MGT 6203 Group Project Proposal

TEAM INFORMATION

Team #: 10

Team Members

1. Chris Messer; Edx username: chrislmesser

I am an analytics consultant at a big 4 consulting firm. My day to day work is varied across analytics and software development projects. My projects primarily relate to helping clients understand their data needs, and helping them get the data where they need it. I have an undergraduate degree in finance, and a graduate degree in accounting. I have completed 6501 and am in progress completing 6040 and 6203. Past analytical projects include overhauling employee evaluations to take a data driven approach.

2. Erin Abbott; Edx username: enaena44

I currently work as a business consultant for a supermarket chain where my role is to maintain a financial dataset and support all reporting that utilizes that data. My undergraduate degree was in industrial engineering from Georgia Tech with a minor in computer science, and I have taken two other edX courses as part of the micromasters in analytics (CS6040 and ISYE6501). In my current role, I often work on analytics projects involving bad data detection and correction where machine learning models are used to detect the bad data and various statistical models are employed to impute those data points.

3. Devyn Byrd; Edx username: DevynPByrd42

I am a computer science graduate student at Clemson University with an undergraduate degree in audio technology. I worked as an audio intern while at Clemson, leading recording productions for student ensembles and surveying students on preferred software implementations. I completed Google's Data Analytics Professional Certificate, which involved an analytics/marketing case study on women's health and wellness technology.

OBJECTIVE/PROBLEM

Project Title: Bitcoin and Sentiment Analysis

Project Background

Bitcoin is a cryptocurrency, a type of virtual currency utilizing encryption tools to facilitate secure digital transactions on the blockchain. Bitcoin does not have inherent worth and is instead based on supply and demand due to its decentralized nature. Although cryptocurrencies have been thoroughly studied recently, cryptocurrency is still a relatively new concept, and as a result, the

general public's perception of the technology is divided. The value of Bitcoin has endured drastic peaks and troughs throughout its brief history.

Sentiment analysis is a field of computing that utilizes programming to understand and process human emotions. The goal of such programming is often to train artificial intelligence on human interactions to improve "emotional intelligence" in AI. By codifying human feelings, businesses can better understand consumer responses to improve their data-driven decision making.

Farell, Ryan, "An Analysis of the Cryptocurrency Industry" (2015). Wharton Research Scholars. 130. https://repository.upenn.edu/wharton_research_scholars/130

Cambria, E., Das, D., Bandyopadhyay, S., & Feraco, A. (Eds.). (2017). Chapter 1: Affective Computing and Sentiment Analysis. In *A Practical Guide to Sentiment Analysis* (pp. 1–2). essay, Springer.

Problem Statement

This analysis seeks to understand the correlation between the sentiment score of tweets about bitcoin and the price of bitcoin. In particular, this analysis will investigate the impact of news and other factors on the price of bitcoin and determine if there is a significant correlation between the sentiment score of tweets and other sources about bitcoin and the price of bitcoin.

Research Questions (RQs)

1. (Primary) Is there a correlation between sentiment of public perception of bitcoin, and the price of bitcoin?
2. What source of public perception has the strongest correlation?
3. What kind of sentiment has the strongest correlation with price?

Business Justification

By understanding the correlation between the sentiment score of tweets about bitcoin and the price of bitcoin, investors and traders can make more informed decisions when it comes to investing in bitcoin. Additionally, this analysis can provide insight into the impact of news and other factors on the price of bitcoin, helping investors better understand the potential direction of asset prices.

DATASET/PLAN FOR DATA

Data Sources

1. Bitcoin 17.7 million Tweets and price | Kaggle
2. Reddit Comments Containing "Bitcoin" 2009 to 2019 | Kaggle

Data Description

1. Bitcoin 17.7 million Tweets and price: This data source contains information captured hourly on the cost and volume of bitcoin as well as details on the tweets mentioning bitcoin during that hour.

Date	Compound_ Score	Total Volume of Tweets	Count_ Negatives	Count_ Positives	Count_ Neutrals	Sent_Negatives	Sent_Positives	Count_ News	Count_ Bots	Open	High	Low	Close	Volume (BTC)	Volume (Currency)
8/1/17 0:00	0.08289332	1027	139	347	397	-0.516620144	0.417881844	695	144	2855.81	2863.06	2823	2825.92	184.02	522951,00
8/1/17 1:00	0.05315978	778	111	193	320	-0.423401802	0.415384974	493	154	2823.01	2860.02	2821.01	2853.38	77.3	219605.16
8/1/17 2:00	0.1242508	836	89	273	264	-0.440938202	0.428661172	510	210	2846.27	2858.04	2837.31	2841.6	135.83	386739.15
8/1/17 3:00	-0.0210365	984	250	236	314	-0.4411732	0.396034322	683	184	2841.84	2863.88	2837.73	2862.93	143.2	408360.03
8/1/17 4:00	0.05543746	751	114	195	305	-0.41497807	0.417159487	517	137	2862.92	2876	2848.11	2874.99	222.53	637045.88
8/1/17 5:00	0.09344372	842	120	286	303	-0.424965833	0.409956294	563	133	2872.03	2929.17	2865.01	2900	415.97	1208323.94

(Header rows from dataset "Bitcoin 17.7 million Tweets and price)

2. Reddit Comments Containing "Bitcoin" 2009 to 2019: This dataset has reddit comments across all subreddits that mention bitcoin. This dataset is not tagged for sentiment, so we will need to do some preprocessing ourselves.

```
btred.head()
```

	Unnamed: 0	datetime	date	author	subreddit	created_utc	score	controversiality	body
0	0	2014-06-26 16:49:41	2014-06-26	HonorConnor	AskReddit	1.403801e+09	162.0	0.0	How do you feel about Bitcoin? I don't really ...
1	1	2014-05-17 23:30:36	2014-05-17	Talvoren	AskReddit	1.400369e+09	75.0	0.0	[The guy who blew a huge portion of his and hi...
2	2	2014-12-18 04:54:48	2014-12-18	ninjoe87	Outdoors	1.418878e+09	-4.0	0.0	Because you are a damn liar.\n\nFrom my other ...
3	3	2014-04-30 00:24:16	2014-04-30	fathergrigori54	tf2	1.398817e+09	13.0	0.0	This was the bitcoin hat guy that ACTUALLY del...
4	4	2014-07-07 16:06:21	2014-07-07	Draber-Bien	starcraft	1.404749e+09	19.0	0.0	I'm sure this is "good" for bitcoin

(Header rows for dataset Reddit Comments Containing "Bitcoin" 2009 to 2019)

Key Variables

The two datasets are unique in that the Twitter Dataset already is tagged for sentiment and aggregated at an hourly level. Also included in this dataset is price and trading volume for bitcoin at each of the hourly increments.

As such, there will be a few existing independent variables we will want to use from the twitter dataset, and also a few new variables we will need to generate out of the raw reddit comment data.

Lastly, the pricing data provided in the twitter dataset is only provided at an hourly basis. Our analysis will also observe the correlation of sentiment at different time horizons, i.e. at the per day level. As such, we will need to bring in independent pricing data for the open/close/volume from an independent source or extract it from the twitter data provided.

- Independent:
    - Twitter
        - Existing Variables
            - Tweet volume per hour
            - Count of Tweets with negative sentiment
            - Count of Tweets with positive sentiment

                - Tweets with neutral sentiment
        - Reddit
            - New Variables
                - Count of comments per hour
                - Count of negative, positive, neutral comments per hour
        - Pricing Data
            - Existing Variables
                - open/close at hour intervals
                - open/close at day intervals
                - Price in previous intervals
- Dependent
    - Bitcoin Price

APPROACH/METHODOLOGY

Approach

This analysis will examine the correlation of bitcoin price with the sentiment of public discourse around bitcoin. We will evaluate the correlation on two general regression models.

Linear Regression Model: We will use a general linear regression model that is fit on the dependent variables mentioned above to predict the price of Bitcoin at a given hour.

Logistic Regression Model: A logistic regression model will be used to determine the odds that the price of bitcoin will increase or decrease in the next time period.

Transformations

Time lag: Since we are using past data to predict a future price, each model will require transforming the dependent variable time series by t-1 to predict the price of bitcoin at time t.

Time aggregation: The twitter data is provided at a per hour interval basis, and the reddit data is provided at a per occurrence aggregation. As such, some transformations will need to take place in order to aggregate the data at the appropriate level.

Model Comparison

For both models, we will separate our data into train, test, and validation sets.

For the linear regression model, we will alternate the variables included in order to find the best model, and then will compare the results of the testing data using the r^2 value.

For the logistic regression model, we will vary the variables included in the model as well as the cutoff threshold for considering what is an up signal vs. a down signal. I.e. if there is a 20% chance or higher that the price will move down in the next hour, consider it a down signal, otherwise consider it an up signal. We will then compare the AUC against the various approaches and determine the best model.

Anticipated Hypothesis/Conclusions and Business Decision Impact

We expect to see a positive correlation between the sentiment of public discourse and the price direction of bitcoin.

As a result of this analysis, we will determine whether public sentiment about bitcoin is useful information when developing trading strategies. This could result in increased profits for a data-driven trader.

PROJECT TIMELINE/PLANNING

Date	Milestone
3/25	Complete 1-2 Regression Analyses
3/29	Progress Report
4/1	Progress Report Video
4/8	Complete Remaining Regressions
4/15	Final Report
4/18	Final Presentation

Appendix (any preliminary figures or charts that you would like to include):