Machine Learning for Cryptocurrency Price Prediction

Team: Karan Jaisingh (419, Shubham's Cohort), Brendan McAffery (519, Yang's Cohort), Raphael Iskra (519, Yang's Cohort)
Project Mentor TA: ?

Introduction

In 2020, Bitcoin led a massive revolution in the growth of cryptocurrency, with its market cap surpassing $1 Trillion for the first time in history. This revolution was met with institutional adoption: JP Morgan and Goldman Sachs reversed its previous stance which declared Bitcoin/crypto "is not an asset class," and Tesla became the first company in the S&P 500 to convert a portion of its balance sheet to Bitcoin. Most importantly, the events of 2020 have forced individuals across the globe to deeply investigate the question, "What is money?"

The members of this project group understand that a decentralized network of proof-of-work-based value recognition is far superior to the synthetic solutions (targeted inflation, TBill Rates, etc.) offered by centralized financial institutions. Brendan has experience trading crypto currencies with algorithms that employ technical analysis strategies. Raph, who boasts an Bitcoin portfolio with a Cost Basis of $14/BTC, is also an active crypto-trader. Karan has an interest in using quantitative/ML methods to predict market behavior, and is intrigued to investigate ML capabilities on blockchain networks. Thus, the members of this group are interested in using Machine Learning methods to predict short-term price fluctuations of Bitcoin and other crypto currencies.

Related Prior Work

A Deep-Learning Based Cryptocurrency Price Prediction Scheme:
Mohil Maheshkumar Patel, Sudeep Tanwar, Rajesh Gupta, Neeraj Kumar, A Deep Learning-based Cryptocurrency Price Prediction Scheme for Financial Institutions, Journal of Information Security and Applications, Volume 55, 2020, 102583, ISSN 2214-2126, https://doi.org/10.1016/j.jisa.2020.102583.
In this paper, a LSTM and GRU-based hybrid cryptocurrency prediction scheme is proposed, which focuses on only two cryptocurrencies, namely Litecoin and Monero. The results depict that the proposed scheme accurately predicts the prices with high accuracy, revealing that the scheme can be applicable in various cryptocurrencies price predictions.

Anticipating Cryptocurrency Prices Using Machine Learning:
Laura Alessandretti, Abeer ElBahrawy, Luca Maria Aiello, Andrea Baronchelli, "Anticipating Cryptocurrency Prices Using Machine Learning", Complexity, vol. 2018, Article ID 8983590, 16 pages, 2018. https://doi.org/10.1155/2018/8983590.
This study analyzed daily data for cryptocurrencies for the period between Nov. 2015 and Apr. 2018. We show that simple trading strategies assisted by state-of-the-art machine learning algorithms outperform standard benchmarks.

Both LSTM and GRU-based approaches have shown to be effective. Additionally, regression trees built by the XGBoost Algorithm used in the latter study showed to be effective in short term price prediction. The XGBoost Algorithm built different regression trees for each cryptocurrency, suggesting it's effective across currencies.

Formal Problem Setup (T, E, P)

T:

For each day in our time frame, its state is on day t is denoted as $s_t = [o_t, h_t, l_t, c_t, v_t, w_t]$, corresponding to Open, High, Low, Close, Trading Volume, and Weighted_Price respectively. Our task is to classify BTC as a buy/sell (whether the price will increase/decrease) at the start of each day based on the state of the coin each morning. We propose to learn a predictor $f_\theta(s_{0:t})$ to approximate s_t.

E:

We have explored both historical BTC datasets and pulling live, daily data from the Coinbase API as a source for the input information defining the state of the coin each day. These datasets have all of the defined attributes of the coinstate above, as well as others which we may wish to incorporate into our machine learning experiment in the next phase. We propose to use data from 30 days worth of BTC trading from Coinbase's live dataset, and test our model's performance over the course of another 5 or 10 days from this live dataset (test set). If this dataset is unavailable, difficult to incorporate, or found to be inappropriate for training our model, we will explore training our model on 30 days worth of BTC historical data from some time in the past, but still use live, daily datasets to test our model's predictive accuracy in classifying the coin as a buy or sell.

P:

We define the Mean Average Error (MAE) over the days of BTC trading as follows: On a given day t, if our model classifies the coin state s_t as a buy and the BTC price does in fact increase, this will be a success. if our model classifies the coin state s_t as a buy/sell and the BTC price does in fact increase/decrease (respectively), this will be a success, while the opposite is a failure. We will then calculate the MAE of different classification approaches and use this as our standard metric of performance.

Progress Report (Methods, Results):

Our initial data exploration has involved us working with a suite of both classification and regression models. The classification task revolved around predicting whether the price of Bitcoin would increase or decrease on a day-to-day basis, and was tested with baseline models like Random Forest and Support Vector Machines with limited hyperparameter optimisation. The results were reasonably standard across the historical period used when predicting future results, whether it was just a 5 day period or a 90 day period.

When attempting to extend this task into the regression domain, we did experience relatively higher error rates when using the mean absolute error metric. The algorithms tested were Linear Regression and Support Vector Machines, though again minimal hyperparameter optimisation was performed.

A summary of the current results can be seen in the table below. The plan going forward is to attempt to use models that account for sequential data, such as Recurrent Neural Networks and LSTMs, in order to generate predictions. This is crucial because the current models being tested do not account for the fact that certain samples of each input vector represent a less significant period of time, as they are more outdated than other samples - taking this into account structurally should ideally lead to a lower MAE.

Method Name	Training Accuracy/MAE	Test Accuracy/MAE
Random Forest Classifier	91.24%	87.32%
Support Vector Machine	88.91%	81.43%
Linear Regression	41.59	45.63
Support Vector Regression	34.84	41.26

Remaining Challenges, and work plan over the next ~4 weeks:

- Gaining intuition on sequential-based models in order to implement them to account for differences in time amongst the samples that are fed into the model.
- Finding a suitable library to implement more complex models, as well as potentially finding pre-built models in the given library to apply transfer learning techniques on.

Prior Work / References (this part can be over the 1-pg limit):

1. Mohil Maheshkumar Patel, Sudeep Tanwar, Rajesh Gupta, Neeraj Kumar, A Deep Learning-based Cryptocurrency Price Prediction Scheme for Financial Institutions, Journal of Information Security and Applications, Volume 55, 2020, 102583, ISSN 2214-2126, https://doi.org/10.1016/j.jisa.2020.102583.
2. Laura Alessandretti, Abeer ElBahrawy, Luca Maria Aiello, Andrea Baronchelli, "Anticipating Cryptocurrency Prices Using Machine Learning", Complexity, vol. 2018, Article ID 8983590, 16 pages, 2018. https://doi.org/10.1155/2018/8983590.
3. https://www.kaggle.com/kaushiksuresh147/bitcoin-prices-eda-and-prediction-r2-0-99
   A simple implementation of an ML algorithm on a BTC dataset
4. https://www.kaggle.com/jeongbinpark/lstm-can-we-predict-the-bitcoin-price
   A good source of initial intuition for recurrent neural networks to find long-term dependencies in our dataset.