, which are active forums of discussion.
- **Wikipedia** views of Ripple's and Coinbase's pages. Obtained using the Mwviews API.
- **Volume, Exchange rate, and Momentum factor**, obtained by aggregating the hourly data coming from Cryptocompare's API
- **Coin Metrics** economic indicators, downloaded from Coin Metrics' website. Among them:
    - **Active addresses,** which is the number of unique addresses that were active in the network, by either being the receiver of the originator of a transaction.
    - **Mean Fee** per transaction, which increases when the network experiences an unusual load.
    - **ROI** at 30 days, the return on the investment in XRP assuming a purchase was made 30 days before. A positive number may be a signal that short-term investment positions will be unloaded.
    - The number of **transactions** made that day

All the data was collected from January 1st, 2017 to December 31st, 2019, and therefore contains the period where the price of cryptocurrencies skyrocketed, followed by the big crash at the beginning of 2018. This allows us to analyze both a very volatile period and a more stable one, following the crash.

Image 1: XRP historical price and volume
Source: https://cointelegraph.com/xrp-price-index

Data cleaning

Before proceeding with further analysis, we transform our dependent variable (exchange rate between XRP and USD) into a binary variable, which takes the value of 1 if the value of XRP has increased with respect to the dollars in the previous hour/day and 0 if it has decreased. This allows us to predict price movements instead of the actual price value, which will be a non-stationary dependent variable.

Furthermore, we proceed with a feature scaling of our dataset to normalize the range of values of the different columns, since some classifiers might otherwise not work properly. In particular, a mean normalization was used. In addition to that, we handle missing values by substituting them with the mean of each feature.

Last, since we want to perform a timeseries forecasting, we shift the independent variable column up one row. This way we will use data we are in possession of this hour/today to predict the price in the next hour/tomorrow.

Data correlation

For hourly data, we observe the following Pearson correlation:

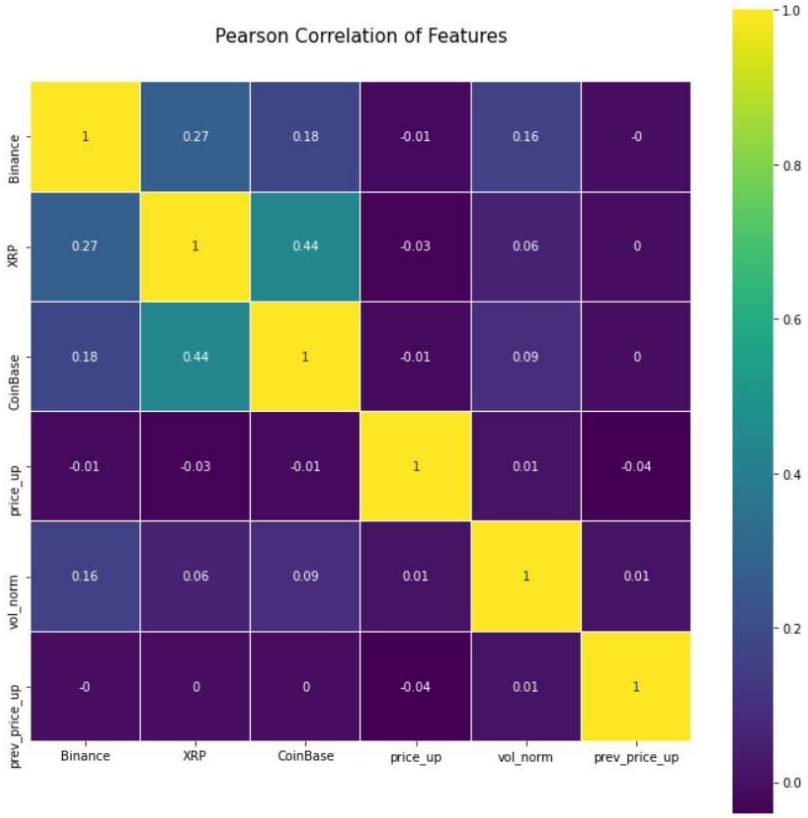

Image 2: Features correlation - Hourly data

It’s worth noticing how Google queries for the keywords Binance, XRP, and Coinbase have some correlation between each other, especially Coinbase and XRP.

For daily data, we observe the following Pearson correlation:

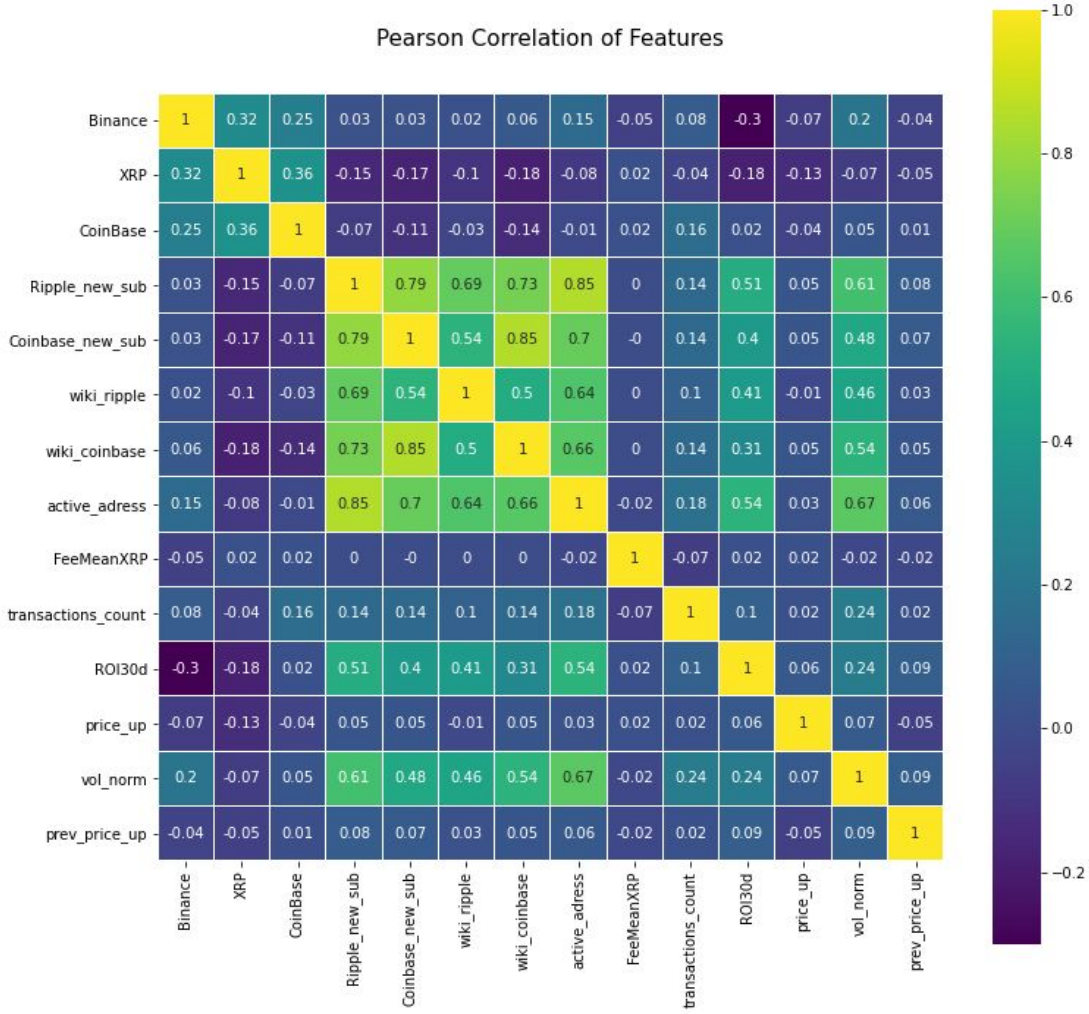

Image 3: Features correlation - Daily data

This graph is much more interesting than the previous one and shows several highly correlated ‘boxes’. In particular, we notice:

- A medium positive correlation between Google trends data
- A high positive correlation between the number of daily new subscribers to Ripple and Coinbase subreddits, Wikipedia page views, and the number of active addresses on the XRP Ledger.
- A medium positive correlation between the ROI at 30 days, the volume in XRPs exchanged on the Ledger, and the previously mentioned highly correlated block.

Models and results

For both hourly and daily data we run different classification models. We used the following split for the data:

- Training set: first 70% of the data in chronological order (2017-01-01/2019-02-06)
- Validation set: subsequent 15% of the data (2019-02-07/2019-07-20)
- Test set: last 15% of the data (2019-07-21/2019-12-30)

Our data is well balanced on all the different sets, with around 50% observations having the price going up and 50% with the price going down. Our baseline model would be predicting the most occurring class on the training set (price going down 51% of the times) on the validation and test set

The results for each model are summarized in the following tables. We performed cross-validation on the validation set and reported the best results.

	Validation set accuracy
Baseline model	51.65%
Logistic regression	52.36%
K-Nearest Neighbors (k=74)	**53.68%**
Support Vector Machine	53.20%
Perceptron	49.57%
XGBoost (# of estimators = 14)	**53.68%**
Random Forest (# of estimators = 30)	52.81%

Table 1: Classification results - hourly data

	Validation set accuracy
Baseline model	52.44%
Logistic regression	51.83%
K-Nearest Neighbors (k=19)	**60.98%**
Support Vector Machine	55.49%
Perceptron	46.34%
XGBoost (# of estimators = 80)	54.27%

Random Forest	56.70%

Table 2: Classification results - daily data

By checking the variable importance for the Random Forest we notice that some variables (wiki_coinbase, Binance, and FeeMeanXRP) are not significant predictors. Therefore we drop them and recompute the validation accuracy for our best-performing model (KNN), noticing an increase in it. At this point, we also check for the accuracy on the test set.

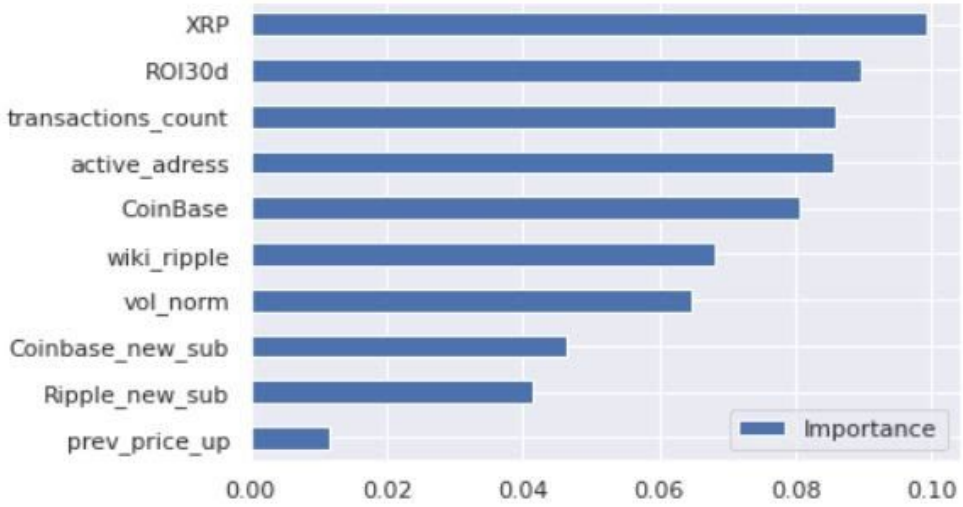

Image 4: feature importance

	Validation set accuracy	Test set accuracy
KNN (k=33)	61.58%	60.74%

Table 3: Best model

	Prediction: price down	Prediction: price up
Price down	75	16
Price up	48	24

Table 4: Test set confusion matrix

Investment strategy

Based on the previous results, we now use the best-performing model (KNN) on the daily data to build an investment strategy. Let's assume we have in our portfolio 200 XRP and 200 USD at the beginning of the validation (2019-02-07) and test set period (2019-07-21). Our investment strategy would be to buy XRPs with dollars if our model predicts XRP to appreciate with respect to the dollars tomorrow, and to sell XRPs for dollars if XRP is predicted to depreciate.

To assess the quality of our model we calculated the dollar value of our portfolio each day. As a benchmark, we use the value of the portfolio at the end of the two periods (validation: 2019-07-20, test: 2019-12-30) if we didn't implement the trading strategy, namely if we still had 200 XRPs and 200 dollars.

Our strategy gives the following results:

<u>Validation</u>

Initial dollar value of our portfolio: $258.27
Final dollar value of our portfolio: $267.76
Benchmark final value: $266.17
Final XRPs: 102

Image 5: Trading strategy - validation set

We notice a general increase in the portfolio value over time, with a period of high potential profits before the end of the considered time windows. XRP appreciates with respect to the dollar, and the final portfolio value is just slightly higher than the benchmark.

Test

Initial dollar value of our portfolio: $265.73
Final dollar value of our portfolio: $244.39
Benchmark final value: $238.85
Final XRPs: 117

Image 5: Trading strategy - test set

We notice a general decrease in the portfolio value over time. XRP considerably depreciates with respect to the dollar during the considered time window, and our investment strategy offsets particularly well this decrease. In fact, the final portfolio value is 2.5% higher than the benchmark, a good result given the relatively short time span.

Conclusions

Transaction volume on a particular cryptocurrency's blockchain has always been thought to be a driver of that cryptocurrency's price. Within that volume, there's often a significant percentage of what is referred to as "fake volume". For example, one big player can easily create two accounts A and B and move large quantities of money back and forth from A to B. It is easy to see how this volume is not representative of the value of one cryptocurrency. Nevertheless, it can be a driver of poor investment strategies, especially by small investors, since this information is easy to find online on several different websites. In contrast, it's hard to find reliable information on how much fake volume is contained in that volume information.

This study demonstrated that, together with more traditional economic data, social network and search engine data can be used to predict XRP's price movements. We conclude that this data can be a signal of people's interest in a particular cryptocurrency and can contain willingness-to-buy information. The factors considered in this analysis could be included in more complex trading strategies where forecasting cryptocurrencies' price is essential.

Bibliography

[1] Phillips, R. C., & Gorse, D. (2018). Cryptocurrency price drivers: Wavelet coherence analysis revisited. *PloS one*, *13*(4), e0195200. https://doi.org/10.1371/journal.pone.0195200

[2] GeneralMills. "GeneralMills/Pytrends." *GitHub*, github.com/GeneralMills/pytrends.

[3] Al-Vincent. "Al-Vincent/Reddit-Analytics." *GitHub*, github.com/al-vincent/reddit-analytics/blob/master/ExtractRedditMetricsData.py.

[4] Mark Needham "Finding Famous MPs Based on Their Wikipedia Page Views · Mark Needham." *Mark Needham*, 1 Apr. 2019, markhneedham.com/blog/2019/04/01/famous-mps-wikipedia-pageviews/.

[5] Pethani, Ajay. "CCCAGG Exchange Selection Methodology." *Medium*, CryptoCompare Research, 4 Oct. 2019, blog.cryptocompare.com/cccagg-exchange-selection-methodology-5f53a847761c.

[6] CryptoCompareLTD. "CryptoCompareLTD/Api-Guides." *GitHub*, github.com/CryptoCompareLTD/api-guides/tree/master/python.