Machine Learning Nanodegree
Capstone Proposal

Jeroen F.L. Schmidt

May 12, 2018

1 Introduction

This capstone project aims to apply convolutions networks to time series data too predict Etherium price movement using historical Bitcoin and Etherium trade data and social media sentiment analysis from Twitter tweets and Reddit comments.

1.1 Problem Statement

Can the price of Etherium be predicted by using Etherium and Bitcoin trade data along with sentiment analysis performed on Twitter tweets and Reddit comments?

2 Domain Background

Bitcoin is a peer-to-peer electronic cash system that uses a decentralised ledger to establish trust and the largest global open-source crypto-currency by market cap. Ethereum in the same vain as Bitcoin is also a peer-to-peer decentralised ledger but it's use case is more general as it allows for smart contracts. Smart contracts allow for the digital negotiation and execution of a contract between two or more parties while ensuring trust between said parties. Etheriums smart contracts allows it to be used as a crypocurrency; it is currently the second largest crypo-currency by market cap. Bakar and Rosbi[2] showed that there exists a strong positive correlation between the price of Etherium and Bitcoin.

Time series classification and regression problems are an important area of study within academia and industry. It remains a hard topic of study because of the temporal dimensional of the data and the high levels of dimensional that has to be dealt with as a result. Within the field of statistics many head ways have been made in time series analysis through stuff, stuff, stuff. Often when machine learning models are used on temporal data, the temporal data is dimensional reduced in-order for the machine learning method to handle it; this is done through feature engineering by creating bucketed features like move averages, aggregated counts and lagged values.

In recent years several researchers have started to tackle time series problems by using neural network methods. A. Radityo and Budił[1] use a variate of artificial neural networks (ANNs) to try do a regression predict of Bitcoin prices. They achieve varying degrees of success and offer ideas of what features to use when training an ANN.

In Zebik et al.[8], it was shown that Convolutions Neural Networks (CNN) could be applied to temporal data and in many cases outperform traditional time series modelling methods. Zebik et al. argued that CNNs are able to create their own complex features that retain most of the temporal information of the data. In this way CNNs are superior to ANN models.

Research has also been done on the use of sentiment analysis to predict market behaviour. Bollen, Mao, and Zeng[3] was able to predict market movement four days in advance with a 87% success rate by combining twitter and chat form data. This research was extended by Matta, Lunesu, and Marchesi[7] too Bitcoin price prediction, by using 60 days worth of Twitter data (40 000 tweets a day) they were able to show that positive twitter sentiments had a 30% cross correlation.

2.1 Motivation

My personal goal for this capstone project is to further refine my deep learning and sentiment analysis skills in the domain of time series data. I personally find the application of machine learning to time series problems very interesting because the vast majority real world problems are framed with respect to temporal behaviour and because these types of problems are rarely correctly solved unless they employ advanced time series analysis techniques.

3 Datasets and Inputs

The following section will outline the basic charactersitics of the data that will be used and what their sources are. Two years worth of data will be collected from each data source.

3.1 Crypto Currency Ticker Data

The Poloniex public api can be used to obtain historic Etherium and Bitcoin price data [6]. The ticker data will be collected in intervals of 5 minutes. This will be done for a span of 2 years. This will result in 210240 observation per cryptocurrency.

Column Name	Date Type	Description
close	FLOAT	closing price of observation time span
date	integer	unix time of observation
high	FLOAT	max price observed in the time interval of observation
low	FLOAT	min price observed in the time interval of observation
open	STRING	price of cryptocurrecy at the beginning of observation

Table 1: Data Structure of Poloniex Ticket Data - Both Etherium and Bitcoin

3.2 Twitter Data

Twitter tweets can be scraped using a public github library [5] by Henrique. This project will try collect the same number of tweets as in Bollen, Mao, and Zeng[3], where they used on average 40000 tweets per day. This will result upto 29200000 tweets per cryptocurrency over 2 years. It should be expected that there will be a discrepancy of the number of tweets obtained for Bitcoin vs the number of tweets obtained for Etherium will be different because of the larger popularity and Bitcoin.

Column Name	Date Type	Description
username	STRING	author of tweet
date	DATETIME	date and time that tweet was tweeted
retweets	INTEGER	number of times someone else retweeted the tweet
favorites	INTEGER	number of times another user favourited
text	STRING	text content of tweet
geo	STRING	geographic location of tweet
mentions	INTEGER	how many times the tweet was mentioned
hashtags	STRING	hashtags present in tweet
id	INTEGER	unique ID of tweet
permalink	STRING	direct URL link to tweet

Table 2: Data Structure of Twitter Date

3.3 Reddit Data

Historical reddit comments can be accessed on the public Googles big query data repositories [4]. The Bitcoin subreddit produces between 2000000 to 10000000 words per month, this results in a total count of between 48000000 and 240000000 words over two years. The Etherium subreddit produces between 500000 to 1000000 words per month, this results in a total count of between 12000000 and 24000000 words over two years.

Column Name	STRING	Description
body	STRING	text of comment
name	STRING	tier and comment id
author	STRING	comment author user name
author_flair_text	STRING	flair attached to comment authoer
created_utc	INTEGER	unix time of when comment was posted
subreddit_id	STRING	id of subreddit where comment was posted
parent_id	STRING	comment id that this observation replied to
score	INTEGER	comment score assigned by community
retrieved_on	INTEGER	unix time of comment retrieval
controversiality	INTEGER	reddit metric measuring controversiality of comment by the score behaviour of comment
gilded	INTEGER	number of guilds the comment recived
id	STRING	system id of comment
subreddit	STRING	subreddit the comment was posted in

Table 3: Data Structure of Reddit Data

4 Solution Methodology

Building off the research in research in [3], [7] it might be possible to use sentiment analysis of tweets and reddit comments to build features that will aid in the prediction of Etherium price. From what was outlined in [1], [8], [2] it might be possible to predict Etherium price movement by using CNNs on historical Etherium and Bitcoin prices.

The solution will try predict two different types of outputs that indicate market movement. The most desired solution would be to predict the actual price value of Etherium for some point in the feature, the

second desired solution would be to predict a categorical result which informs us if the price will go up or down for some point in the future.

5 Evaluation Metrics

Time series data requires some special considerations in how it is handled when creating train, cross validation (CV) and testing data sets. A common base assumption made with temporal data is that an observation x_t at time t has some relation to past observation(s) x_{t-i} at time $t-i$. The data must be slit according to distinct time groupings in order to respect the temporal order of the data; in other words, the model should be trained on only past data and tested against hold out data from the future.

The implication of having to preserve the temporal order of our data means that we can not use k-folds cross validation, instead we will have to use Walk-Forward Validation.

The model will be evaluated on two metrics of prediction:

- Can the model predict if the price of the cryptocurrecy will go up or down? (categorical prediction)
- Can the model predict the value that the cryptocurrecy will go to for the next time interval? (regression prediction)

Metrics well suited to asses regression predictions are:

Absolute Measures

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{x_i} - x_i)^2} \tag{1}$$

$$\text{MAD} = \frac{1}{n}\sum_{i=1}^{n}|\hat{x_i} - x_i| \tag{2}$$

Relative Measures

$$U = \frac{\sqrt{\sum_{i=1}^{n}(\hat{x_i} + x_i)^2}}{\sqrt{\sum_{i=1}^{n}(x_i + x_{i-1})^2}} \tag{3}$$

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\hat{x_i} - x_i}{x_i}\right| \tag{4}$$

The F_β score is the best metric to evaluate the classification model,

$$F_\beta = (1+\beta^2)\frac{\text{precision} \cdot \text{recall}}{(\beta^2 \cdot \text{precision} + \text{recall})} \tag{5}$$

6 Benchmark Model

The following benchmark models could be used to asses the catergocial model:

- Baseline model: 50% guess that the crypocurrency will go up or down
- Logistic Regression Model

The following benchmark models could be used to asses the regression model:

- Baseline model: Mean or Medium of the data for a specific time interval
- Regression Model

7 Project Design

Both desired solutions will be concurrently solved by using the following methodology:

Step 1: Build a CNN model that only uses the ticker trade data from Bitcoin and Ethereum. This will involve experimenting with the architecture presented by Zebik et al. and building features that don't reduce the temporal dimensionality of the data.

Step 2: Explore, understand and perform sentiment analysis on the Twitter and Reddit data. Step 2 will start by using the sentiment techniques used in [3].

Step 3: Build features from the insights gained in step 2. A lot of feature experimentation will have to be conducted, like the bucket time intervals that the tweets are aggregate into.

Step 4: Integrate sentiment features into the CNN and then repeat all the steps from the beginning.

References

[1] Q. Munajat A. Radityo and I. Budił. "Prediction of Bitcoin exchange rate to American dollar using artificial neural network methods". In: *2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*. Oct. 2017, pp. 433–438. DOI: 10.1109/ICACSIS.2017.8355070.

[2] Nashirah Abu Bakar and Sofian Rosbi. "STATISTICAL DIAGNOSTICS FOR BIVARIATE CORRELATION AND REGRESSION ANALYSIS BETWEEN CRYPTOCURRENCY EXCHANGE RATES OF BITCOIN AND ETHEREUM". In: ().

[3] Johan Bollen, Huina Mao, and Xiaojun Zeng. "Twitter mood predicts the stock market". In: *Journal of computational science* 2.1 (2011), pp. 1–8.

[4] Google. *Google Big Query Public Data Repo*. 2018. URL: https://bigquery.cloud.google.com/dataset/fh-bigquery:reddit_comments (visited on 05/21/2018).

[5] Jefferson Henrique. *GetOldTweets-python*. https://github.com/Jefferson-Henrique/GetOldTweets-python. 2018.

[6] Poloniex LLC. *Poloniex Public API*. 2018. URL: https://poloniex.com/support/api/ (visited on 05/21/2018).

[7] Martina Matta, Ilaria Lunesu, and Michele Marchesi. "Bitcoin Spread Prediction Using Social and Web Search Media." In: *UMAP Workshops*. 2015.

[8] Mariusz Zebik et al. "Convolutional Neural Networks for Time Series Classification". In: *International Conference on Artificial Intelligence and Soft Computing*. Springer. 2017, pp. 635–642.