SUBREDDIT IDENTIFICATION IN ONLINE FORUMS

Presented by Yetunde Obasade

CONTENTS AND FLOW

01

PROJECT OVERVIEW

02

PROBLEM STATEMENT

03

DATA ANALYSIS

Cleaning and Understanding the data

04

MODELING AND INSIGHTS

05

CONCLUSIONS/RECOMMENDATIONS

PROJECT OVERVIEW

The goal of this project is to take data from Reddit and build a model to identify whether or not a given post is from a rollerskating subreddit or a rollerblading subreddit.

We want to build and train multiple models in order to find the best optimization for these predictions.

Image from tumblr.com

BLADES VS SKATES

Roller blades

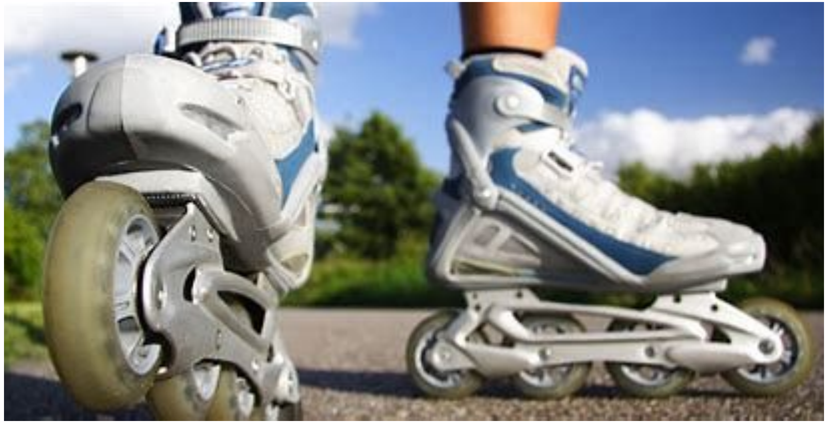


Image from www.centennialparklands.com.au

Roller skates

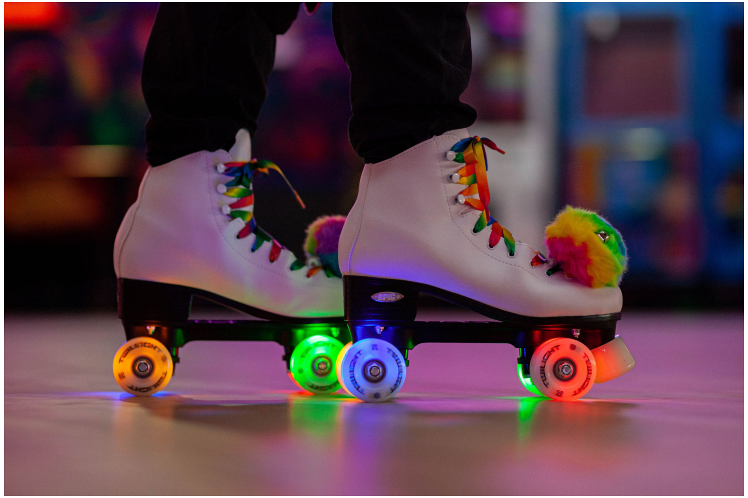


Image from www.devaskation.com/

“Can we build a model that accurately predicts which subreddit a rollerblading/skating post comes from, while also optimizing for the best predictions?”

PROBLEM STATEMENT

01

DATA ANALYSIS

WEBSCRAPING

We scraped the information we needed from two subreddit posts.
We gathered over 4,000 rows of data for both subreddits.

CLEANING DATA

We cleaned the data using Natural Language Processing techniques. This includes: removing unnecessary columns (focusing solely on the posts), removing stop words, removing special characters, lowercasing all words and lemmatizing the words.

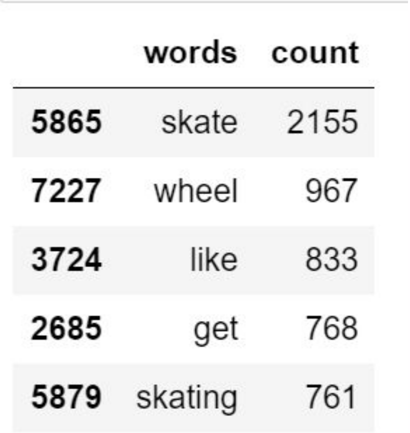
CLEANING DATA

We also tokenized the words (separating by whitespace or line breaks), removed NaN values and replaced any “deleted” inputs with a whitespace.

UNDERSTANDING THE DATA

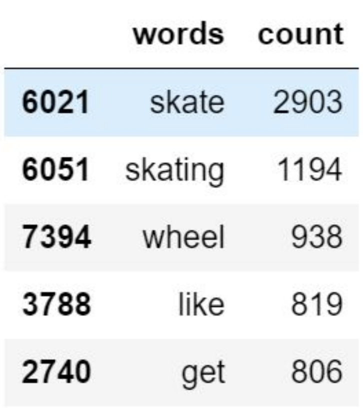
Top five words for rollerblading (1) and rollerskating (2).

Skating is used much less in the rollerblading dataset compared to the rollerskating dataset.



	words	count
5865	skate	2155
7227	wheel	967
3724	like	833
2685	get	768
5879	skating	761

(1)

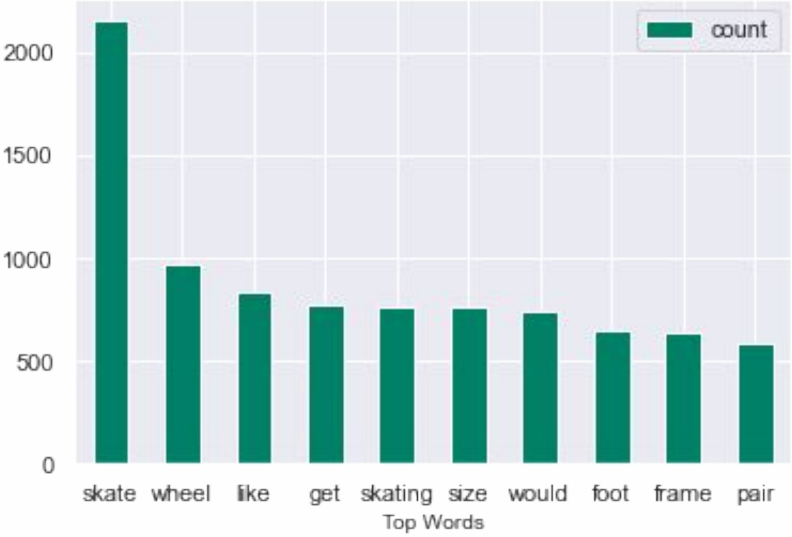


	words	count
6021	skate	2903
6051	skating	1194
7394	wheel	938
3788	like	819
2740	get	806

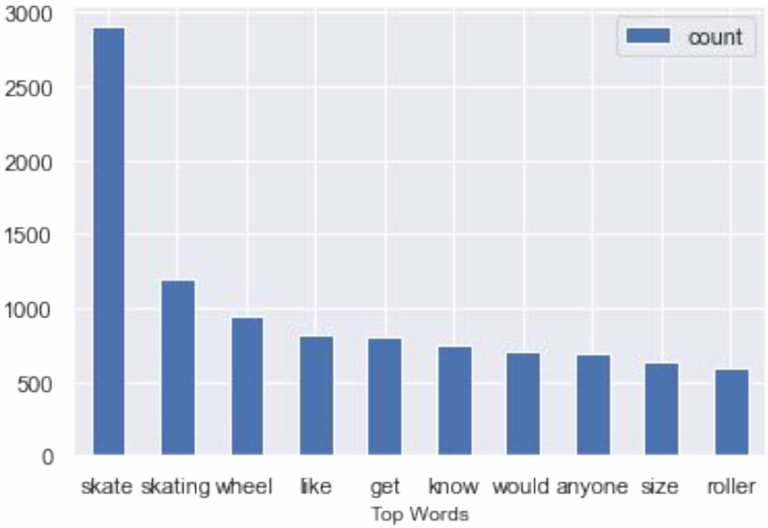
(2)

UNDERSTANDING THE DATA

The word *skate* is used more than double the word *wheel* for "blade" category. The word *skate* is used more than double the word *skating* for the "skate" category.



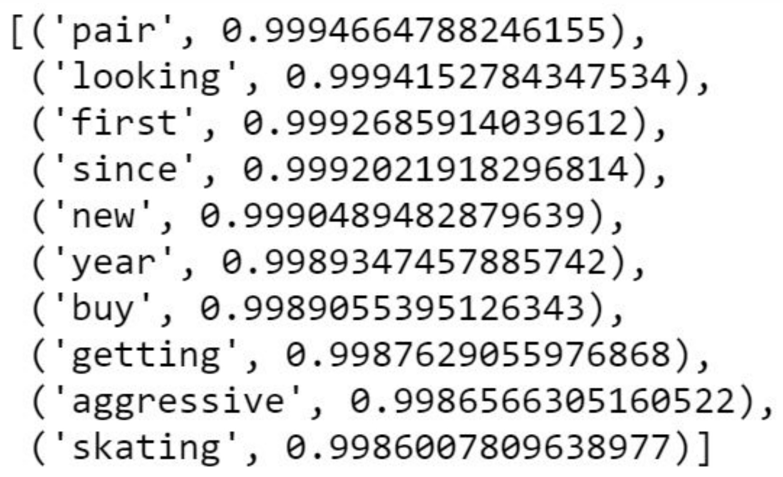
(1)



(2)

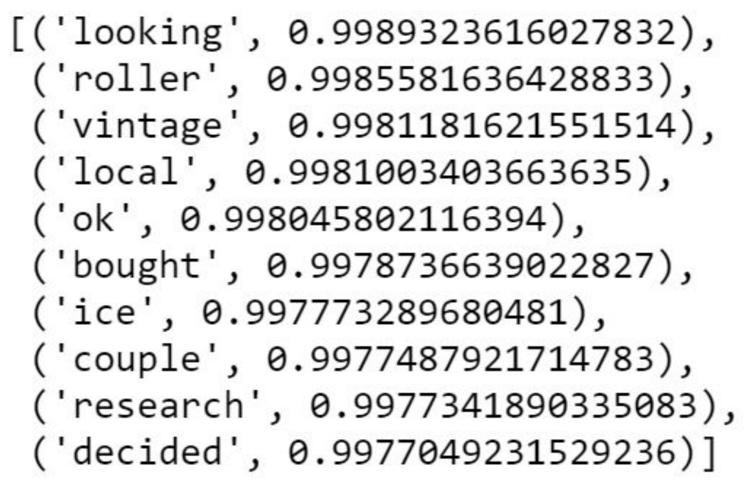
UNDERSTANDING THE DATA

- Top 10 most similar words for the blading category (1).
- Top 10 most similar words for the skating category (2).
- They differ immensely with only one word showing up in both.



```
[('pair', 0.9994664788246155),
 ('looking', 0.9994152784347534),
 ('first', 0.9992685914039612),
 ('since', 0.9992021918296814),
 ('new', 0.9990489482879639),
 ('year', 0.9989347457885742),
 ('buy', 0.9989055395126343),
 ('getting', 0.9987629055976868),
 ('aggressive', 0.9986566305160522),
 ('skating', 0.9986007809638977)]
```

(1)

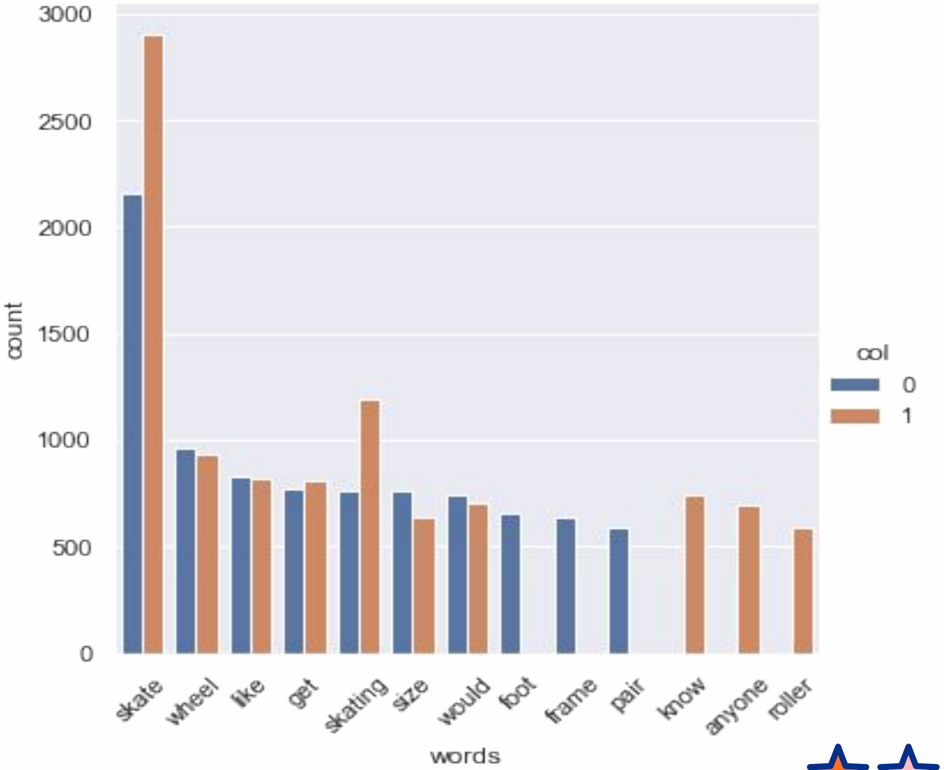


```
[('looking', 0.9989323616027832),
 ('roller', 0.9985581636428833),
 ('vintage', 0.9981181621551514),
 ('local', 0.9981003403663635),
 ('ok', 0.998045802116394),
 ('bought', 0.9978736639022827),
 ('ice', 0.997773289680481),
 ('couple', 0.9977487921714783),
 ('research', 0.9977341890335083),
 ('decided', 0.9977049231529236)]
```

(2)

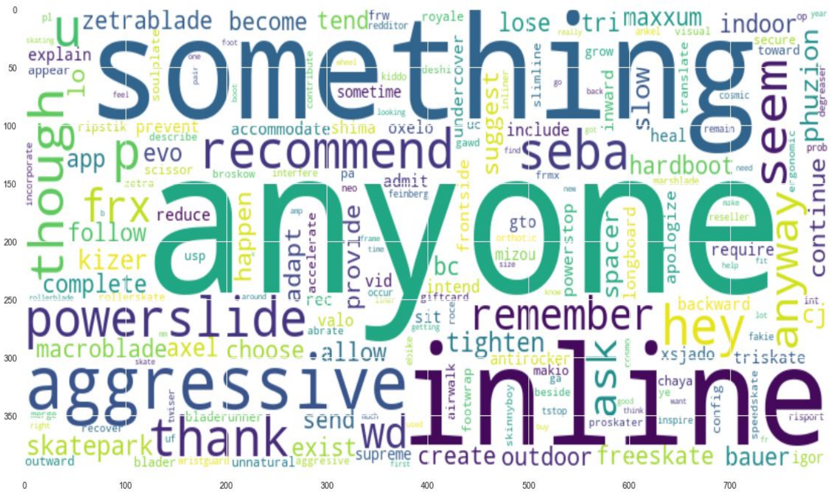
UNDERSTANDING THE DATA

- Top 10 word comparisons.
- Last 3 words of both are different for each subreddit.
- After the top 10 words, words begin to no longer match up.

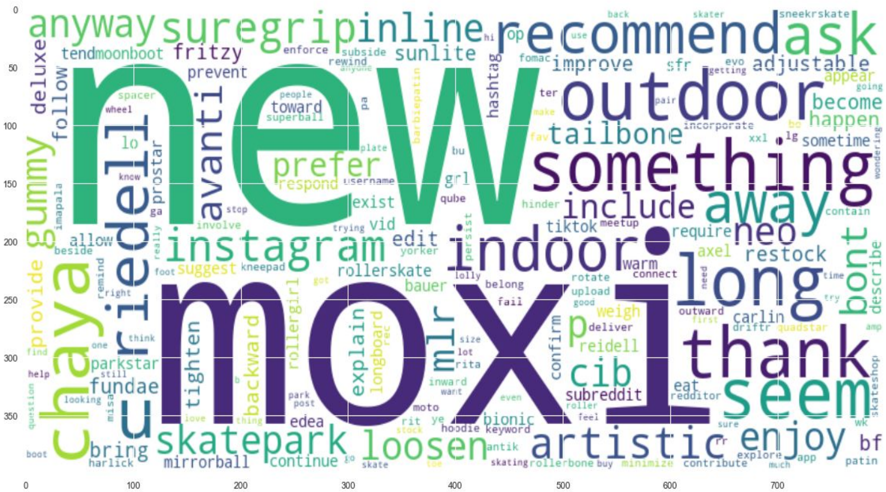


UNDERSTANDING THE DATA

- Word cloud comparison.
- (1) - rollerblading, (2) is rollerskating.
- In comparing the clouds to the graphs above, these are the combined words.

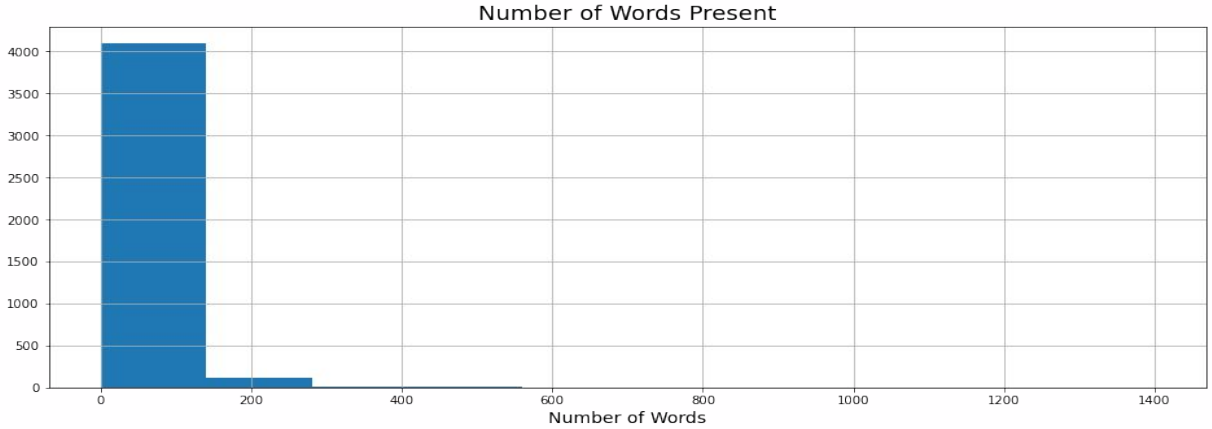


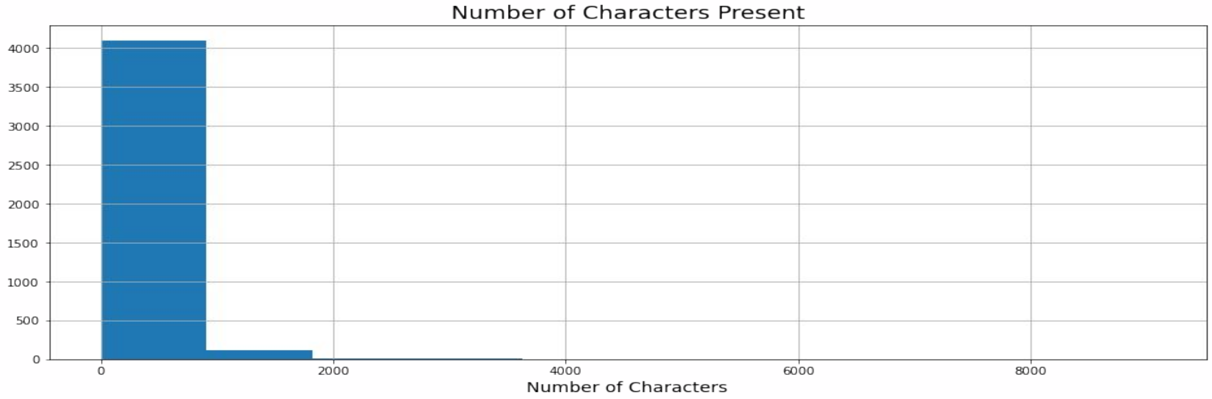
(1)



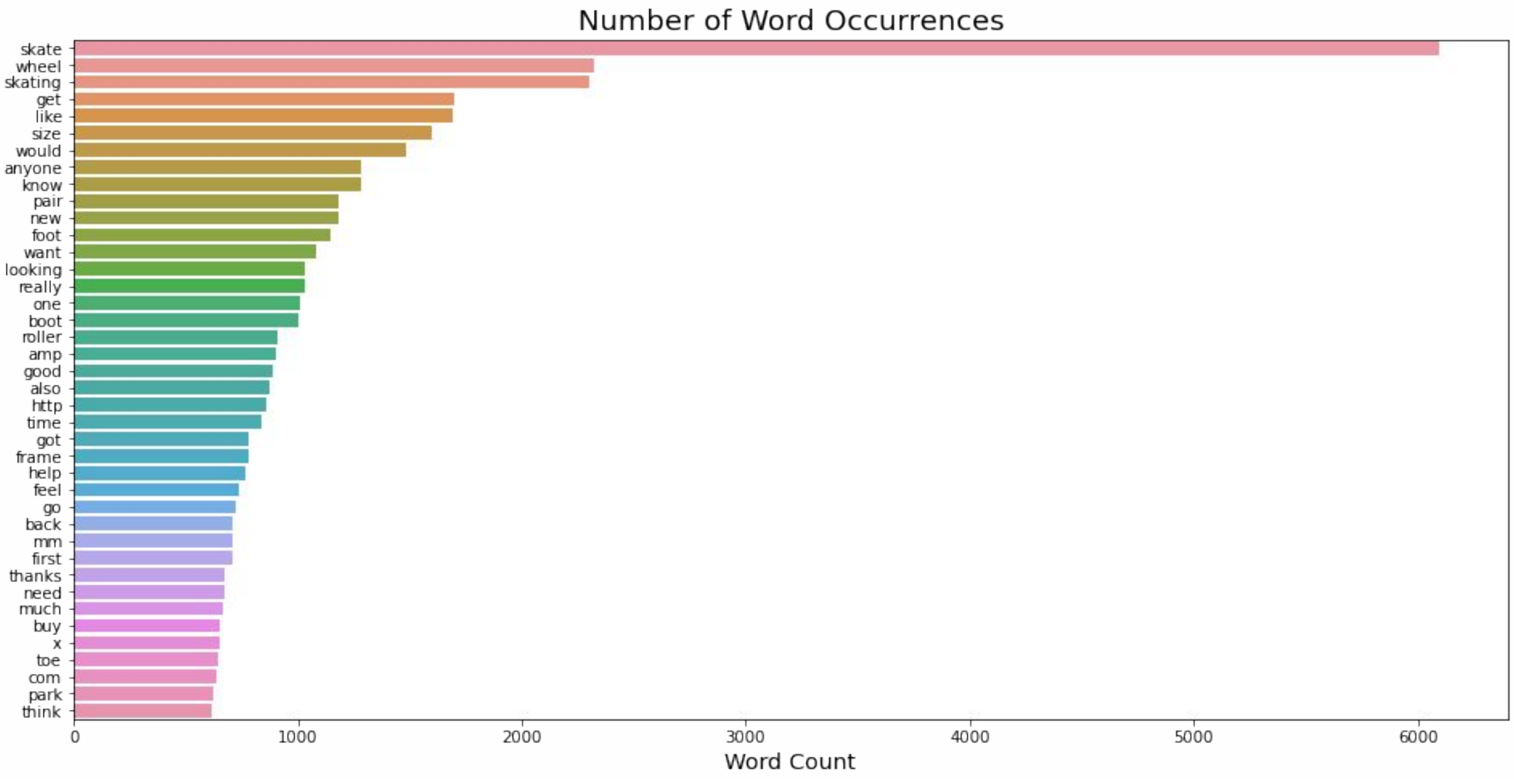
(2)

UNDERSTANDING THE DATA





UNDERSTANDING THE DATA



Number of word occurrences between both models.

02

MODELING AND INSIGHTS

MODELING

- We used a multinomial Naive Bayes model for our baseline
- Our baseline model did not have a pipeline nor a grid search applied to it.
- Our other 2 models were the Random Forest Classifier and Logistic Regression.
- We did apply a Count Vectorizer (transforms text into a vector based on occurrence of word count) to all models.

INSIGHTS

- Baseline model accuracy score: 89.29%
- Baseline Training score: 99.05%
- Baseline Testing score: 89.29%

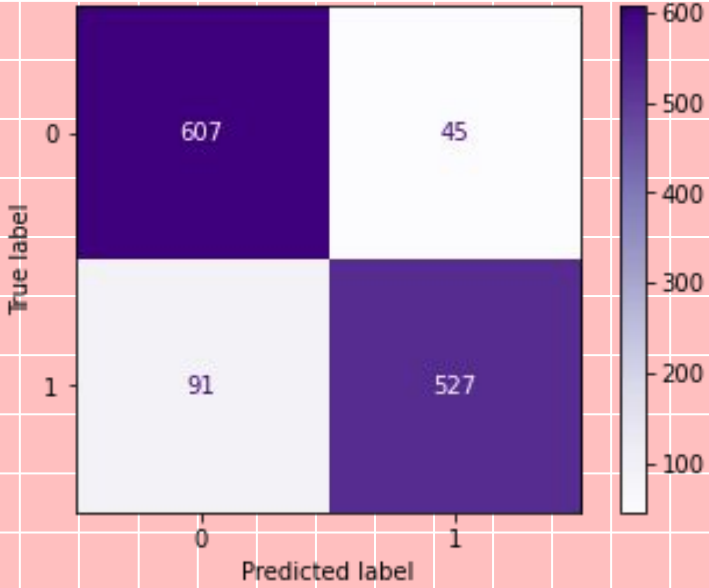
Though we have a high train score here, our test score is lower by 10%. This model is very overfit.

(Overfitting means the model is “learning” the training data too much).

INSIGHTS (BASELINE MODEL)

CONFUSION MATRIX

- 527 true positives
- 607 true negatives
- Higher number of false negatives compared to false positives



INSIGHTS (RANDOM FOREST)

- RF model accuracy score: 89.44%
- RF Training score: 98.22%
- RF Testing score: 96.59%

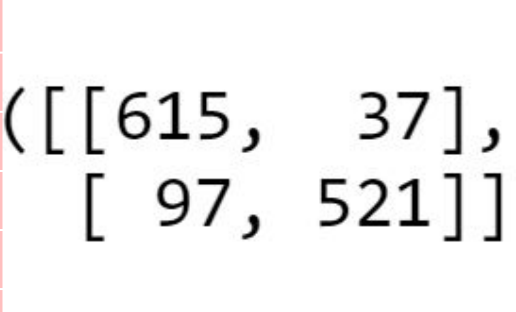
There is some overfitting, however, this is a good model result.

The margin is much lower between the testing and training scores.

INSIGHTS (RANDOM FOREST MODEL)

CONFUSION MATRIX

- 521 true positives
- 615 true negatives
- Lower number of false negatives compared our baseline model



```
([[615,  37],
  [ 97, 521]]
```

INSIGHTS (LOGISTIC REGRESSION)

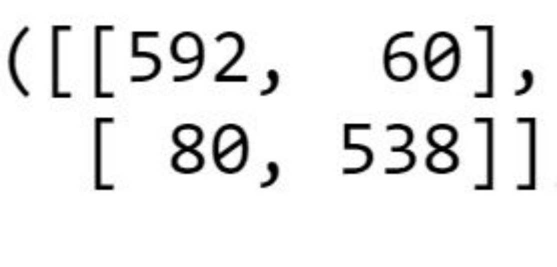
- LR model accuracy score: 89%
- LR Training score: 99.99%
- LR Testing score: 95.69%

- The training score went up by .1 compared to the random forest-grid search model. Our testing score went down by .1.
- In comparison to our Multinomial NB model, this model outperforms on the training score.
- The model is more overfit than the Random Forest.

INSIGHTS (LOGISTIC REGRESSION MODEL)

CONFUSION MATRIX

- 538 true positives
- 592 true negatives
- Higher number of false negatives compared previous models.



```
([[592,  60],
  [ 80, 538]]
```

CONCLUSIONS AND RECOMMENDATIONS

03

CONCLUSION

In relation to our problem statement, we can predict- fairly accurately- which submission a subreddit came from.

- For best results, we should use our random forest classifier model.
- This model has the smallest margin between the training and testing scores.
- The predictions are still above 95%. RF model = good!

CONCLUSION

- Though logistic regression had the highest training score, there was a larger margin of overfitting between the train and test scores.
- 7 out of our top 10 words overlapped between both models
- Our random forest classifier had the lowest number of false positives at 37, however our false negatives were the highest at 97

RECOMMENDATIONS

- Our next steps would be to change and optimize the random forest parameters to get a higher training score with a lower margin of overfitting (example- TFIDF vectorizer).
- Can also explore the parameters for logistic regression to close the gap between the training and testing scores.
- Try a KNN model through a pipeline with gridsearch and adjust the KNN neighbors.
- Change the data to view comments vs submissions in subreddit.

GET SOME BLADES!

05

QUESTIONS?

Image from giphy.com