Capstone Project Proposal: Estimating Cryptocurrency Future Price Using Machine Learning in the Latin American Markets.

Nicolas Rodriguez Celys

June 1, 2018

1 Domain Background

The interest in blockchain has received much attention in several domains since its introduction by [Nakamoto, 2008]. Blockchain is supported by the join of technologies of peer to peer, symmetric cryptography and cryptohgraphic hashing where the information of transactions (Ledgers) is distributed among computers without the intervention of a central authority. A cryptocurrency is a digital currency exchanged on the distributed ledger of blockchain; bitcoin was the first to be implemented in 2009. In the financial sector, individuals see an oportunity for personal investment in the growing economy of cryptocurrencies. The price of bitcoin has climbed to 5000%([Urquhart, 2008]) since 2009. There is a divided opinion on the academic field about the benefits or the drawbacks of cryptocurrencies. [Dwyer, 2014] argues that bitcoin exhibits a high volatility compare to stable assets such as gold or some set of fia money. [Dwyer, 2014] presents the existence of bubbles in the bitcoin market. Yet, [Cheung et al., 2013] argues bitcoin can be used for diversification while [Briere and Osterlicnk, 2015] shows that the criptocurrency has hedge capabilities such as gold and dollar. [Urquhart, 2008] concludes that bitcoin, in general, is an inefficient on the perspective of efficient market hypothesis (EMH) [Fama, 1970] however, it might has a tendency to be efficient on the weak form, meaning that investors cannot use past information to predict future returns.

In areas such as machine learning (ML), artificial inteligence, data mining, big data or business inteligence, the use of computer science has emerged in the last two decades as a prediction tool for the financial sector([Vityaev, 2005]). [Pang and Song, 2011] uses Support Vector Machine (SVC) for Forex Time series prediction. Moreover, clustering is also used to price prediction for the Forex market ([Sopranzetti and Datar, 2002], [H-J.Ahn and Cheung, 2005]). Due to its novelty, in the cryptocurrencies market, there are few studies on the application of ML for price prediction. [Peng and Albuquerque, 2018] uses a SVC to forecast the volatility of cryptocurrencies while [Georgoula et al., 2015] uses SVN to measure the sentiment analysis on twiter and its correlation with the

price of bitcoin. Finally [Indera et al., 2017] uses NARX model combined with neuronal networks to forecast bitcoin price.

Despite the volatility of the cryptocurrencies, they are becoming alternative for investing in the Emerging Markets specially in goverments where the fiscal and monetary policy is irresponsible. Latin America is characterise for its high inflation. Argentina is still struggling to control the inflation created but by the Kitchner government and Venezuela is suffering an hyperinflation due to its socialist totalitarian regime. For this reason, I am personally encourage to research more about cryptocurrencies as alternative for investment.

2 Problem Statement

The problem to be solved is to forecast the daily closing price of bitcoin (BTC) and ethereum (ETH) for suramerican fia currencies Colombian Peso (COP), Brazilian Real (BRL), Chilean Peso (CHL) and peruvian soles (SOL) accesing historical data from local exchange markets. The problem it is a time series price forecasting where the inputs for the ML model are a set of historic time series of different dimensions.

3 Datasets and Inputs

The price information is going to be collected from the local exchanges buda and mercadobitcoin. Buda opened business in 2014 in Chile and they have gradually expanded to Colombia and Peru. Therefore, for each of the currency pair BTC-COP, BTC-CHL, BTC-SOL, ETH-COP, ETH-CHL and ETH-SOL hosted in buda, there will be a daily time series dataset of 3 years. Futhermore, Mercadobitcoin opened in 2011 in Brazil being the first exchange in Latin America for the pair BTC-BRL; hence, it is expected a timse series dataset of 6 years. Each time series will contain the following fields:

Date Day with the information.

Closing Price The price at the end of the day.

Opening Price The price at the beginning of the day.

High Price The maximum price during the day.

Low Price The minimum price during the day.

Total circulation of bitcoin Total daily number of bitcoin in circulation.

In addition, I will extract the pair BTC-USD from the site `www.bitcoincharts.com`. It provides a complete history of various Bitcoin exchanges. The pair BTC-USD will be usefull to see if there is an arbitrage strategy or to find out the correlation between the different prices.

There are other technical time series that can be extracted from the data represented here. Moving Averages, Bollinger bands and pivots are timeseries information that is being used for professional traders among the world [Menkhoff, 2010].

4 Solution Statement

The solution consist of a nonlinear autoregressive exogenous model (NARX). Using a ML, we are going to find a function H such $P_t = H(\tau, \theta, \omega)$ where:

P_t The price estimate for the date t.

τ Represent the dimension of the factual information of the price. this information includes information of previous days such as: closing price, opening price, min price, max price and volume.

θ Represent the dimension of the complementary technical information of the price. This information includes information of previous days such as: Moving averages, Moving Standard Deviations, Bolinger Bands etc.

ω Represent the market information that may enrich the analysis, for instance, the representative rate against USD.

To find the H, It is going to be used the set of ML algorithms use it for regression provided by the softwares Skelearn [Pedregosa et al., 2011] and Tensorflow[Abadi et al., 2015]:

Generalized Linear Models: Ordinary Least Squares, Ridge Regression, Lasso and Logistic Regression.

Support Vector: SVR and Kernel Ridge Regression.

Decision Trees: Decision Tree Regression.

Ensemble Methods: Gradient Boosted Regression Trees and Extreme Gradient Boosting.

Deep Learning: Multilayer Perceptron, Long Short-Term Memory.

5 Benchmark Model

The problem is a time series forecast. As benchmark we are going to use the autoregressive integrated moving average (ARIMA) Model and the model proposed in [Indera et al., 2017]. ARIMA model combines the dependency between the current observation and a number of lagged observations (AR), the dependency with the trending and the seasonality of the time series (I) and the dependency between the moving average applied to lagged obervations (MA). [Indera et al., 2017] uses a NARX model to forecast the bitcoin price with NN.

6 Evaluation Metrics

As a regression metric valuation we are going to use the **coefficient of multiple determination** R^2 and the **adjusted** R^2 or R_a^2. The R^2 is defined as:

$$R^2 = 1 - \frac{SSE}{SST} \tag{1}$$

where SSE is the residual sum of squares and SST is s measure of total variation of the target values. For R_a^2 is:

$$R_a^2 = 1 - \frac{n-1}{n-(k+1)} \cdot \frac{SSE}{SST} \tag{2}$$

where n is the number of observations and k the number of features use to predict the target variable.

7 Project Design

This is the proposed workflow for the project based on **Cross-industry standard process for data mining** (CRISP-DM):

1. **Data Retrieval**: Extract the data from the public APIs of the exchange and organize them in csv files for future use in pandas. This data will be known as the raw data.
2. **Data Preparation**: Prepare the data for the ML models. In addition of the raw data, some additional features are going to be added to the data. Some of the features are technical trade variable such as Moving averages, Moving Standard Deviations, Bolinger Bands. Then we proceed to feature selection can be used with the PAC model or some other. Moreover we scale/normailize each of the features to be included in the ML model. Finally, we build the training, validation and test datasets.
3. **Pre Modeling**: Use some machine learning algorithms for regression and check how good they perform on the validation dataset. the idea is to have an insight how good different algorithms behaves with the regression task of the model. The algorithms to be used are highlited in section 4.
4. **Training and cross validation**: Train the best ML algorithms chosen from the Pre Modeling Stage and use cross validation to tune each of the models. Select the specific parameters of each of the ML so the model do not overfits or underfits.
5. **Evaluation**: Compute the evaluation metrics R^2 and R_a^2 each of the models on the test dataset.

The researcher may iterate several times on the stages Data Preparation, Premodeling and Training and cross validation in orther to tune better the final model.

something

References

[Abadi et al., 2015] Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

[Briere and Osterlicnk, 2015] Briere, M. and Osterlicnk, K. (2015). Virtual currency, tangible return: Portfolio diversification with bitcoin. *J. Asset MAnag. 47,2348-2358.*

[Cheung et al., 2013] Cheung, A., Roca, E., and Su, J. (2013). Cypto-currency bubbles: An application of the phillips-shi-yu (2013) methodology on mt.gox bitcoin prices. appl. econom. *Appl. Econom. 47,2348-2358.*

[Dwyer, 2014] Dwyer, G. (2014). The economics of bitcoin and similar private digital currencies. *J.Financ.Stab. 17,81-89.*

[Fama, 1970] Fama, E. F. (1970). Efficient capital markets: A review of theory and emprirical work. *J. Finance. 25,383-417.*

[Georgoula et al., 2015] Georgoula, I., Pournarakis, D., Sotiropoulos, D. N., and Giaglis, G. M. (2015). Using time-series and sentiment analysis to detect the determinants of bitcoin prices.

[H-J.Ahn and Cheung, 2005] H-J.Ahn and Cheung, Y. (2005). Price clustering on the limit-order book: Evidence form the stock exchange of hong kong. *Finan. Rev 8(4),421-451.*

[Indera et al., 2017] Indera, N. I., I. M. Yassin, A. Z., and Rizman, Z. I. (2017). Non-linear autoregressive with exogenous input (narx) bitcoin price prediction model using pso-optimized parameters and moving average technical indicators. *J Fundam Appl Sci. 2017, 9(3S), 791-808 792.*

[Menkhoff, 2010] Menkhoff, L. (2010). The use of technical analysis by fund managers: International evidence. *Journal of Banking and Finance 34 (2010) 2573 2586.*

[Nakamoto, 2008] Nakamoto, S. (2008). Bitcoin: A peer-to-peer electronic cash system.

[Pang and Song, 2011] Pang, S. and Song, L. (2011). Correlation-aided support vector regression for forex time series. *Neuronal Computing and Applications, 20,1193-1203.*

[Pedregosa et al., 2011] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

[Peng and Albuquerque, 2018] Peng, Y. and Albuquerque, P. H. M. (2018). The best of two worlds: Forecasting high frequency volatility for cryptocurrencies and traditional currencies with support vector regression. *Expert Systems With Applications 97 (2018) 177–192.*

[Sopranzetti and Datar, 2002] Sopranzetti, B. and Datar, V. (2002). Price clustering in foreign exchange spot markets. *J. Financ Mark, 5(4),411-417.*

[Urquhart, 2008] Urquhart, A. (2008). The inefficiency of bitcoin. *Economics letters 148 2016 (80-82).*

[Vityaev, 2005] Vityaev, B. K. E. (2005). Data mining and knowledge discovery handbook. *Data Mining and Knowledge Discovery Handbook.*