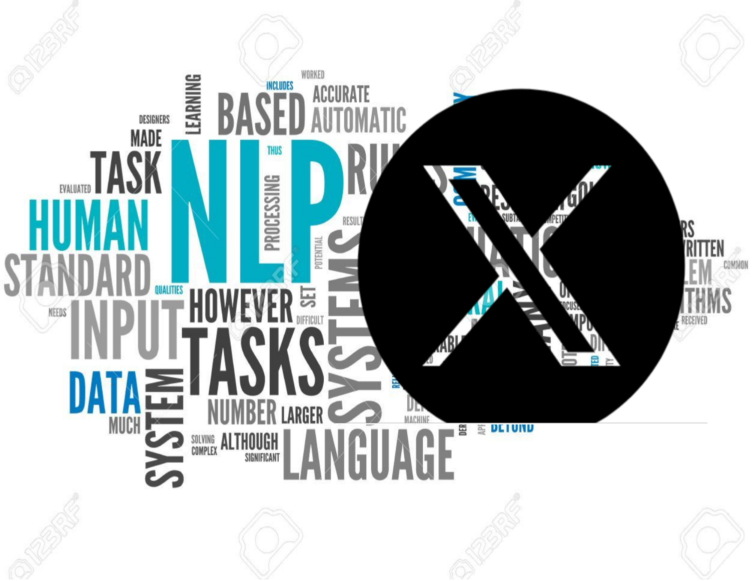
NLP with Disaster Tweet

Andy Wang, Jie Huang



Topic and Objective

What

Improve disaster tweet detection with advanced NLP for emergency response

How

Apply NLP methods, convert texts to numerical data, using logistic regression

Why

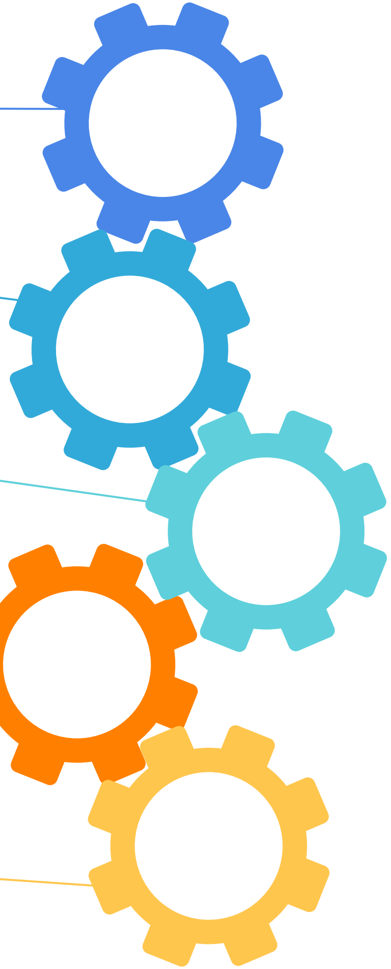
Increase the reliability and usefulness of social media information

Data Feature Label

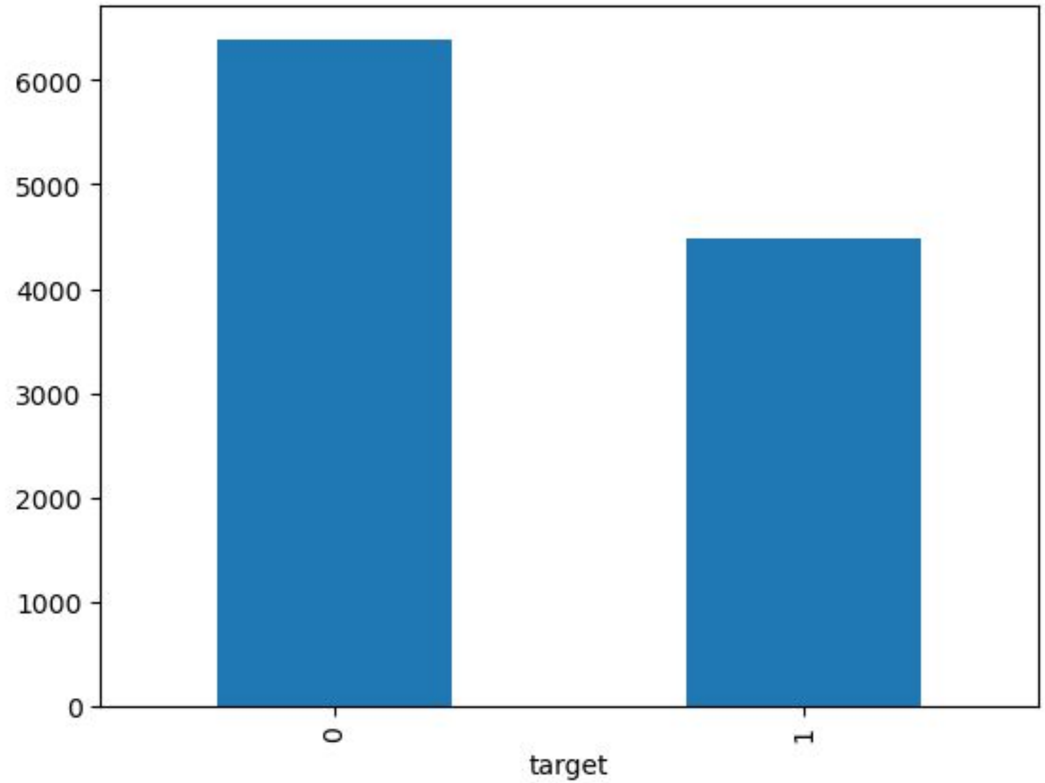
Id: list of numbers for each tweet
Keyword: particular word from the tweet
Location: where the tweet was sent from
Text: the content of the tweet
Target: 1 → disaster, 0 → not

Balance/ Imbalance

Our data is imbalance
→ Use recall, precision, and F1



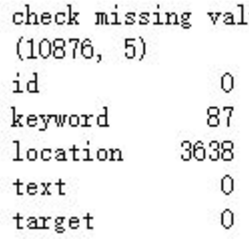
Data is Imbalanced



Description of data

Check Missing Value

Check for missing value with “is null” and fill them with ‘No_data’



```
check missing values:
(10876, 5)
id             0
keyword       87
location    3638
text           0
target         0
```

Preprocessing Text

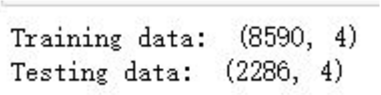
- Removing stop words/special char/URLs
- Stemming
- Lemmatizing
- Tokenization

Check Categorical Values

Categorical values are keyword, location, and tweet text

Split Dataset

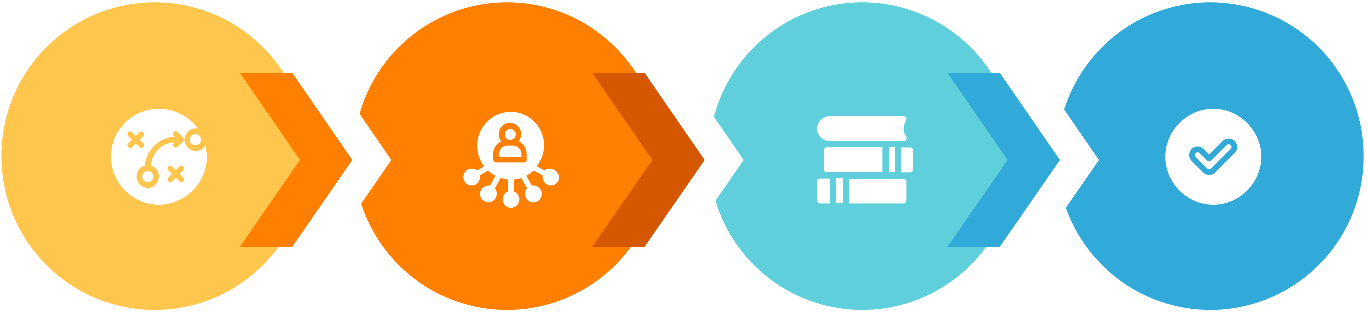
Split dataset to training set and testing set



```
Training data:  (8590, 4)
Testing data:  (2286, 4)
```

Description of data

Preprocessing Text



Removing

Take out URLs, stop words, and special characters to make the text cleaner

Stemming

Make words shorter to their root form

For example, “studies”, “studied”, and “studying” all change to “stud”

Lemmatizing

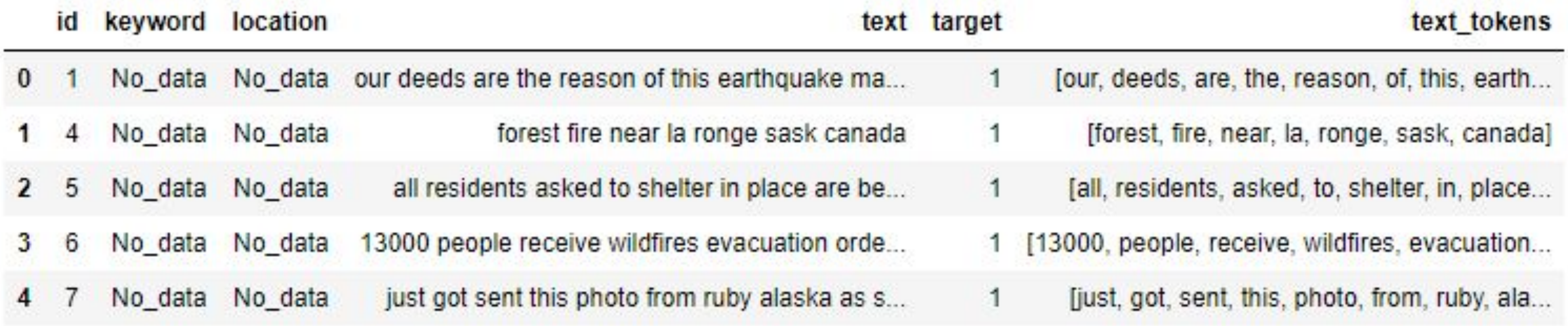
Change words to their basic dictionary form

For example, “studies”, “studied”, and “studying” all change to “study”

Tokenization

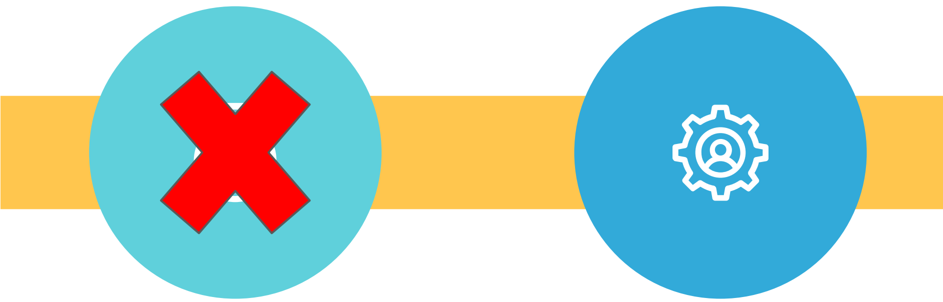
Break the text into single words or tokens

Tokenization



	id	keyword	location	text	target	text_tokens
0	1	No_data	No_data	our deeds are the reason of this earthquake ma...	1	[our, deeds, are, the, reason, of, this, earth...
1	4	No_data	No_data	forest fire near la ronge sask canada	1	[forest, fire, near, la, ronge, sask, canada]
2	5	No_data	No_data	all residents asked to shelter in place are be...	1	[all, residents, asked, to, shelter, in, place...
3	6	No_data	No_data	13000 people receive wildfires evacuation orde...	1	[13000, people, receive, wildfires, evacuation...
4	7	No_data	No_data	just got sent this photo from ruby alaska as s...	1	[just, got, sent, this, photo, from, ruby, ala...

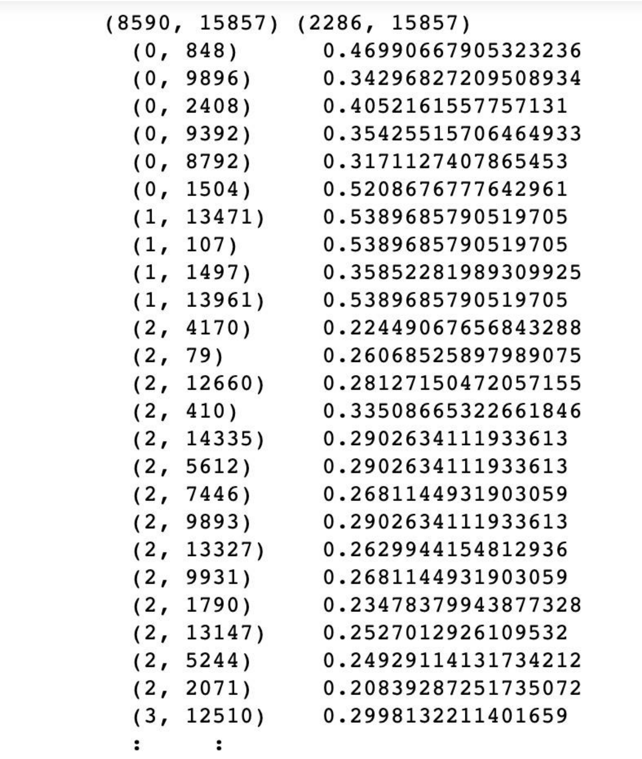
TF-IDF method



Bag-of-Words

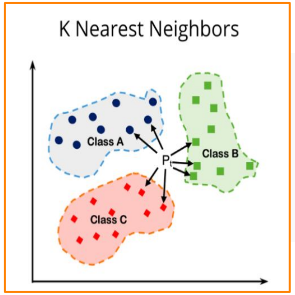
TF-IDF

No need additional normalization for the input



```
(8590, 15857) (2286, 15857)
  (0, 848)	0.46990667905323236
  (0, 9896)	0.34296827209508934
  (0, 2408)	0.4052161557757131
  (0, 9392)	0.35425515706464933
  (0, 8792)	0.3171127407865453
  (0, 1504)	0.5208676777642961
  (1, 13471)	0.5389685790519705
  (1, 107)	0.5389685790519705
  (1, 1497)	0.35852281989309925
  (1, 13961)	0.5389685790519705
  (2, 4170)	0.22449067656843288
  (2, 79)	0.26068525897989075
  (2, 12660)	0.28127150472057155
  (2, 410)	0.33508665322661846
  (2, 14335)	0.2902634111933613
  (2, 5612)	0.2902634111933613
  (2, 7446)	0.2681144931903059
  (2, 9893)	0.2902634111933613
  (2, 13327)	0.2629944154812936
  (2, 9931)	0.2681144931903059
  (2, 1790)	0.23478379943877328
  (2, 13147)	0.2527012926109532
  (2, 5244)	0.24929114131734212
  (2, 2071)	0.20839287251735072
  (3, 12510)	0.2998132211401659
  :	:
```

Description of the models

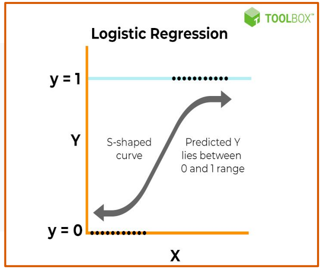


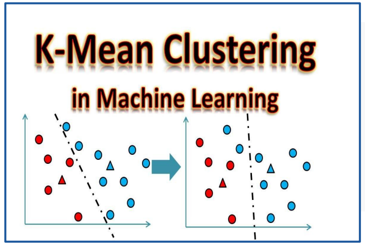
K Nearest Neighbors

Find the k nearest neighbors of sample x, our text in this case

Logistic Regression

Binary classification for only two class





K-means clustering

Initialize cluster centroid, compute distance, compute mean values

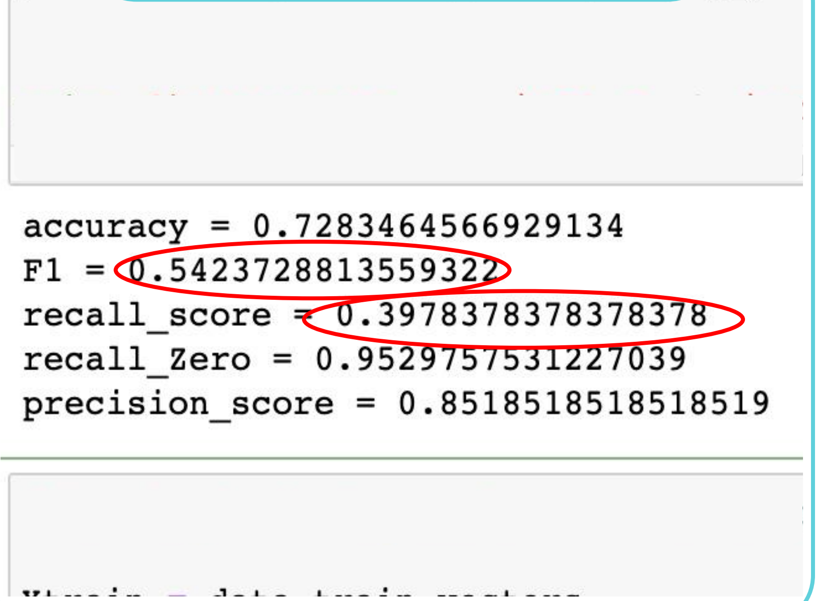
Description of the results

KNN

Not the best model

The F1 and Recall score are too low

Evaluation Score



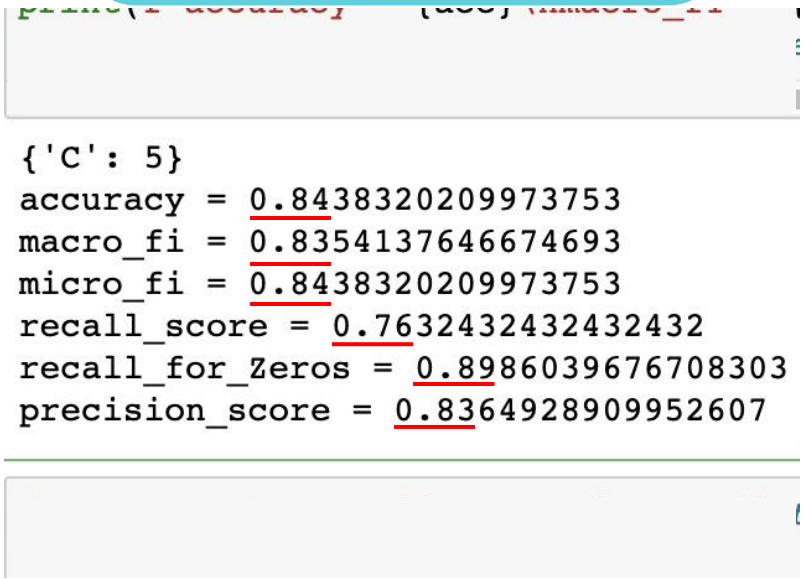
```
accuracy = 0.7283464566929134
F1 = 0.5423728813559322
recall_score = 0.3978378378378378
recall_Zero = 0.9529757531227039
precision_score = 0.8518518518518519
```

Logistic Regression

Best model

The score for each evaluating method are high enough

Evaluation Score



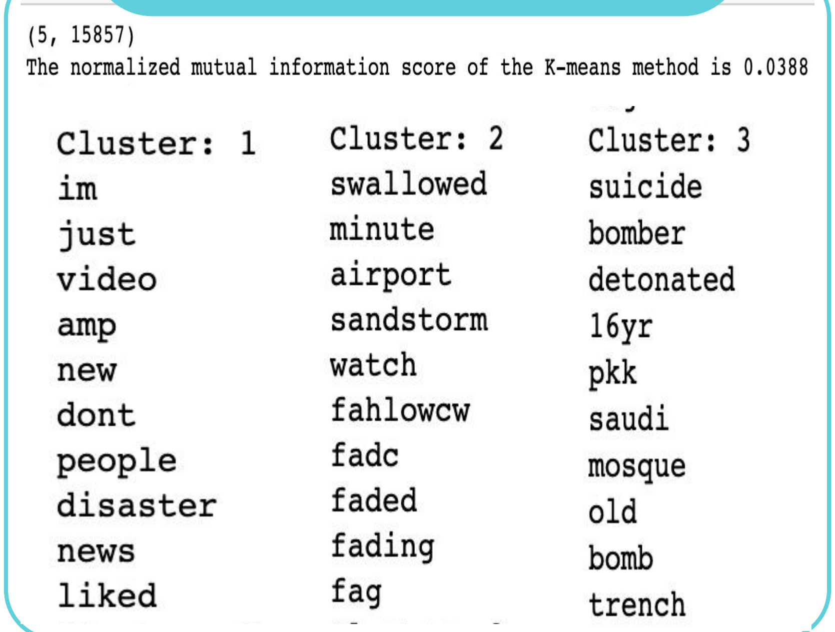
```
{'C': 5}
accuracy = 0.8438320209973753
macro_fi = 0.835413764667469 3
micro_fi = 0.8438320209973753
recall_score = 0.7632432432432432
recall_for_Zeros = 0.8986039676708303
precision_score = 0.8364928909952607
```

k-means

Not the best model

Low NMI score (**0.0388**).
Hard to find representative words

Evaluation Score



```
(5, 15857)
The normalized mutual information score of the K-means method is 0.0388

Cluster: 1       Cluster: 2       Cluster: 3
im               swallowed        suicide
just             minute           bomber
video            airport          detonated
amp              sandstorm        16yr
new              watch            pkk
dont             fahlowcw         saudi
people           fadc             mosque
disaster         faded            old
news             fading           bomb
liked            fag              trench
```

Visualization: Comparison of predicting 1s and 0s

76.32%

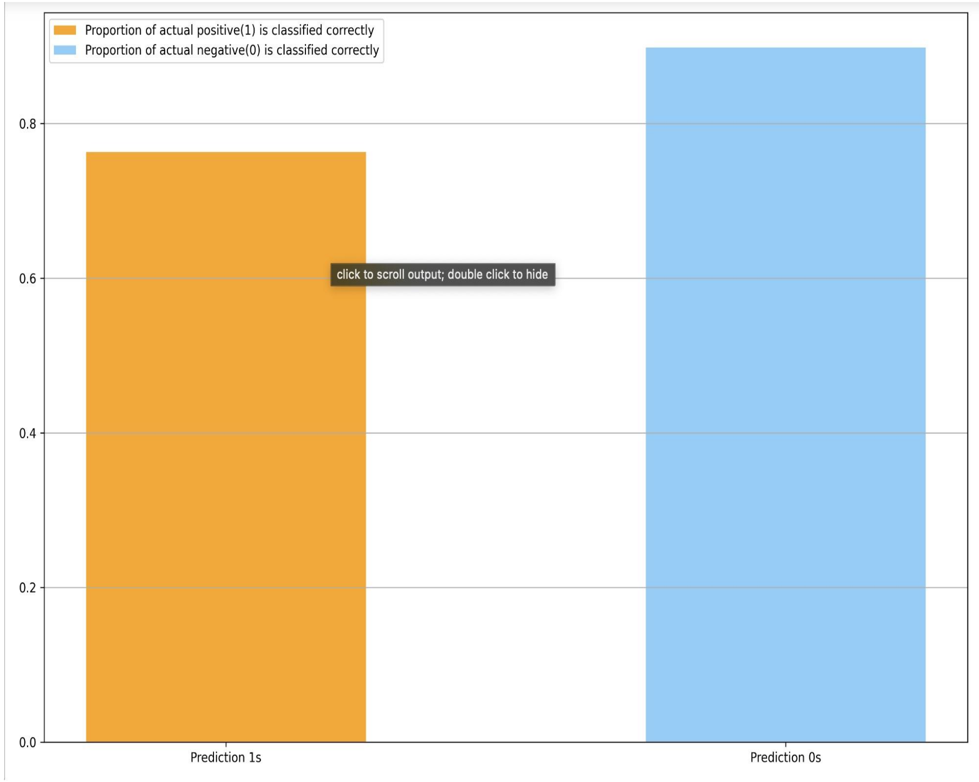
Positive 1s are classified correctly

Of all positive results **(1)**, the proportion that is actual positive

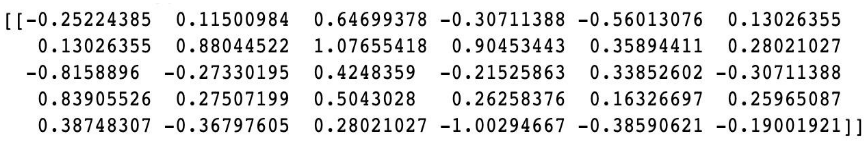
89.86%

Negative 0s are classified correctly

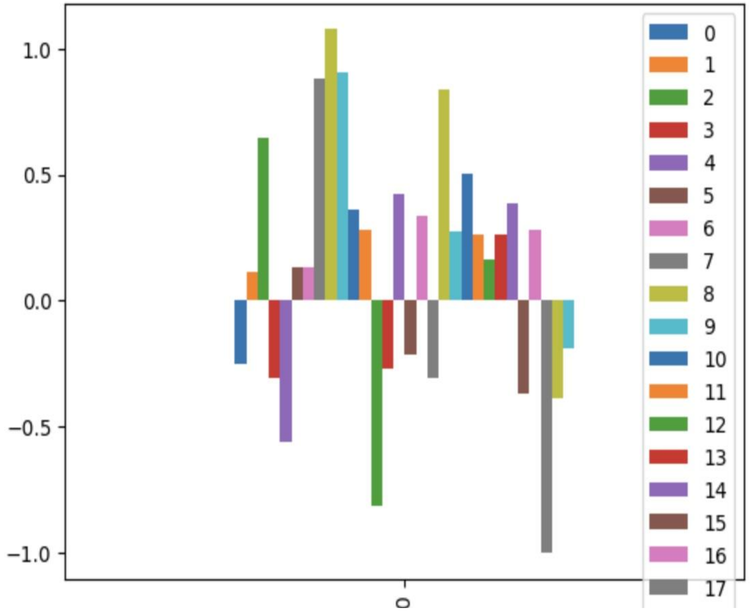
Of all negative results **(0)**, the proportion that is actual negative



Visualization: Influence on parameters



```
[[-0.25224385  0.11500984  0.64699378 -0.30711388 -0.56013076  0.13026355
   0.13026355  0.88044522  1.07655418  0.90453443  0.35894411  0.28021027
  -0.8158896  -0.27330195  0.4248359  -0.21525863  0.33852602 -0.30711388
   0.83905526  0.27507199  0.5043028   0.26258376  0.16326697  0.25965087
   0.38748307 -0.36797605  0.28021027 -1.00294667 -0.38590621 -0.19001921]]
```



$\mathcal{W}$ with larger absolute values

The parameters with a larger values indicate that it has a bigger influence in that feature

Thanks for listening!

Q&A