Twitter Can Predict Market Drops

Mingxuan Wu, Shuying Wu, Yumeng Zhang, Yuyao Zhang

May 4, 2018

Agenda

- Problem Statement
- Data Analysis
- Model Performance
- Model Validation
- Results Interpretation

Problem Statement: twitter and big data

Problem Statement: twitter and crash

- Users’ sentiment
- Stock market performance

The Framework of Our Project

- Data Preparation
- Data Analysis
- Model Development
- Validation and Interpretation

Tweets **S&P 500 index**	**Build word bag,** **Label major drops**	**Fit and evaluate different models**	**Understand coefficients and decision boundary**
Twitter API, **tweepy library**	**nltk library**	**sklearn library**	**Lasso,** **feature importance**

Data Analysis: build our tweets library

Famous investors
Medias
Financial institutions

Data Analysis: build our word bag

- more than 2 characters with regular expression.
- NLTK (Natural Language Toolkit):
  - Filter stop words
  - stemming using Porter Algorithm
- top 1000 words to build our word bag
- Further feature selection will be implemented within each model.

	co	http	market	amp	year	via	stock	fed	today	week	...
0	21	21	6	5	11	3	9	0	14	6	...
1	17	11	0	3	4	0	0	0	13	2	...
2	38	38	3	12	3	2	6	0	0	5	...
3	32	23	0	20	6	3	0	0	0	0	...
4	39	39	3	22	6	3	6	0	5	0	...

Data Analysis: label a drop in S&P500

- Calculate the weekly log return of S&P 500 index.
- The top 10% worst drop labeled as “1”.
- Base rate will be 10%. (10.4% because of numerical error)

	market shock
0	1
1	0
2	0
3	0
4	1

Model Selection

- We tested various families of classification model.

Supervised learning	Logistic regression	Ordinary logistic
		Logistic GAM
		Lasso logistic
	Random forest	
	Boosting	Gradient descent boosting
	Naïve Bayes	
	SVM	Linear and rbf kernel
	KNN	
Unsupervised learning	Clustering	K-means
		Hierarchical clustering

Data Preprocessing: scaling

- Scaling: Because we have more data in later years, scaling is necessary to make sure the mean of each row is the same.
- We used the built-in scale function in sklearn library.
- However, we also compared with the results without scaling.
- In most cases, scaling is helpful for prediction.

Feature Selection: PCA

- We have 1000 features on only 172 samples, so feature selection is very necessary.
- Some models can do feature selection themselves. (E.g. lasso logistic regression)
- For those who cannot, PCA was implemented.
- Among the 1000 features, PCA shows that 35 dimensions are able to capture 90% of the information.
- We also compared with the result without dimension reduction.

Model Evaluation: metrics

- We looked at 3 metrics to judge if a model is decent.
- **Misclassification rate**: not important due to heavily unbalanced data. (Base rate = 10%)
- **Top 10/20 predictions**: the probability that a market drop happens if our model believes it is very likely to happen.
- **AUC score**: important especially in our case.

Model Evaluation: summary

Model	Misclassification Rate	Top 10/20 predictions	AUC score
Logistic Regression	0.1731	predict 5 in top 20	0.5909
Logistic GAM	0.1154	predict 4 in top 10	0.642
Lasso Logistic	0.1538		
Random Forest	0.1538	predict 3 in top 10	0.723
Gradient Boosting	0.1538	predict 4 in top 20	0.6023
Naïve Bayes	0.1538	predict 2 in top 10	0.5
K-means Clustering	0.1676	-	-
SVM	0.1538	Predict 6 in top 20	0.7798
KNN	0.1538	predict 3 in top 10	0.6478

Validation: lasso logistic regression

- Lasso is able to select features that are most important in prediction.
- Our lasso picks 15 words. They can be divided into 2 groups: twitter account names and descriptive words.
- Below are some selected words and their coefficients.

Word	Coefficient
GDP	-0.00311922
remain	-0.0372262
highlight	-0.02579761
soft	-0.06621725
morganstanley	0.08584726
eagle	0.00410814
crash	0.03951969

Interpretation: lasso logistic regression

- When the tweet is filled with the word “crash”, a real crash/drop is more likely to happen, which indicates that the market is in panic.

- “Soft”, “remain”, “highlight” are safe words, typically referring to a prosperous macroeconomic environment. Under these good news, it is less likely that a crash will happen.

- Interesting fact: why Morgan Stanley has a positive coefficient?

Validation: random forest

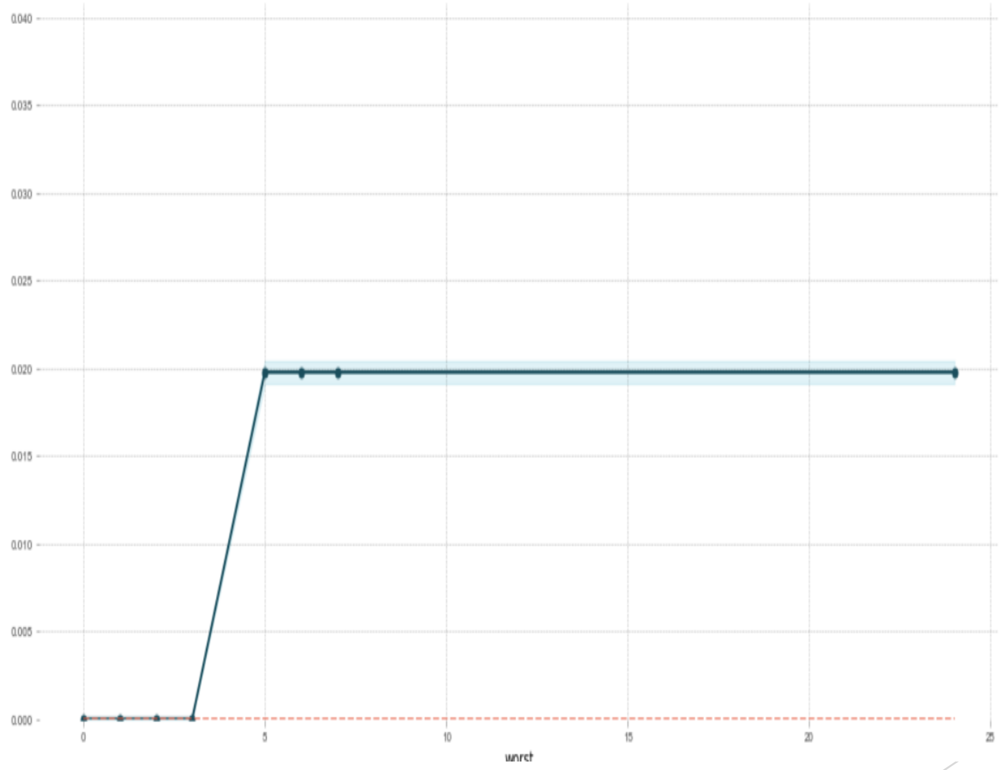
- Random forest is able to describe the importance of every feature.
- We found that random forest picked 2 kinds of words: high-frequency words and descriptive words.
- Below are some examples:

words
continue
remain
worst
market
hk
read
today

Interpretation: random forest

For example, we can plot how the word “worst” affects our prediction.

This also shows that if the market is in panic, a crash/drop is very likely to happen.



Summary: emotion and prediction

- Tweets can reflect the emotion and expectation of the market.
- If the market is expecting or worrying about a crash, it is very likely that a crash will truly happen.

Improvement

- More data: we need to have more twitter accounts to better cover the dates in 2015 and 2016.
- More tuning: we can spend more time adjusting parameters for certain models like gradient descent boosting.
- Bag-of-word: we can find some other method to represent the text to get more context than bag-of-word