Political tweets: Milestone 3

2018-11-20

Group 6: Martin Jin, Qimei Lin, Yingxue Lu, Yuhan Zhang

Introduction

The goal of our project is to explore the relation between Trump's tweets and performance of U.S. financial system. The reason that propels us to conduct this project is that due to the advancement of technology, especially the prevalence of mobile phones, social media has been playing an increasingly significant role in reflecting and influencing political and social events which in turn affect U.S. financial system. U.S. president Donald Trump's continuous tweeting about politics and society becomes harder to ignore when it comes to investment decisions[1]. We want to select the features of Trump's tweet that are more likely to make an impact and seek out specific financial market data that are highly likely to be impacted.

For this milestone, we are asking whether the S&P 500 index could be predicted by the number of tweets Trump produces, the number of tweets he retweets from others, the total number of retweets he gets for his tweets, and the total number of 'favorite' he received. We chose S&P 500 index because of two reasons: first, for a sporadic/abrupt tweet which could have sentiment effect on the market, due to the amplification of the social media and the information exchanges, it tends to affect more ordinary traders in a stock market, than, say, the seasoned traders for U.S. treasury bond or future or options; second, S&P 500 index reflects the performance of 500 stocks, whose potential association with Trump's tweet is more likely to reflect a real relation between his tweet and the whole stock market.

This is the first step we took to understand the phenomenon, which would inform us with other potential interesting questions along the way.

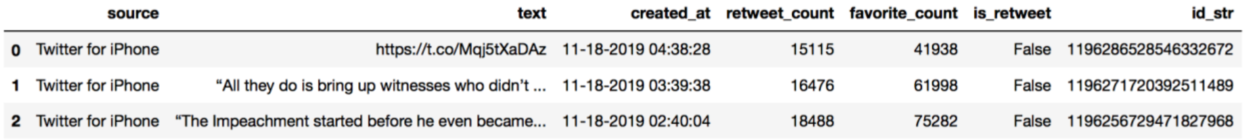
Data Acquisition

We downloaded Trump's tweets from *TrumpTweeterArchive* to prepare our predictor set. The response variable is the market price of S&P 500 index, which was obtained by the ImportData(). function (please see the notebook). Both datasets cover the period from 2017-01-20 to 2019-11-18.

Data Manipulation

The raw data from Trump's tweet look like the table below.

Table 1 Raw data

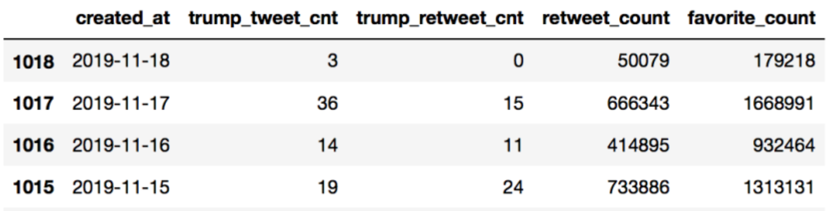


	source	text	created_at	retweet_count	favorite_count	is_retweet	id_str
0	Twitter for iPhone	https://t.co/Mqj5tXaDAz	11-18-2019 04:38:28	15115	41938	False	1196286528546332672
1	Twitter for iPhone	"All they do is bring up witnesses who didn't ...	11-18-2019 03:39:38	16476	61998	False	1196271720392511489
2	Twitter for iPhone	"The Impeachment started before he even became...	11-18-2019 02:40:04	18488	75282	False	1196256729471827968

[1] Refer to the article: 'The Volfefe Index, Wall Street's new way to measure the effects of Trump tweets, explained'

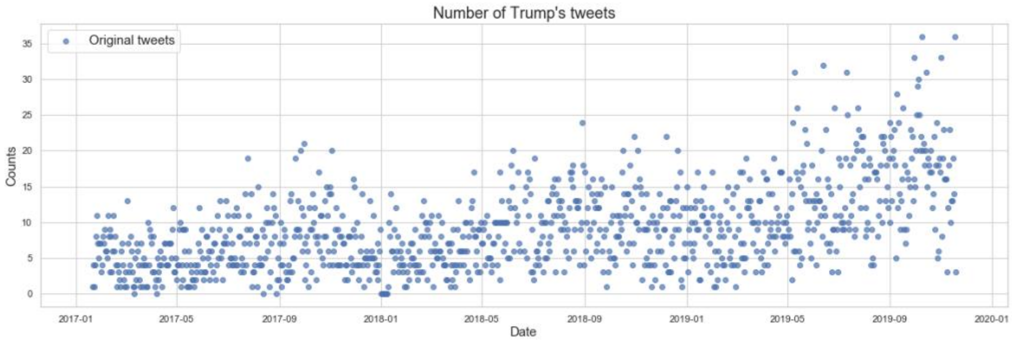
We aggregated the data so that each line represents the number of tweets/retweets/favorites for a single day (Table 2). We got our **three predictors**: (1) **‘trump_tweet_cnt’** means the number of original tweets Trump created per day; (2) **‘retweet_count’** means how many people retweeted Trump’s tweet, including his original tweets and the tweets he got from others; (3) **‘favorite_count’** means how many ‘favorites’ Trump received for his own tweets.

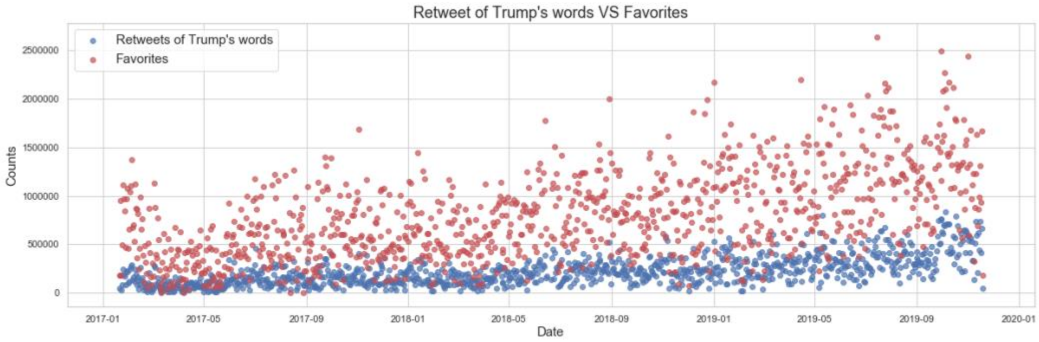
Table 2 Aggregated by Day



	created_at	trump_tweet_cnt	trump_retweet_cnt	retweet_count	favorite_count
1018	2019-11-18	3	0	50079	179218
1017	2019-11-17	36	15	666343	1668991
1016	2019-11-16	14	11	414895	932464
1015	2019-11-15	19	24	733886	1313131

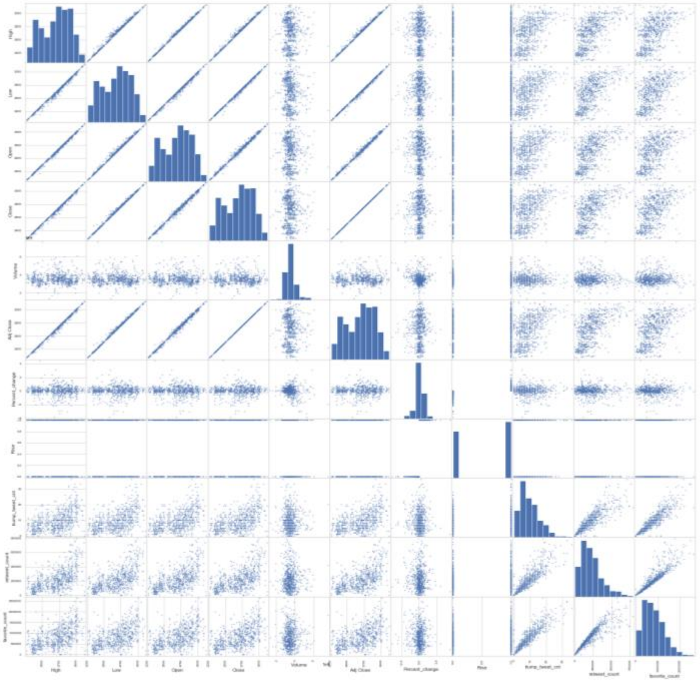
The distribution of the three predictors are shown below. We could see that since 2019, Trump has tweeted more frequently per day than previous years. So has the number of people who retweeted his words and who marked his words as ‘favorite’.





As for the response variable, S&P 500 has provided several variables for us to choose from. We made a paired scatter plot (see next page) of all these stock variables and the four predictors. We found that at first glance, **‘high’**, **‘low’**, **‘open’**, **‘close’** and **‘adj close’** have some linear relationship with **‘trump_tweet_cnt’**, **‘retweet_count’**, and **‘favorite_count’**. Other columns that are not directly reflecting the price, such as **‘volume’** and **‘percent_change’** seem to be unrelated to **‘trump_tweet_cnt’**, **‘retweet_count’**, or ‘**favorite_count**’.

After the speculation, we chose ‘high’ as our response variable, indicating the highest price during the day.



Basic modelling

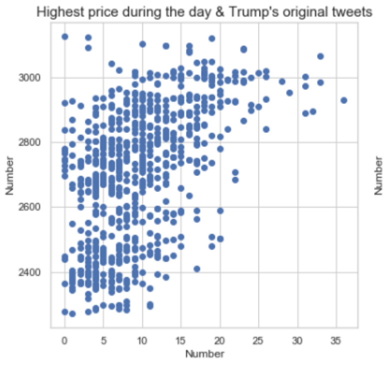
Because the response variable is continuous, we will adopt regression models.

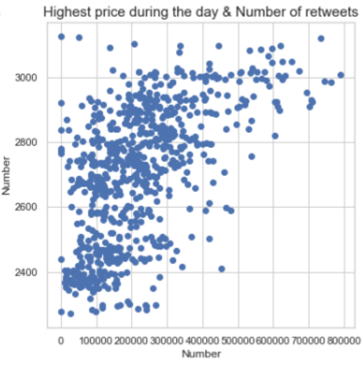
1. Simple Linear Regression

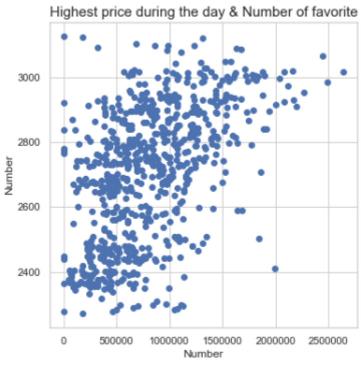
We performed a simple linear regression model using **‘trump_tweet_cnt’**, **‘retweet_count’**, and **‘favorite_count’** as predictors and **‘high’** as response variable and calculated the score. The score for the training set is 0.310 and the score for the test set is 0.340.

2. KNN

Then we performed a series of k-NN model using **‘trump_tweet_cnt’**, **‘retweet_count’**, and **‘favorite_count’** as predictors and **‘high’** as response variable and calculated the score. The best scores we get is when number of neighbors is 20. The training set score is 0.350. The test set score is 0.346.







We also visualized the relation between the response variable and each predictor. It seems that the distribution and shape of the response variable doesn’t change much across the three plots, indicating there is multicollinearity between the predictors, which explains the low score. Since these scores of both basic models are pretty bad, we will further process the data and choose other predictors and response variables for our model.

Future direction

1. The baseline model clearly gives terrible fitting. The failure is due to the lack of proper predictors. The number of tweets by Trump does not mean much to the market; rather, it's the content in those tweets that matters more. A simple modification on the current dataset is to have a column named ave_retweet or ave_favorite, whose value is `retweet_count`/`trump_tweet_cnt` or `favorite_count`/`trump_tweet_cnt`. This column would reflect the importance of the tweets at a particular day. We expect important tweets with more retweets and favorites to have more influence on the market.

2. An important task to do is to categorize the tweets based on the content. Possible categories are, whether a tweet mentions market (e.g. a company's name or an industry), whether a tweet mentions trade war, whether a tweet expresses positive or negative attitude, whether a tweet uses strong language or not. After categorization, we may use models such as decision tree to find out the most important features that influence the market.

3. We will find out the most frequent phrases that Trump uses in the tweets and check how the appearance of these phrases influence market behavior.

4. There are other ways to refine our response variables. We can include more features into the market data. For example, we can have a column named big_change, indicating whether the stock/index experiences a change that is way larger than 2*standard deviation. Or we can show the volatility of the market by measuring the range from the lowest price to highest price.