Machine Learning Engineer Nanodegree

Predicting Ethereum prices using supervised machine learning

Nicolás Kittsteiner
November 16st, 2018

Proposal

Domain Background

In the world of finance, particular on investment and stock trading, the disruption of cryptocurrencies[[1]] or crypto assets[[2]] related to the blockchain[[3]] technology, created a new ecosystem of possibilities for investors. Based on the quantity of public information related to cryptocurrencies available on the Web it's possible to perform different analysis related of valuation and future price predictions using the appropriate key data and supervised machine learning algorithms.

Ethereum[[4]] is a cryptocurrency that handles a technology called Ethereum Virtual Machine (EVM)[[5]]. This allows to process *smart-contracts*[[6]] which is custom code statements that enable multiple use cases like, creating autonomous organizations, making crowdfunding projects, or applications that can transfer value automatically if the rules defined in the contract are processed accordingly. In this way it's possible to understand this invention as a commodity like for instance a cloud-based provided for computing.

Problem Statement

Investment on cryptocurrencies, could be a problem if we analyse the information about this assets manually. Supervised machine learning (SML)

can provide a way to simplify the analysis of the future prices based on past information of the market. There are vast information related to the prediction of stocks prices in finance (including cryptocurrencies) that use approach related to SML techniques [7]. In this way the approach followed to analyse and give a potential solution to the problem is to take a historical dataset, and use different SML techniques (Linear regression, K-Nearest Neighbors and Ensemble Methods) making a benchmark with the result on each technique.

Datasets and Inputs

The dataset used is a composition of different sources that has historical information of different cryptocurrencies, but in this analysis only two sets are considered (ethereum_price and ethereum_dataset) [8]. The fields considered on each dataset are:

- Ethereum Dataset (ethereum_dataset.csv):
  - Date(UTC) : Date of transaction
  - UnixTimeStamp : unix timestamp
  - eth_etherprice : price of ethereum
  - eth_tx : number of transactions per day
  - eth_address : Cumulative address growth
  - eth_supply : Number of ethers in supply
  - eth_marketcap : Market cap in USD
  - eth_hashrate : hash rate in GH/s
  - eth_difficulty : Difficulty level in TH
  - eth_blocks : number of blocks per day
  - eth_uncles : number of uncles per day
  - eth_blocksize : average block size in bytes
  - eth_blocktime : average block time in seconds
  - eth_gasprice : Average gas price in Wei
  - eth_gaslimit : Gas limit per day
  - eth_gasused : total gas used per day
  - eth_ethersupply : new ether supply per day
  - eth_chaindatasize : chain data size in bytes
  - eth_ens_register : Ethereal Name Service (ENS) registrations per day
- Ethereum prices (ethereum_price.csv):
  - Date : date of observation
  - Open : Opening price on the given day
  - High : Highest price on the given day
  - Low : Lowest price on the given day
  - Close : Closing price on the given day (*)
  - Volume : Volume of transactions on the given day
  - Market Cap : Market capitalization in USD

Solution Statement

The proposed solution includes an exploration to the data, preparing the data, merging and normalize the datasets, preprocessing and creating training and testing data. Also it's considered to make a random predictor to compare with each technique used, implementing the supervised machine learning pipeline for each algorithm (Linear regression, K-Nearest Neighbors and Ensemble Methods) described and finally make a model evaluation.

Benchmark Model

The benchmark consist on the evaluation of the accuracy score and F-score for each model, this in response to the close price which are expected to be predicted. The main idea is to figure out if any approach could have a good performance (more than 75% accuracy) so can be used on real trading scenarios.

Evaluation Metrics

As defined later, the evaluation for analysing the performance of each ML technique is using the accuracy score and F-score. In this way it's possible to determine which algorithm has the better results trying to predict close prices for the next day.

Project Design

Also as described before in the solution statement section. There are tasks for data exploration, preprocessing and normalization of each dataset, merges of information and the implementation of predictors for each ML technique explained in the solution statement.

---

[1] Burniske, Tatar: "Cryptoassets: The Innovative Investor's Guide to Bitcoin and Beyond" ISBN: 978-1-26-002668-9
[2] Ídem nº1
[3] Nakamoto: "Bitcoin: A Peer-to-Peer Electronic Cash System", https://bitcoin.org/bitcoin.pdf
[4] Ethereum, "Ethereum White Paper", https://github.com/ethereum/wiki/wiki/White-Paper
[5] Ídem nº4
[6] Ídem nº4
[7] Madan, Saluja, Zhao: "Automated Bitcoin Trading via Machine Learning Algorithms" http://cs229.stanford.edu/proj2014/
[8] Kaggle: "Cryptocurrency Historical Prices" https://www.kaggle.com/sudalairajkumar/cryptocurrencypricehistory/home