NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

School of Computer Science and Engineering
College of Engineering

Time-Series & Sentiment Analysis of Top Cryptocurrencies

Student: Chester Yeoh Fu Soon Supervisor: Assoc Prof Anwitaman Datta

Introduction

Bitcoin (BTC), the world's first cryptocurrency, was created in 2009 by Satoshi Nakamato. Powered by a blockchain framework with no central authority or middlemen, BTC has seen a meteoric rise in its popularity and now exceeds 300,000 transactions per day on top of a USD 1Tn market capitalisation (as of 9 March 2021). The benefits of cryptocurrencies such as accessibility and autonomy (Reiff,2021) are well documented. However, most cryptocurrencies suffer from a key drawback- high price fluctuations. Bitcoin, for example, increased its value by 1900% in 2017 before plunging below USD 8,000 per coin from a high of USD 19,000 in the 2018 Jan/Feb Crash. Since cryptocurrency price research is still a unexplored area and equilibrium market prices are poorly understood (Conrad, 2018), this project seeks to conduct a sentiment analysis to explore possible drivers such as social media and news reports that leads to these price fluctuations. Additionally, other non-sentiment factors such as Public Interest and Developer Data will also be analysed for their possible price impacts.

Objectives

Objective 1

Explore correlation between the price of cryptocurrencies and social media/news sentiment.

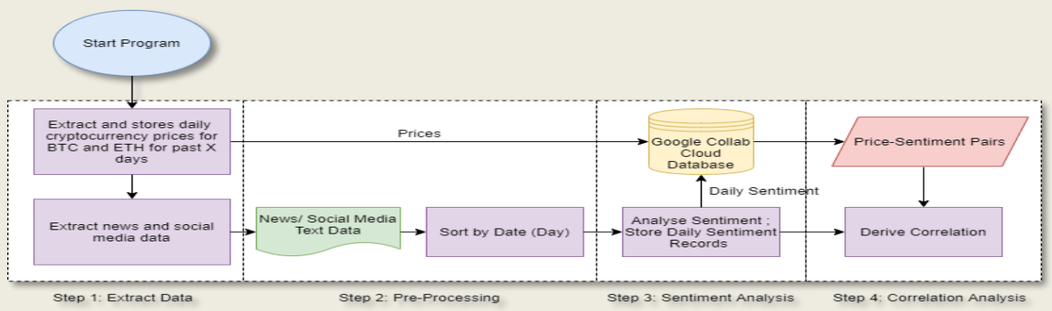
Objective 2

Examine non-sentiment related price drivers to create regression (Predict Daily BTC Price) & classification (Predict Buy, Sell, Hold decisions) models.

Methodology

Overview

1. Extraction of Data
    1. Sentiment Data from Reddit, Twitter and News Sources
    2. Daily Price Data of Ethereum (ETH) and BTC
    3. Non-Sentiment Data such as ***Alexa Search Rankings***
2. Pre-process data by removing nulls and grouping on dates
3. Aggregate Daily Sentiment using a Valence Aware Dictionary for Sentiment Reasoning (VADER) analyser
4. Calculate Correlation using Pearson's Correlation Coefficient
5. Use 1.3 as inputs for regression/classification models (Objective 2 Only)



Results of Sentiment Analysis

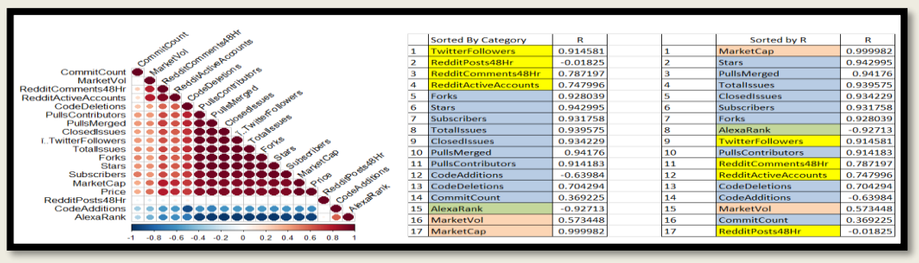
There are **no** correlations between the price of BTC and social media/news sentiment. (|r|<0.1).
There are **weak** correlations between the price of ETH and social media/news sentiment (0.1< |r|<0.3)
A lead-lag analysis of 2 days increased the correlation coefficient in certain scenarios (i.e. |r|=0.269 when a 2 day lead on ETH price was introduced on Reddit sentiment data). However, there were still no situations in which at least a **medium** (0.3<|r|<0.5) correlation was observed.

	BTC Norm	Eth Norm
BTC Norm	1	
Eth Norm	0.701040026	1
News Norm	-0.04932699	0.132783105
Reddit Norm	0.025630669	0.143285397
Twitter Norm	-0.04963587	-0.12541159

	BTC	BTC Lead 1	BTC Lead 2	BTC Lag 1	BTC Lag2
News Norm	-0.0493	-0.0487	-0.1036	-0.0283	0.0280
Reddit Norm	0.0256	0.0014	0.0371	0.1362	0.1225
Twitter Norm	-0.0496	-0.0039	0.1904	-0.0449	0.0303
	ETH	ETH Lead 1	ETH Lead 2	ETH Lag 1	ETH Lag 2
News Norm	0.13278	0.12404	0.10065	0.00081	-0.01306
Reddit Norm	0.14329	0.22211	0.26924	0.18212	0.17572
Twitter Norm	-0.12541	-0.03347	0.11706	-0.22194	-0.03386

Results of Non-Sentiment Related Factor Analysis

17 other Non-Sentiment related factors were examined for their correlations to BTC prices and it was discovered that 13 of them had **strong** price correlations (with |r|>0.7).



	Sorted By Category	R
1	TwitterFollowers	0.914581
2	RedditPosts48Hr	-0.01825
3	RedditComments48Hr	0.787197
4	RedditActiveAccounts	0.747996
5	Forks	0.928039
6	Stars	0.942995
7	Subscribers	0.931758
8	TotalIssues	0.939575
9	ClosedIssues	0.934229
10	PullsMerged	0.94176
11	PullsContributors	0.914183
12	CodeAdditions	-0.63984
13	CodeDeletions	0.704294
14	CommitCount	0.369225
15	AlexaRank	-0.92713
16	MarketVol	0.573448
17	MarketCap	0.999982

	Sorted by R	R
1	MarketCap	0.999982
2	Stars	0.942995
3	PullsMerged	0.94176
4	TotalIssues	0.939575
5	ClosedIssues	0.934229
6	Subscribers	0.931758
7	Forks	0.928039
8	AlexaRank	-0.92713
9	TwitterFollowers	0.914581
10	PullsContributors	0.914183
11	RedditComments48Hr	0.787197
12	RedditActiveAccounts	0.747996
13	CodeDeletions	0.704294
14	CodeAdditions	-0.63984
15	MarketVol	0.573448
16	CommitCount	0.369225
17	RedditPosts48Hr	-0.01825

The predictive models built using these 17 factors had the following results:

- Linear Regression (LR): Test Set RMSE of 538, Adjusted R-Squared of 0.9497
- Random Forest (RF) Classifier: Test Set Confusion Matrix Accuracy of 90%
- Naïve Bayes (NB) Classifier: Test Set Confusion Matrix Accuracy of 60%

Important and significant features highlighted by the models include ***Number of Pull Contributors*** on a cryptocurrency's GitHub Repository(LR) and ***Number of Active Reddit Accounts*** on a cryptocurrency's subreddit (RF).

Discussion

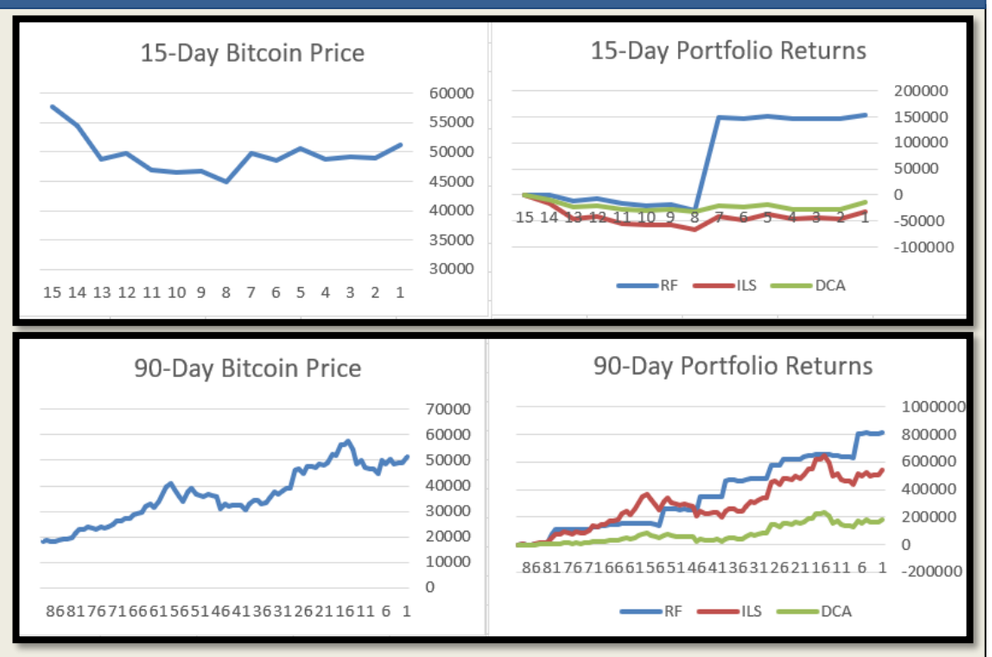
The findings of this FYP indicate that news/social media data are poor price drivers of cryptocurrencies like BTC and ETH. This is due to some of its limitations – which includes an inability to detect modern slang and contextual words as well as the presence of noise (Promotions and Ads) in the data extraction phase which cannot be fully eliminated. On the other hand, non-sentiment factors proved to be strong drivers of cryptocurrency prices with almost perfect correlations and high modelling accuracies.

Business Implication & Future Work

To test for the potential real-world application for the RF classifier, a portfolio stimulation of 3 different investing strategies was carried out:

Name of Strategy	Investment Methods
Random Forest (RF)	"Buy" Prediction: Investor will invest $100,000 into buying BTC "Sell" Prediction: Investor will liquidate half of their BTC holdings into cash "Hold" Prediction: No action taken by investor
Initial Lump Sum (ILS)	Investor invests $300,000 on day 0 into buying BTC.
Dollar Cost Averaging (DCA)	Every 10 days, the investor will invest a fixed X amount into buying BTC. • For 90-day stimulation, X=300,000/ (90/10=9 periods) = $33,333.33 • For 15-day stimulation, X=300,000/ (15/10= 2 periods) = $150,000

In both 15-day and 90-day stimulations, the RF strategy yielded the highest profits of $154,141 and $812,158 respectively. The results further indicated two key advantages of the RF model- the ability to make profits in a downward BTC price trend as well as providing a systemic method to sell BTC to realise profits. ($655,208 realised profits vs 0 for ILS/DCA).



Future Work

With promising results obtained in the 15-day and 90-day stimulation, a future direction would be to include other price drivers in the developed RF model with the hopes of improving model accuracy. Another aspect that can be worked on further is the streamlining of end-to-end processes used in model development and prediction since it is currently split across multiple platforms due to the constraints of platform-specific APIs and libraries. Upon finetuning the RF model both in terms of accuracy and efficiency, it can be benchmarked against more advanced investing strategies as potential research topics to further extend the work of this FYP