Introduction to Data Science

PREDICTING BITCOIN PRICE MOVEMENTS

E4: Yangfan Tan, Gleb Kudrin, Rasmus Soome

December 2022

1. INTRODUCTION

Business value: Trading Bitcoin has immense potential in building personal wealth.

Erik Finman: a self-made millionare at the age of 18 by purchasing about $1,000 worth of Bitcoin

But predicting the future price of Bitcoin is not easy, given its **volatility**, both in long term and short term (Chart 1), due to limited supply and lack of a central bank

Goal: To predict the next-day movement direction of Bitcoin prices

Label: a binary movement label

- 1 for up, 0 for same/down, compared to tomorrow price → **Classification** problem

Features:

- 9 other financial variables (Table 1)
- + People's sentiments about Bitcoin (section 5)

Stock indexes	S&P 500, Dow Jones, NASDAQ
Other cryptocurrency prices	Ethereum, Tether
Other crucial prices	Oil, Gold, EUR-USD

Table 1. Financial variables

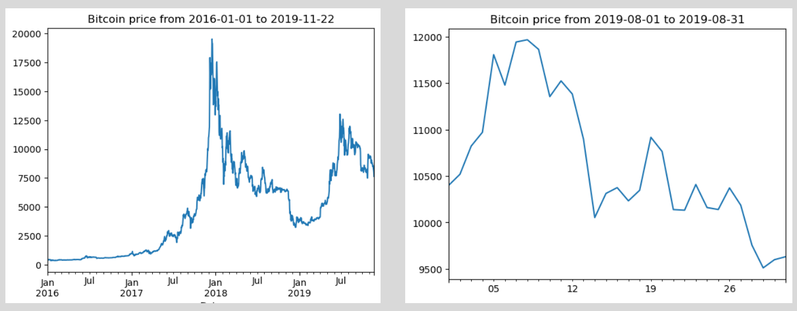


Chart 1. Bitcoin price changes (absolute value)

2. Data

Get data: Use *yfinance* to download over 3-year (2016/01/01-2019/11/22) financial data from *Yahoo! Finance*.

Process data: Create binary movement labels for each variable

- Price movement compared to tomorrow (future movement, prediction goal)
- Price movement compared to yesterday (lagged information/memory)

→ Rather balanced label: *btc_change_tmr*

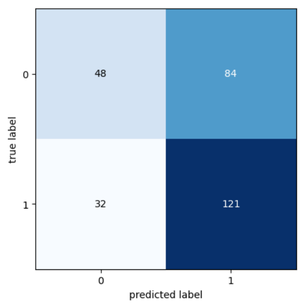
1: 793
0: 628

3. CO-MOVEMENTS

If we *know* how other financial variables will move tomorrow,

NOT THE REALITY

```
Accuracy :  0.5929824561403508
Precision :  0.5902439024390244
F1 Score :  0.6759776536312848
Recall :  0.7908496732026143
ROC AUC :  0.577243018419489
```



- Correlation

  btc_change_tmr highly correlates with *ethereum_change_tmr* (most important feature)

	btc_change_tmr
btc_change_tmr	1.000000
oil_change_tmr	0.008005
ethereum_change_tmr	0.314460
euro_change_tmr	-0.024928
gold_change_tmr	0.013059
tether_change_tmr	-0.013809
dow_change_tmr	-0.022534
sp_change_tmr	-0.005359
nasdaq_change_tmr	-0.008124
tenyear_change_tmr	-0.034045

Table 2. Correlation between concurrent movements

	importance	feature
1	0.370993	ethereum_change_tmr
4	0.101108	tether_change_tmr
3	0.094433	gold_change_tmr
8	0.079969	tenyear_change_tmr
0	0.077716	oil_change_tmr
2	0.073266	euro_change_tmr
7	0.071994	nasdaq_change_tmr
5	0.070145	dow_change_tmr
6	0.060375	sp_change_tmr

Table 3. Feature importance in the RF model

- Dropping *ethereum_change_tmr* , accuracy drops to 0.49 (worse than random guessing!)

- *Notes: If include longer data coverage (2016-2022), the accuracy of RF will increase to 0.77 due to 0.6 correlation between btc_change_tmr and ethereum_change_tmr.*

4. PREDICTION WITH ONE DAY LAG

Label: *btc_change_tmr*

Features: all variable change (including Bitcoin) compared to yesterday

Classification algorithms

- Random forest
  - Tuned RF [Best]
- Decision Tree
- KNN
- XGBoost

  eXtreme Gradient Boosting

Results: Better than random guessing (Table 4)

Model	Accuracy	Precision	F1_Score	Recall	AUC
Random Forest	0.557895	0.555556	0.681818	0.882353	0.532086
Tuned RF	0.561404	0.561404	0.671916	0.836601	0.539513
Decision Tree	0.529825	0.550802	0.605882	0.673203	0.518419
KNN	0.487719	0.518919	0.568047	0.627451	0.476604
XGBoost	0.536842	0.554974	0.616279	0.692810	0.524436

Table 4. Model performance comparison

5. Sentiment analysis

Why: Sentiment might matter

2021/01/19 Elon Musk placed #Bitcoin in his Twitter profile, tweeting "In retrospect, it was inevitable"
→ Bitcoin price briefly rose about $5000 in an hour

How:

Computationally determining whether a piece of writing is positive, negative, or neutral

- *"Bitcoin is the ruler of the world!"* 👍→ estimated sentiment score: 0.96
- *"I'm certain that Bitcoin will fail."* 👎→ - 0.999

INPUT
a dataset containing daily tweets from 2016 to 2019

VADER

Daily Sentiment Scores (Chart 2)

Valence Aware Dictionary and sentiment Reasoner

- a pre-trained sentiment analyzer from Python's *nltk.sentiment* package
- a lexical database and rule-based sentiment analysis tool that is optimized for **social media sentiments**

Prediction with sentiment Score as an additional feature

- Didn't see improvement (Table 5)

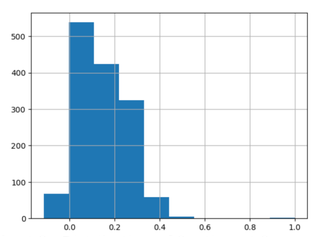


Chart 2. Distribution of Sentiment Scores

Model	Accuracy	Precision	F1_Score	Recall	AUC
Random Forest	0.512281	0.538793	0.642674	0.796178	0.480120
XGBoost	0.512281	0.551136	0.582583	0.617834	0.500323
Best RF	0.512281	0.542857	0.621253	0.726115	0.488057

Table 5. Model performance comparison, with sentiment as a feature

6. Conclusions

- The world of Bitcoin is highly volatile, not only its prices → Recent collapse of FTX

- It co-moves with other financial variables concurrently → Using nine other variable movement results in relatively accurate prediction, especially when we cover a longer time range.
  - But mostly due to its high correlation with another Cryptocurrency movement: Ethereum.
- It is hard to predict its movement direction using lagged movement information, with evaluation metrics slightly over 50% (~ random guessing?).
  - This is expected in financial analysis (+ Bitcoin volatility), otherwise it would be so easy to become millionaires.
  - Also possible we didn't capture what really affects Bitcoin price movement (other than market fluctuations / sentiment).
- We attempted to incorporate sentiment into our prediction, but the model performance didn't improve much.
  - Possible explanations
    - Doubtful link between sentiment affecting bitcoin price
    - Sentiment score from Tweets as an incomplete proxy for sentiment

UNIVERSITY OF TARTU
Institute of Computer Science