Predicting Cryptocurrency Prices using Twitter Sentiment Analysis

Author: Ricardo Sans Cipollitti
Supervisor: David Amorós Alcaraz
MSc in Computational Engineering and Mathematics
Rovira i Virgili – Universitat Oberta de Catalunya

Index

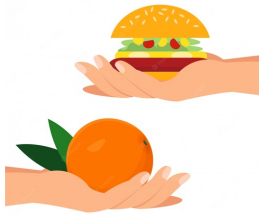
1. Introduction

The main goal of this project is to predict cryptocurrency prices using tweets

In this context, we need to understand why do **cryptocurrencies** exist in the first place…

1. Introduction

Bartering



Precious metals

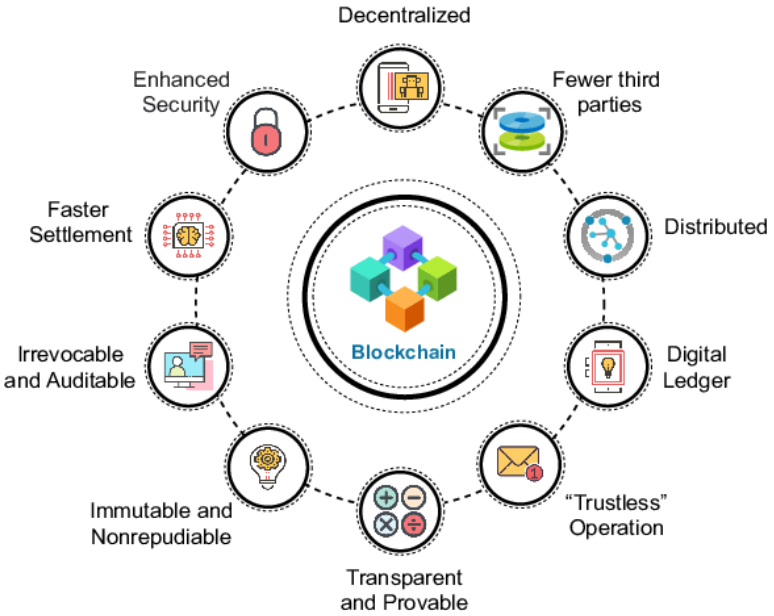
Gold Standard and fiat money

Cryptocurrencies

2. Understanding the technology

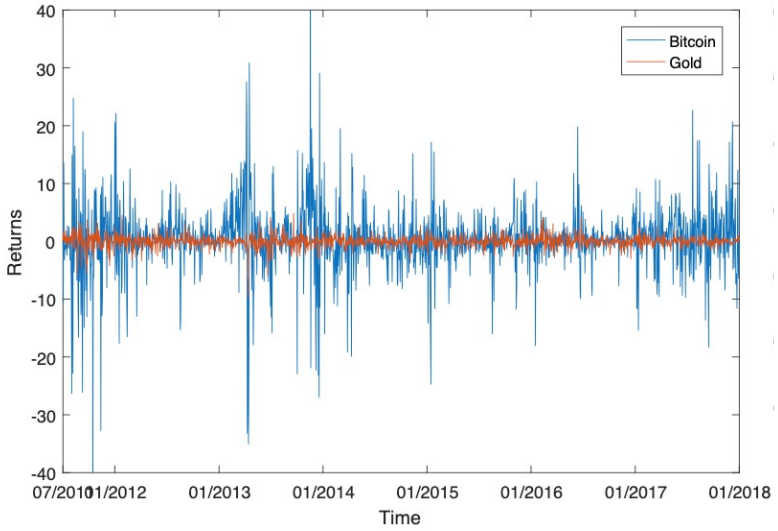
Blockchain is a distributed ledger with the following features:

- Everyone is allowed to write transactions
- Cryptography secures the network and transactions
- No third party needed
- Immutable
- Based on consensus



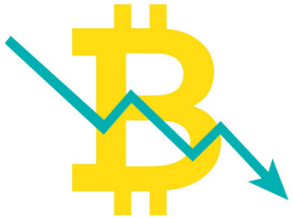
3. Price and volatility of cryptocurrencies

Bitcoin vs. gold daily price returns



There are several drivers that explain this variability:

- Economic
- Transactional
- Technical
- Sentimental



4. Overview of the model

Introduction and gathering of data

- We needed to collect several information from different platforms:
  - ✓ cryptocurrency USDT pair prices
  - ✓ tweet's texts

Gathered from these platforms:

CryptoArchive

Twitter API

- Data was processed using:

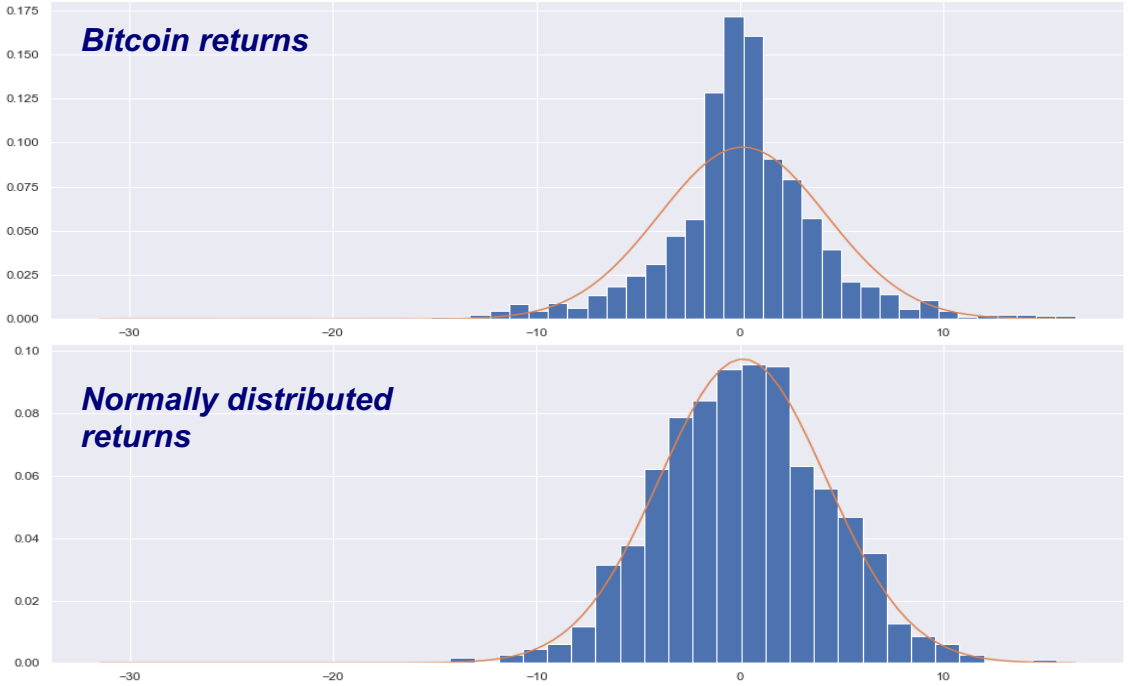
python™

PostgreSQL

jupyter

4. Overview of the model

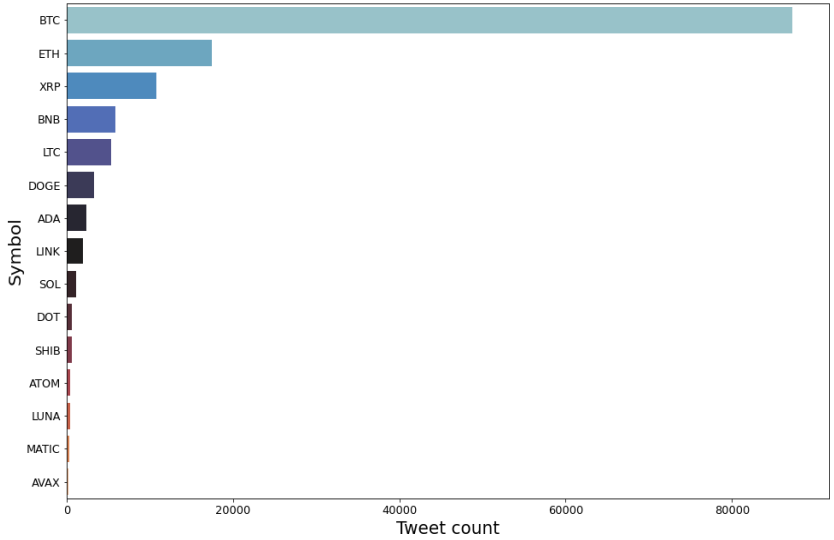
Understanding the data: Crypto price analysis

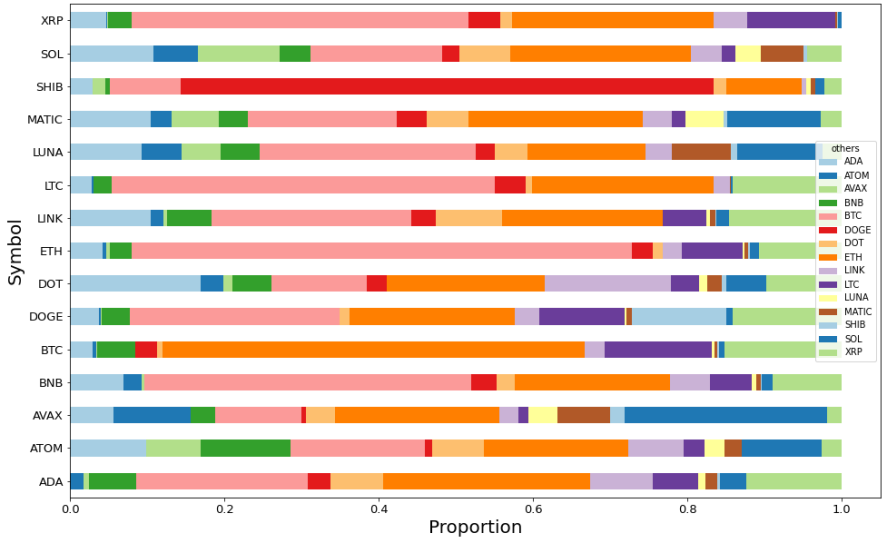


	Test statistic	p-value
x:	1.02	0.3100
BTC:	13.80	0.0000

. Overview of the model

Understanding the data: Tweet analysis





4. Overview of the model

Important considerations

Our limitations and assumptions to this project have been the following:

- Predict price movements
- No time period segmentation
- Unsupervised learning
- Tweets from crypto gurus
- High % of 0 sentiment scores
- No spread nor commissions are assumed

influencer	username	influencer	username
Vitalik Buterin	@VitalikButerin	Coin Bureau	@coinbureau
Roger Ver	@rogerkver	BoxMining	@boxmining
Andreas M. Antonopoulos	@aantop	Lark Davis	@TheCryptoLark
Tim Draper	@TimDraper	BlockchainLeaks	@LeaksBlockchain
Charlie Lee	@SatoshiLite	CryptoLove	@TheCryptoLove
Anthony Pompliano	@APompliano	Aimstone	@Aimstone5
Erik Voorhees	@ErikVoorhees	Hashoshi	@hashoshi4
Tone Vays	@ToneVays	Philakone	@PhilakoneCrypto
John McAfee	@officialmcafee	Cryptonauts	@CryptonautsShow
Ivan on Tech	@IvanOnTech	Jason Pizzino	@jasonpizzino
CryptoBrekkie	@BVBTC	Andreas Antonopoulos	@aantonop
Dan Held	@danheld	Roger Ver	@rogerkver
Layah Heilpern	@LayahHeilpern	Nick Szabo	@NickSzabo4
Kenn Bosak	@KennethBosak	CryptoCred	@CryptoCred
Ben Horowitz	@bhorowitz	Erik Voorhees	@ErikVoorhees
Elon Musk	@elonmusk	PlanB	@100trillionUSD
Ty Smith	@TyDanielSmith	Brian Armstrong	@brian_armstrong
CryptoWendyO	@CryptoWendyO	Loomdart	@loomdart
Euclid and Oaks	@EuclidAndOaks	Naval	@naval
David Gokhshtein	@davidgokhshtein	Credible Crypto	@CredibleCrypto
Hailey Lennon	@HaileyLennonBTC	Josh Olszewicz	@CarpeNoctom
Justin Sun	@justinsuntron	Marty Bent	@MartyBent
Ivan on Tech	@IvanOnTech	Tim Draper	@TimDraper
LayahHeilpern	@LayahHeilpern	Documenting Bitcoin	@DocumentingBTC
Coinbound	@coinboundio	Adam Back	@adam3us
Sheldon Evans	@SheldonEvans	Messari	@MessariCrypto
CryptoBusy	@CryptoBusy	Nick Szabo	@NickSzabo4
JRNY Crypto	@JRNYcrypto	Cred	@CryptoCred
BitBoy Crypto	@Bitboy_Crypto	Changpeng Zhao	@cz_binance
Whale Panda	@WhalePanda	Gavin Andresen	@gavinandresen
Camila Russo	@CamiRusso	Balaji Srinivasan	@balajis
Nicholas Merten	@Nicholas_Merten	The Wolf Of All Streets	@scottmelker

4. Overview of the model

Explanation of the models

Our algorithm included several Machine Learning models. These are:

- Logistic Regression
- Support Vector Machines
- K-Nearest Neighbor
- Naïve Bayes
- Decision Tree Classifier
- Random Forest Classifier

Two scenarios are assumed:

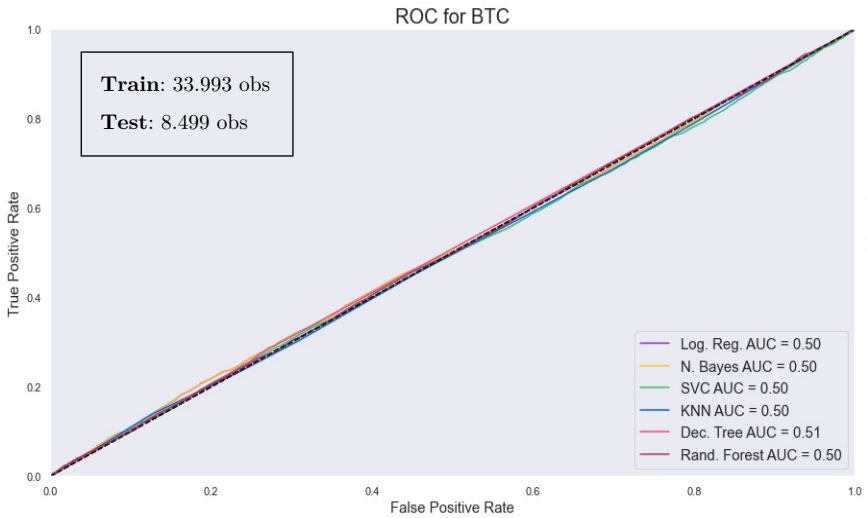
	Scenario A	Scenario B
Features	Sentiment score	Sentiment score and number of retweets
Target	Price movement	Price movement

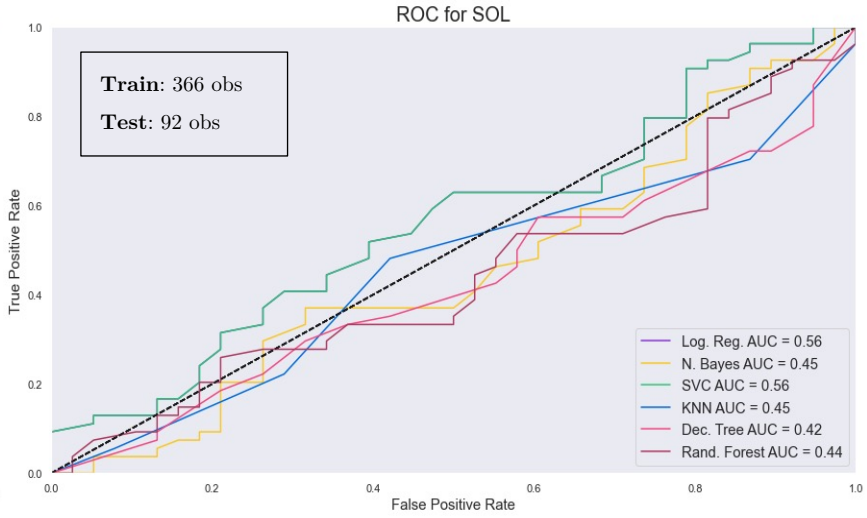
Extensions

1	2	5	- minute delays

4. Overview of the model

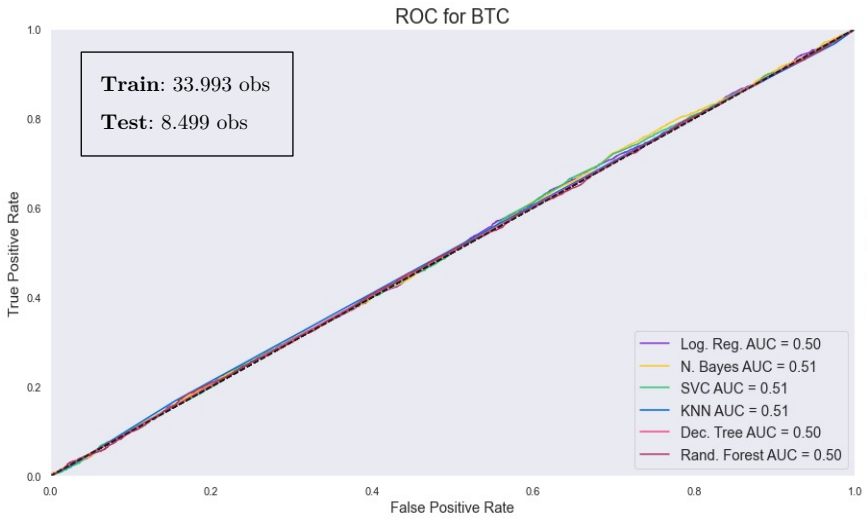
Results: scenario A

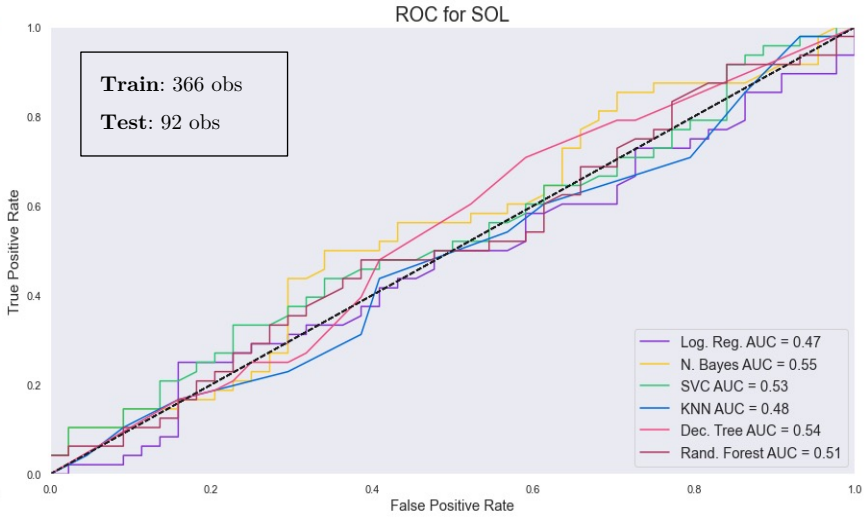




4. Overview of the model

Results: scenario B





4. Overview of the model

Results: extensions

Delay: 1 minute						
Crypto	Logistic regression	Naive Bayes	SVC	KNN	Decision tree	Random Forest
BTC	49,8%	49,8%	50,2%	50,0%	50,4%	50,4%
ETH	46,6%	53,4%	46,6%	50,1%	51,3%	51,1%
BNB	52,4%	52,4%	52,4%	52,8%	52,4%	51,5%
LTC	51,4%	50,7%	51,4%	55,9%	50,1%	53,0%
ADA	52,5%	50,8%	50,0%	53,2%	49,3%	51,4%
XRP	51,2%	48,4%	50,0%	50,2%	49,1%	49,9%
LINK	48,8%	48,5%	50,5%	47,7%	49,5%	52,7%
MATIC	46,8%	46,8%	46,8%	41,5%	45,8%	37,6%
ATOM	56,5%	56,5%	42,0%	42,9%	56,2%	48,3%
DOGE	54,7%	54,7%	50,0%	51,4%	53,2%	50,9%
DOT	55,5%	52,6%	55,5%	54,2%	47,4%	37,6%
SOL	56,5%	50,0%	41,6%	50,6%	48,3%	50,3%
LUNA	57,1%	57,1%	42,9%	59,4%	53,6%	54,8%
AVAX	50,0%	35,8%	41,7%	39,6%	31,7%	38,3%
SHIB	56,5%	56,5%	43,9%	46,6%	42,3%	48,1%

Delay: 2 minutes						
Crypto	Logistic regression	Naive Bayes	SVC	KNN	Decision tree	Random Forest
BTC	49,7%	49,7%	50,0%	51,0%	49,3%	49,7%
ETH	49,4%	49,4%	50,6%	51,5%	49,9%	50,2%
BNB	49,8%	49,8%	50,2%	52,0%	50,3%	51,8%
LTC	52,9%	47,1%	52,9%	47,6%	50,1%	45,7%
ADA	46,6%	54,1%	53,4%	58,5%	50,3%	58,4%
XRP	51,0%	51,0%	47,2%	50,9%	51,2%	51,7%
LINK	52,5%	47,5%	47,5%	53,7%	50,6%	50,6%
MATIC	50,4%	49,6%	49,6%	42,5%	50,0%	54,0%
ATOM	60,8%	40,4%	29,0%	61,2%	50,0%	43,7%
DOGE	59,1%	59,1%	40,9%	46,1%	44,4%	44,2%
DOT	39,9%	55,3%	39,9%	45,2%	51,7%	59,3%
SOL	46,0%	54,0%	46,0%	50,7%	53,6%	54,5%
LUNA	41,9%	58,1%	41,9%	61,9%	61,7%	86,4%
AVAX	45,4%	54,6%	54,6%	76,2%	49,2%	48,8%
SHIB	32,4%	41,0%	67,6%	46,3%	56,9%	56,3%

4. Overview of the model

Results: extensions

Delay: 5 minutes						
Crypto	Logistic regression	Naive Bayes	SVC	KNN	Decision tree	Random Forest
BTC	49,3%	49,5%	49,3%	49,9%	50,0%	49,6%
ETH	49,7%	49,7%	49,7%	47,0%	50,3%	49,7%
BNB	53,5%	46,5%	46,5%	50,9%	50,4%	50,2%
LTC	48,6%	50,7%	51,4%	49,2%	50,5%	49,0%
ADA	46,9%	50,0%	46,9%	50,7%	50,1%	48,7%
XRP	48,9%	49,2%	48,9%	50,6%	50,9%	50,4%
LINK	49,5%	49,5%	50,7%	52,6%	51,8%	52,5%
MATIC	52,2%	46,3%	47,8%	55,0%	50,3%	47,4%
ATOM	49,3%	56,5%	49,3%	55,7%	50,0%	53,1%
DOGE	51,0%	51,0%	51,0%	53,2%	52,7%	54,9%
DOT	46,6%	46,6%	46,6%	44,8%	49,4%	47,6%
SOL	50,0%	50,0%	45,0%	44,8%	48,1%	45,4%
LUNA	28,3%	28,3%	71,7%	30,0%	41,4%	35,8%
AVAX	47,4%	52,6%	52,6%	40,6%	47,0%	35,5%
SHIB	53,8%	46,2%	46,0%	47,1%	53,4%	49,6%

5. *Conclusions & outlook*

For **scenario A**:

- 0 – minutes delay
- More than 0 minutes delay

For **scenario B**:

- 0 – minutes delay

Further **considerations** should be considered, such as:

- Increase number of influencers
- ML model for relating tweets to crypto
- Further studies for cryptos with few observations