E4_report
Predicting Bitcoin Price

Yangfan Tan
Gleb Kudrin
Rasmus Soome

November 2022

1 Business understanding

1.1 Identifying business goals

1.1.1 Background

Trading Bitcoin has immense potential in building personal wealth. The problem is that predicting the future price of Bitcoin is not easy, given its volatility. When being careless while trading, there is a big risk of losing your entire investment.

In this project, we attempt to predict the future movement of Bitcoin prices using machine learning, with historical prices of Bitcoin, other financial variables, and people's sentiments about Bitcoin.

Our project falls into the field of financial technical analysis. In short, technical analysis is a trading discipline employed to evaluate investments and identify trading opportunities by analyzing statistical trends gathered from trading activity, such as price movement and volume. Machine learning is one of the most powerful tools for predicting the future using past data.

1.1.2 Business goals and success criteria

Our goal is to predict the price movement direction of Bitcoin, either going up or down. This decision is made because predicting the movement of Bitcoin price has more business value, in that it can guide people's trade decisions: to buy in or out. Therefore, we are dealing with a classification problem. (And if we have more time, we will also try to predict the actual prices of Bitcoin, which turns into a regression problem.)

Given the volatility of Bitcoin price and financial markets in general, it would be too ambitious to set too high criteria. But our model should perform better than random guessing. Therefore, we set the success criteria of this project to be 60% prediction accuracy (later in this project we will show this is a balanced dataset) and more than 0.6 AUC.

1.2 Assessing the situation

1.2.1 Inventory of resources

- Bitcoin prices and other financial/economic data from Yahoo Finance, and *yfinance* module to download data in Python
- A dataset of tweets on Bitcoin from Kaggle (more detail in Data Understanding part)
- Python 3 and various data processing and machine learning modules

1.2.2 Requirements, assumptions, and constraints

We should warn that our project is not aiming for investment advice in real life and no financial transactions should be conducted based on this project. Otherwise, there are no legal or security considerations since all the data and tools are openly available.

The following assumptions are made:

- Technical analysis can be useful to predict the price of an asset.
- Historical prices are helpful to predict future prices.
- Financial data from Yahoo finance are accurate.
- Tweets about Bitcoin can work as a valid proxy for people's sentiment on Bitcoin and can be generalized to reflect real-world sentiment about Bitcoin.
- Bitcoin prices may move together with other financial variables/features.

1.2.3 Risks and contingencies

- Risk: *yfinance* API stops working
- Contingency: Find an API that works analogous or download data manually

1.2.4 Terminology

- Cryptocurrency: a digital currency designed to work as a medium of exchange through a computer network that is not reliant on any central authority, such as a government or bank, to uphold or maintain it
- Bitcoin, Ethereum, and Tether: specific examples of cryptocurrency
- Sentiment analysis: use of natural language processing, text analysis, computational linguistics, and biometrics to systematically identify, extract, quantify, and study affective states and subjective information.
- S&P 500: a stock market index tracking the stock performance of 500 large companies listed on stock exchanges in the United States
- Dow Jones Industrial Average: a stock market index of 30 prominent companies listed on stock exchanges in the United States
- NASDAQ Composite: a stock market index that includes almost all stocks listed on the Nasdaq stock exchange

- Treasury Yield 10 Years: used as a proxy for mortgage rates. It's also seen as a sign of investor sentiment about the economy.

1.2.5 Costs and benefits

- Costs: Time spent on it
- Benefits: Models with the potential to predict the movement of Bitcoin prices.

1.3 Data mining goals

1.3.1 Goals

- To obtain the historical price data of Bitcoin
- To obtain other economic/finanial variables
- To obtain past tweets about bitcoin and conduct sentiment analysis on them
- To clean, process and transform the data into a form to conduct machine learning
- To train models to predict price movement direction

1.3.2 Success criteria

- A clean and tidy final dataset
- >= 60% prediction accuracy
- >= 0.6 AUC

2 Data understanding

2.1 Gathering data

2.1.1 Outline data requirements

We mainly need 3 datasets. One for label and two for features.

Label:

1) Historical prices of Bitcoin
   Type: numerical
   Daily price from 2017-09-01 to 2022-09-01 (five years in total).

Features:

2) Other 9 economic/financial variables

including S&P 500, Dow Jones Industrial Average, NASDAQ Composite, Crude Oil price, Gold price, EUR-USD exchange rate, Treasury Yield 10 Years, Ethereum price, and Tether price.
We hypothesize these variables/features may move together with Bitcoin prices.
Type: numerical
Daily price from 2017-09-01 to 2022-09-01 (five years in total).

To make the measurement consistent, all variables are measured in current US dollars (USD).

3) People's sentiment on Bitcoin, proxied by tweets.
Type: text/string
The sentiment database must contain the comments themselves which would show a particular person's attitude towards bitcoin and its trends (whether the person feels the popularity of bitcoin is going up or down). Also for each person, it would be really beneficial to know how many people are associated with them on Twitter (i.e. their number of friends and followers)

2.1.2 Verify data availability

For historical prices of Bitcoin and other 9 economic/financial variables, we gather data from Yahoo Finance, an online platform that records various financial statistics.
To download them, we use *yfinance*, a Python module that allows downloading of market data from Yahoo! Finance's API.

For People's sentiment on Bitcoin, we proxy it by tweets. And we use a dataset from Kaggle, which collects all the tweets with trending #Bitcoin and #btc hashtags.

2.1.3 Define selection criteria

For historical prices of Bitcoin and other economic/financial variables, we download daily data from Yahoo Finance. Among all the columns (see Table 1), the relevant ones are *Date* and *Close*.

Date	Open	High	Low	Close	Volume	Dividends	Stock Splits
2014-09-17 (	359.5462	361.46851	351.58688	355.95737	16389166	0	0
2014-09-18 (	355.58841	355.5054	319.78946	328.53937	26691849	0	0
2014-09-19 (	328.2785	330.93671	298.92102	307.76114	29560103	0	0
2014-09-20 (	307.66525	329.97818	303.93124	318.75897	28736826	0	0

Table 1

Columns *Dividends* and *Stock Splits* are irrelevant as they only apply to publicly traded stocks and they should be dropped. There are four types of price data: *Open*, *High*, *Low*, *and Close*. We choose *Close* for all which is the trading price at the end of the trading day on a financial market.

For People's sentiment on Bitcoin, the only file used for the sentiment analysis is 'Bitcoin_tweets.csv'. There are many attributes, but the ones that we'll need are 'user_followers' of int type, 'user_friends' of int type, 'text' as a string, and 'date' as a time date.

	user_followers	user_friends	date	text
0	8534.0	7605	2021-02-10 23:59:04	Blue Ridge Bank shares halted by NYSE after #b...
1	6769.0	1532	2021-02-10 23:58:48	😎 Today, that's this #Thursday, we will do a "...
2	128.0	332	2021-02-10 23:54:48	Guys evening, I have read this article about B...
3	625.0	129	2021-02-10 23:54:33	$BTC A big chance in a billion! Price: \487264...
4	1249.0	1472	2021-02-10 23:54:06	This network is secured by 9 508 nodes as of t...
...	...	...	...	...

Table 2

2.2 Describing data

The historical prices of Bitcoin and other economic/financial variable data are downloaded from Yahoo Finance. They are measured in USD, from 2017-09-01 to 2022-09-01. The total number of observations is 1826 (days) * 10 (columns) = 18, 260.

Initially, the tweet database has the following attributes:*User_name, user_location, user_description, user_created, user_followers, user_friends, user_favorites, date, text*. But we need only the attributes mentioned above (see Table 2). The number of instances is too huge to estimate on our machines right now, so for the time being we'll take the first 100 000 instances. The datastes also contain NaN values, with which we should deal with caution.

2.3 Exploring data

For Bitcoin prices, we can see from the Figure1 that it dramatically surged after 2021, leading to a large standard variation (see Table 2). That's why it makes sense to predict the movement of Bitcoin prices (going up or down), rather than the absolute values.

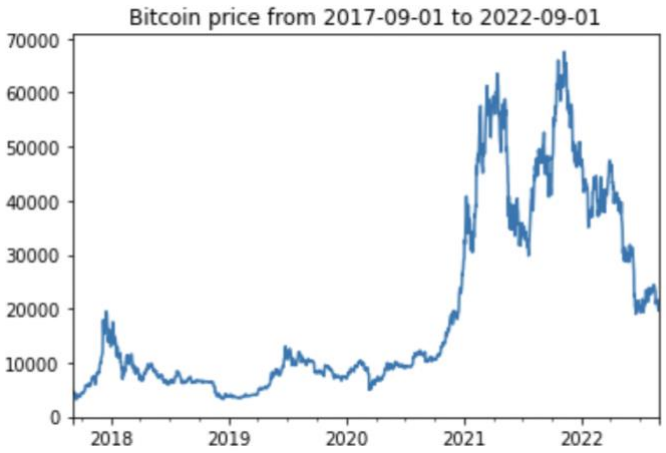

Figure 1

	btc
count	1826.000000
mean	19648.554517
std	17290.028456
min	3154.949951
25%	7245.359375
50%	10183.570801
75%	33740.259766
max	67566.828125

Figure 2

After processing the price data into a binary movement label (1 for going up and 0 for going down), there are 955 observations for 1 and 871 for 0, making it a rather balanced dataset.

For sentiment data, the number of followers and the number of friends in the dataset is normally distributed. The maximum number of followers is 3,977,223, the minimum number of followers is 0, the mean number of followers is 6,921.7605; the maximum number of friends is 145,451, the minimum number of friends is 0, and the mean number of friends is 1,217.7882.

The starting point of gathering the data in the dataset is 2021-02-06, and it is updated every day. The comments themselves vary a lot, so it's possible to see only a few-word comment or, on the other hand, a comment containing several sentences. But in the actual data analysis, we're going to use only the first 200-250 characters of a given comment. That presumably should represent the contents of the comment in the most effective and accurate way

2.4 Verifying data quality

So far, all these data are good enough for this project as they contain all the information we need and no quality data problem has been found.

3 Planning the project

3.1 Subtasks

Since the tweet dataset only starts from 2021-02-06, we will train 2 sets of models, one for the *whole* dataset without the sentiment feature, and one for the dataset covering the period from 2021-02-06 to 2022-09-01 with the sentiment feature.

Index	*Task*	*Required hours*	*Responsible person*
1	Download, clean and process financial data from yahoo finance	10	Yangfan
2	Download, clean and process the tweet data set	15	Rasmus and Gleb
3	Sentiment labelling and analysis	15	Rasmus and Gleb
4	Merge the financial dataset with the sentiment dataset	5	Yangfan and Gleb
5	Train models	30	All
6	Evaluate error and improve models	15	All
7	Make a poster	10	All
	sum	100	

3.2 Overview of models

Baseline model: Random forest
Other classification algorithms:

- Decision tree
- KNN
- SVC

Logistic regression
More advanced tools:

- Neural network
    - LSTM
- XGBoost