Machine Learning Engineer Nanodegree - Final Project Report

Project Report

By Mathieu Deiber
Date :26/07/2021

Introduction

Project Overview

Bitcoin was first introduced in 2009 and is a new type of currency, often referred to as cryptocurrency. There are many other cryptocurrencies but Bitcoin could be considered as the first one that really took off and is gaining in popularity every year. Bitcoin is based on the technology known as 'Blockchain' and can be compared to gold as it has a limited supply (capped to 21 million Bitcoins), it is easily divisible, durable, and can easily be transferred from one user to another via a decentralized network. Currently, there are around 20 million active addresses every month (https://studio.glassnode.com/). In 2009, 1 Bitcoin was worth less than $0.01, and in April 2021 the price of one Bitcoin exceeded $60,000 (https://coinmarketcap.com/currencies/bitcoin/). Bitcoin is also known to be extremely volatile, it is not uncommon to see the price fluctuating by more than 10% in a single day.

Problem Statement

For an investor, it will be extremely valuable to know when to buy and when to sell Bitcoin. The goal of this model would be to predict if the price is going up or down based on the Tweet mentioning Bitcoin or BTC. This project will be completed by using Natural Language Processing NLP techniques and classify the tweets of one day into 2 categories (i.e. 'TOMORROW GOING UP' or 'TOMORROW GOING DOWN').

Metrics

The metric that will be used for this classifier will be Accuracy. We need the highest accuracy to know when to buy or sell a Bitcoin and be profitable over the long run.

Analysis

Data Exploration and Visualization

First, the Bitcoin price historical dataset will be explored. The dataset includes the following variable:

- Data
- cmc_rank: Bitcoin ranking among other cryptocurrencies
- slug: the denomination of the coin (i.e. Bitcoin)
- symbol: Bitcoin symbol (i.e. BTC)
- market_cap: The total value of the Bitcoin market in US$
- price: the current price
- circulating_supply: Bitcoin circulating supply
- volume_24h: volume of Bitcoin exchange the last 24h
- percent_change_1h: price change the last hour
- percent_change_24h: price change the last 24 hours
- percent_change_7d: price change the last 7 days

An example of the dataset is shown in the table below:

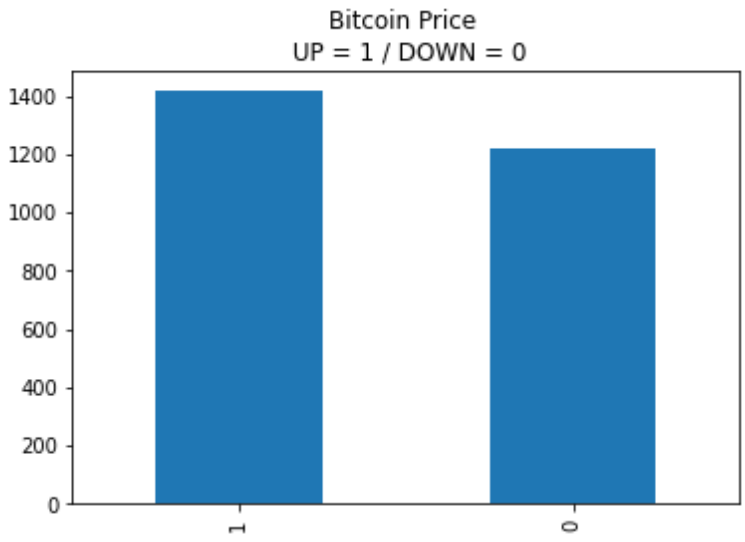
	date	cmc_rank	slug	symbol	market_cap	price	circulating_supply	volume_24h	percent_change_1h	percent_change_24h	percent_change_7d
0	02/02/2014	1	Bitcoin	BTC	10190038594	825.37	12346025	11300875.0	-0.03	-1.25	-6.75
1	03/02/2014	1	Bitcoin	BTC	10174774413	823.83	12350575	13940069.0	0.09	-0.11	4.26
2	04/02/2014	1	Bitcoin	BTC	10229342576	827.96	12354875	16609697.0	-0.13	0.56	1.10
3	05/02/2014	1	Bitcoin	BTC	10034435954	811.91	12359050	22395116.0	-1.47	-2.08	-1.59
4	06/02/2014	1	Bitcoin	BTC	9662790968	781.55	12363625	50108744.0	0.07	-4.19	-4.68

There are no missing values and a few statistics about the data are shown in the table below:

	cmc_rank	market_cap	price	circulating_supply	volume_24h	percent_change_1h	percent_change_24h	percent_change_7d
count	2636.0	2.636000e+03	2636.000000	2.636000e+03	2.636000e+03	2636.000000	2636.000000	2636.000000
mean	1.0	9.742015e+10	5481.743790	1.614905e+07	5.449159e+11	0.001385	0.179450	1.509621
std	0.0	1.446578e+11	7809.864979	1.847997e+06	2.746627e+13	0.822845	3.836825	10.620501
min	1.0	2.444375e+09	178.100000	1.234602e+07	2.857834e+06	-15.890000	-37.140000	-45.310000
25%	1.0	6.754451e+09	472.905000	1.471144e+07	4.570815e+07	-0.180000	-1.250000	-4.192500
50%	1.0	4.663255e+10	2843.505000	1.648004e+07	1.437850e+09	0.010000	0.160000	0.675000
75%	1.0	1.468193e+11	8293.980000	1.771492e+07	1.524146e+10	0.220000	1.650000	6.465000
max	1.0	1.078136e+12	57805.120000	1.865214e+07	1.410183e+15	5.350000	24.890000	79.700000

There are 11 parameters available but we are mainly interested in the percent_change_24h as the goal of this model is to predict if the price will go up or down. So we will focus our attention on this column. We can notice that the price of Bitcoin is very volatile as we can see changes of -37.1% or 24.9% in a single day.

Bitcoin was up 53.7% of the time between 2/2/2014 to 12/3/2021. (See figure below).



We will now be looking at the second dataset which consists of 14M tweets scrapped from twitters between 1/1/2016 and 29/3/2019.
The dataset includes the following parameters :

- id: a unique tweet id
- user: the username
- fullname: the fullname
- url: an url
- timestamp: the time the tweet was posted
- replies: the number of replies
- likes: the number of likes

- retweets : the number of retweets
- text: the tweet

The first lines of the dataset are shown below:

	id	user	fullname	url	timestamp	replies	likes	retweets	text\r
0	1.132977e+18	KamdemAbdiel	Abdiel kamdem	NaN	2019-05-27 11:49:14+00	0	0	0	È appena uscito un nuovo video! LES CRYPTOMONN...
1	1.132977e+18	bitcointe	Bitcointe	NaN	2019-05-27 11:49:18+00	0	0	0	Cardano: Digitize Currencies; EOS https://t.co...
2	1.132977e+18	3eyedbran	Bran - 3 Eyed Raven	NaN	2019-05-27 11:49:06+00	0	2	1	Another Test tweet that wasn't caught in the s...
3	1.132977e+18	DetroitCrypto	J. Scardina	NaN	2019-05-27 11:49:22+00	0	0	0	Current Crypto Prices! \n\nBTC: $8721.99 USD\n...
4	1.132977e+18	mmursaleen72	Muhammad Mursaleen	NaN	2019-05-27 11:49:23+00	0	0	0	Spiv (Nosar Baz): BITCOIN Is An Asset & NO...

The datasets containt more than 16M lines and have 9 columns. We can also notice that some of the tweets are not in english. As we only want to get the daily sentiment of all the tweets for a particular date we will only keep the following columns:

- timestamp
- replies
- likes
- retweets
- text

Something else important to notice is that many tweets have 0 likes. This is probably because the person who tweeted it only has a small audience or the tweet is simply not interesting. At this stage, I am making the decision to only include tweets that have more than 10 likes - this shows that the tweet brought some kind of value to other users and reach at least a small audience.

Algorithms and Techniques

The problem is very similar to a sentiment analysis problem. The techniques that will be used to solve the problem are standard NLP processing techniques such as :

- convert the text to lower case,
- keep only alphanumeric characters,
- remove stopwords,
- stemming and lemmatization,
- bag of words (based on the training dataset only)

The machine learning algorithm that will be used will be a classification algorithm like the random forest classifier, XGBoost or a neural network.

Benchmark

The benchmark will be that the model should perform better than a basic model predicting that the price of Bitcoin will always go up. This model should perform better than that to have a chance to be valuable to investors.

Methodology

Data Processing

After exploring the data, the first step was to work on the tweet dataset by selecting the tweets that had more than 10 likes (this show that the tweet brought some kind of value to other users and reach at least a small audience) and filter by language and retain the tweets that are in English only. This was done by applying a filter on the column likes and using the library spacy to detect le language. An overview of the dataset after this step is shown in the figure below:

3]:

	level_0	index	timestamp	replies	likes	retweets	text\r	language
0	1	10	2019-05-27 11:49:19+00	0	14	2	One of the useful articles of Stefan; here is ...	en
1	2	11	2019-05-21 16:49:45+00	47	81	84	BTC IS STILL GOING STRONG!!\n\nThus, we are gi...	en
2	5	14	2019-05-27 08:13:06+00	5	167	68	Bitcoin Price Hits $8,939 in New 2019 High: Wh...	en
3	7	18	2019-05-27 11:27:22+00	1	19	6	You have roughly 6 days left to get your #Laun...	en
4	8	19	2019-05-27 08:32:08+00	14	40	39	BTC IS GOING CRAZYYY!\n\nThus, we are giving a...	en

As the previous step took a couple of hours, the new dataset was saved and then reopen in a new notebook (Step 2 - Preparing Processing and Modelling the Data).

Then the dataset needed to be cleaned by deleting the unused column and retain only the following ones:

- replies
- likes
- retweets
- text (renamed tweet)

Once this is done, the dataset was resampled to a daily timestep. The replies, likes, and retweets were summed over a day and the tweets were concatenated. After this step, the dataset is shown below:

timestamp	replies	likes	retweets	tweet
2016-01-02	1	52	70	Great infographic on the state of #blockchain ...
2016-01-04	1	25	23	7 Years of Bitcoin: Genesis Block Mined 2009-0...
2016-01-11	5	38	21	2. I'm not sure if this is a risk that @coinba...
2016-01-12	12	75	86	Bitcoin users - please help me understand your...
2016-01-14	31	327	345	#MustRead Full-time Bitcoin developer: Bitcoin...

At this stage, we can separate our dataset into a training and testing dataset. Tweets between 1/2/2016 and 28/2/2019 were used to train the model while the tweets posted after 1/3/2019 were used to predict the price movement and test the model.

The Bitcoin price dataset was used to get the label for the training and testing dataset. The column percent_change_24h was used to determine the price movement. It should be noted that the 24h price movement at a date T+1 needed to be offset by 1 day to be in line with the tweet posted at a date T. If the price went up the label 1 was assigned and if it went down the label 0 was assigned. The label for the training and testing data is now ready.

Now we will process the data and follow standard NLP processing techniques. (convert the text to lower case, keep only alphanumeric characters, remove stopwords, stemming and lemmatization, a bag of words (based on the training dataset only)). Once all these steps are achieved we will reformat our data_training and data_testing data frame accordingly before moving to the classification model. This was done using the nltk, beautifulsoup and sklearn (CountVectorizer) libraries. Vocabulary size of 5000 words was selected and the daily number of likes, replies, and retweets were also included in the training and testing datasets.

The final training dataset has 946 lines and 5003 columns.

Implementation

Three models were considered and two were tested:

1. Random Forest Classifier
2. XGB
3. Neural Network

After cleaning the data and merging all the tweets of the same day. We ended up with a relatively small dataset, therefore, I will only test the random forest classifier and XGB. I will start by testing the two models and optimize the most promising one.

Refinement

As the performance of the 2 models is similar, the Random Forest Classifier was refined to improve its performance. A total of 672 parameters configurations were tested to find the best Random Forest Classifier for this problem. The model was tuned using GridSearchCV from the library Sklearn. Combinations of the following parameters were tested:

- bootstrap: True, False
- max_depth: 3, 5, 8, 12, 20, 50, 100
- max_features: sqrt, log2
- min_samples_split: 2, 3, 4
- n_estimators: 5, 10, 20, 50, 100, 200, 500, 1000

The final parameters selected are:

- bootstrap: True
- max_depth: 3
- max_features: sqrt
- min_samples_split: 3
- n_estimators: 100

Results

Model Evaluation and Validation

The tweets posted before 1/3/2019 were used to train our model while the tweets posted after 1/3/2019 were used to test the model. Approximately 70% of the data will be used to train the model and 30% of the data will be used to test the model.

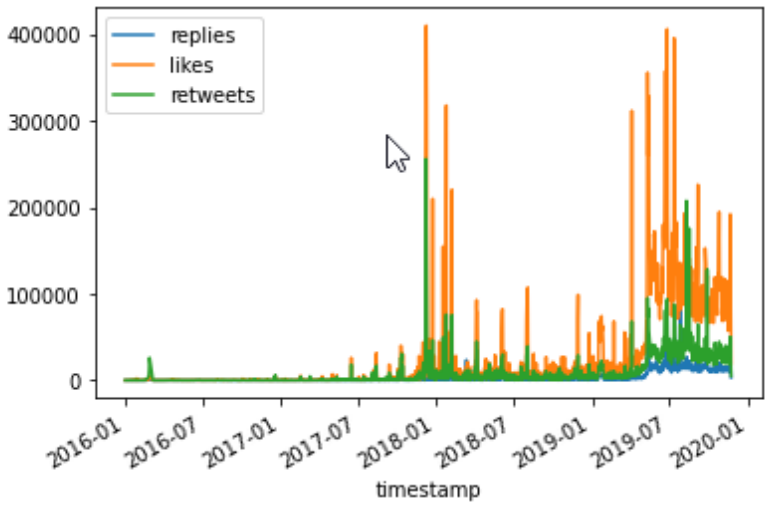
Both preliminary models gave similar results, the random forest model is marginally better. Both models are performing worse than the benchmark model (accuracy of 54.1%) with an accuracy of 51.1% for the random forest classifier model and an accuracy of 50.3% for the XGBoost model. The random forest parameters were tuned to improve its performance but the accuracy didn't improve (accuracy of 49.6%).

The final accuracy of the model is similar to a coin flip. The model is not capable to predict the price movement based on the overall sentiment infer from the tweets.

Justification

Potential reasons for the model poor performances:

- Bitcoin adoption is increasing every year and words use when talking about Bitcoin or the crypto market are evolving. The average tweet in 2016 is completely different from a tweet in 2019.
- Only tweets in English were considered, other languages might contain some useful information (such as Spanish or Chinese)
- The interaction on Twitter was completely different for the testing dataset compared to the training dataset (figure below - we can see a sharp increase in interaction around March/April 2019)



Conclusions

Bitcoin price is known to be extremely volatile and it is very difficult to predict the price over a short period of time. No correlation was found between the daily tweets about Bitcoin and its price movement. This model could potentially be improved by selecting only the tweets from known 'experts' or 'influencers' rather than

including all the tweets. Bitcoin price is also strongly impacted by the latest news related to Bitcoin - this can be positive (a big company accepting Bitcoin as payment) or negative (a country is threatening to ban Bitcoin). News articles might also be a good source of information to predict Bitcoin prices.