Machine Learning Engineer Nanodegree

Capstone Proposal

Predicting Bitcoin Prices
Lukasz Tymoszczuk
September 23rd, 2018

Proposal

Domain Background

Bitcoin is the very first cryptocurrency, an implementation of the Blockchain - P2P distributed ledger. Bitcoin is also the most popular cryptocurrency. One of the biggest reason for this is that it eliminates intermediaries such as banks or financial institutions, governments. There is no single institution which has the power to stop/influence the transactions. This factor is important not only because the commission is reduced, but also by the philosophical reason of Bitcoin - no single institution can stop someone from transferring money to someone else [1].

Based on the recent Bitcoin prices it can be forecasted that it is either a financial bubble or future of payments. The author believes the latter option is very likely. It has the highest security by design, decentralisation and advanced technology used. Based on that reason predicting the behaviour of the currency will be more and more critical. The question how much bitcoin will be worth in the future is a sound question asked by hedge funds, trading institutions and even all those people who bought bitcoins. There were 22 million bitcoin wallets registered by June 2018 [2].

Problem Statement

There is a theory which says that financial instruments follow the random walk

(Random Walk Theory [3]), hence they cannot be predicted as any path is equally likely.

There are multiple hedge funds which started to trade bitcoin [4]. Can Machine Learning predict Bitcoin rise or fall?

If there is a trend in historical data, Machine Learning algorithms can find it, and they would perform better than flipping a coin. Therefore, the expected return from the investment would be positive (investment fees and bitcoin mining fees are ignored for the sake of the project).

Because Bitcoin like other stocks is a time series data, the LSTM seems to be a natural choice for solving the problem [5].

Datasets and Inputs

The data source with Bitcoin prices will be taken from kaggle.com [6].

The data for this problem is typical for stock prices: time series data. The period is over six years: Jan 2012 to July 2018.

The column used will be Date, Closing price and Volume.

Solution Statement

The typical approach for predicting time series data is LSTM (Long short-term memory) [7]. After the model is trained, the prediction can be made for the future prices. The LSTM algorithm will be used from Keras library [8].

There are two metrics I am going to use to determine if the algorithm is performing well.

1. Visualisation - the predicted prices will be combined with the real price on the chart. A suitable algorithm should produce a chart as close as possible to the real prices.
2. Another approach will be calculating the Root Mean Square Error (RMSE).

For replicability, all randomisation will be done around a hard-coded seed.

Benchmark Model

There will be three benchmark models against which the main algorithm will be compared.

The first (the simplest) benchmark model is a simulation of a random walk. It will choose the price (go up or go down) randomly, to simulate flipping a coin before deciding on the price.

The second benchmark model will be simple Machine Learning algorithm: Linear regression, which should aim in finding a line to fit the prices.

The last one is the popular algorithm used by investors and traders to track the price trend: moving averages. It removes day-to-day fluctuations and shows the trend of the instrument [9][10].

Evaluation Metrics

The solution will be measured by calculating RMSE between predicted and actual prices and compared with the RMSE for the benchmark models.

Project Design

The project will use the historical Bitcoin prices and the LSTM algorithm to create a model which will be tested on unseen data.

The below steps are planned for reaching the goal:

1. Understand the data. This step will be to understand if the input data does not have any gaps. The data will be visualised and compared with the different sources.
2. Preprocess the data: remove unused fields and normalise the values. For this algorithm, only three columns will be used: Date, Close Price and volume.

3. Split the dataset into training and testing set.
4. Run the train and test for benchmark models
    - (benchmark model) Random walk
    - (benchmark model) Linear regression
    - (benchmark model) Moving Averages
5. Implement the LSTM algorithm - use the train and test dataset
6. Calculate RMSE and compare which model is better.

Resources

1. https://bitcoin.org/bitcoin.pdf
2. https://www.bitcoinmarketjournal.com/how-many-people-use-bitcoin/
3. https://www.investopedia.com/university/concepts/concepts5.asp
4. https://www.quora.com/Which-hedge-funds-are-trading-Bitcoin
5. https://en.wikipedia.org/wiki/Long_short-term_memory
6. https://www.kaggle.com/mczielinski/bitcoin-historical-data
7. http://trap.ncirl.ie/2496/1/seanmcnally.pdf
8. https://keras.io/layers/recurrent/#lstm
9. https://medium.com/making-sense-of-data/time-series-next-value-prediction-using-regression-over-a-rolling-window-228f0acae363