Machine Learning Engineer UDACITY Nanodegree

Capstone Proposal

By Mathieu Deiber
Date :14/07/2021

Domain Background

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').

Datasets and Inputs

Two datasets will be used, the first dataset will be download from Kaggle (https://www.kaggle.com/alaix14/bitcoin-tweets-20160101-to-20190329) which contains 16 million tweets scrapped from twitters between 1/1/2016 and 29/3/2019

and include the following information: id, user, fullname, url, timestamp, replies, likes, retweets, text.
The second dataset consists of the historical price of Bitcoin. The dataset was obtained by scraping the daily price from the website https://coinmarketcap.com/ between 2/2/2014 to 12/3/2021, this dataset contains the following information: date, price, 24h volume, and 24h price change.

Solution Statement

The Bitcoin market is extremely volatile and strongly correlated to people's sentiment. Over the long term, the price of Bitcoin is predicted to go up by many Bitcoin enthusiast but over the short term, the price of Bitcoin is unpredictable. The model will help investors to make a decision to either buy bitcoins (if the price is predicted to go up) or sell bitcoins (if the price is going down).

Benchmark Model

The benchmark will 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.

Evaluation 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.

Project Design

The model will be using the first dataset (tweets downloaded from Kaggle) and predict if the 24h price change is positive or negative (this information will be extracted from the second dataset).

The following steps will be implemented:

1. Processing of the tweets (language: English, word tokenization, lemmatization, stop word)

2. Split the data into training (1/1/2016 to 1/10/2018 data) and testing (1/10/2018 to 29/3/2019) sets.
3. Create a dictionary of words based on the training data.
4. Processing of the daily tweets (dictionary of words, number of likes, number of retweets).
5. Try different models such as XGboost or neural network.
6. Evaluate the performance of each model (Accuracy)