

The Sorting

Project 3: Web APIs & NLP

Melvin Chandra

DSIF-2

Problem Statement:

Sorting posts from two
similar subreddits

r/LifeProTips

r/YouShouldKnow

When a cat shows you its belly, pet around the cheeks and not the belly itself.

Cats show their belly as a sign of trust, so going for their belly is seen as an immediate violation of that trust and may result in biting.

r/LifeProTips

If your cat has shown in the past that it likes belly rubs, that's an exception of course, but the average kitty just wants you to know that you are trusted by exposing its most vulnerable area to you.

You can freeze fresh bread and reheat it in an oven or an air fryer when you need it.

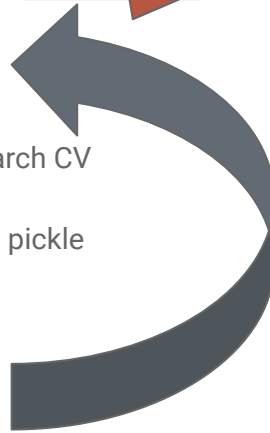
When you buy fresh bread there is often too much of it, especially if you live alone. If you leave the bread at room temperature (about 70°F), it will become stale (lose its texture and taste).

r/YouShouldKnow

If you put it in a freezer, you can take it out after many months, reheat it, and have crispy slices that are still soft on the inside with most of the taste and texture preserved.

Methodology

- Data Acquisition
 - Pushshift API
- Data Cleaning
 - Remove tags
 - Remove URL
- Preprocessing
 - Count Vectorizer
 - Train Test Split
- Modelling
 - Hyperparameter tuning via Gridsearch CV
 - Multiple models
 - Export/import models via BZ2 and pickle
- Evaluation & Visualization
 - Confusion Matrix
 - ROC plot
 - Feature Importance



Feedback
stop words

Data Acquisition

- Pushshift's API
- 100 x 30 cycle (15 each)
- 20s delay per cycle to avoid overloading API
- Parameters:
 - Subreddit
 - Size
 - Before
 - Is_video
 - Stickied
 - Score

Data Cleaning

Title

LPT if you have to give your email to a company, use the company's name as your listed first name. That way if they sell your information to third parties, the spam emails will be auto-filled with the original company's name instead of your own so you know who sold your info.

Selftext

Obviously only do this with unimportant accounts and only if your name isn't important for legal reasons, but I've done this for apps, conferences, demos, etc. Just recently I got spam email that said "Dear XYZ App Trial, don't miss out in these deals!" - so I knew XYZ App sold the data.

Edit: To those asking what you can do, if a company sells your data without your permission you can file a formal complaint with the FTC here:
<https://www.consumer.ftc.gov/blog/2015/08/want-privacy-tell-us-about-it>

2nd Edit: I've been made aware in the comments that others have had this idea and posted it to the sub before and that I'm an idiot. My apologies for not knowing that, I hope I didn't ruin anyone's evening too badly.

Count Vectorizer

Parameters = {

- stop_words : english +
- max_df : 0.7
- ngram_range : (1 , 4)
- min_df : 2
- max_features : 50,000

}

Extra Stop Words = {

- https
- com
- www
- org
- http
- don
- ll
- gt
- ve
- edit
- Really

}

Mapping = {

- LifeProTips : 1
- YouShouldKnow : 0

}

13,488

2,183

Train

728

Test

Random Forest

Parameters = {

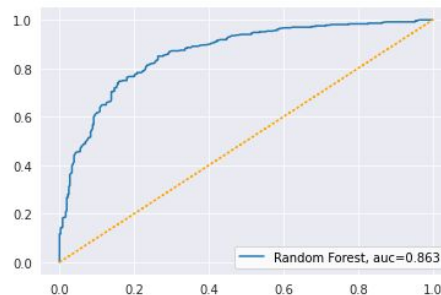
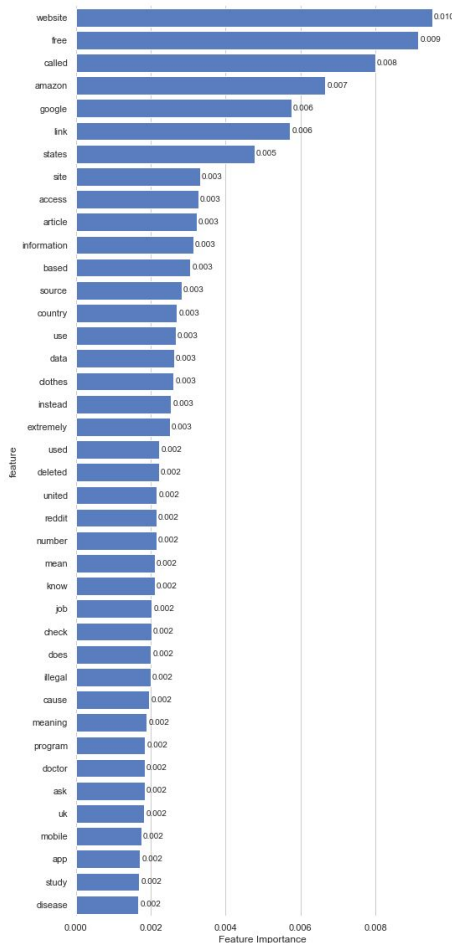
- criterion : gini
- max_features : log2
- n_estimators : 1,000

}

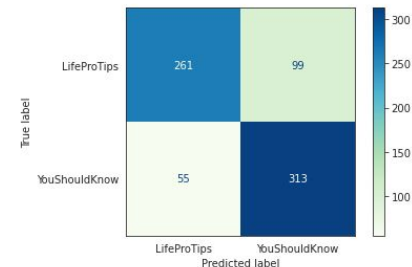
Accuracy of 79%

Recall of 85%

Specificity of 72%



	precision	recall	f1-score	support
0	0.83	0.72	0.77	360
1	0.76	0.85	0.80	368
accuracy			0.79	728
macro avg	0.79	0.79	0.79	728
weighted avg	0.79	0.79	0.79	728



Stacking Classifier

Parameters = {

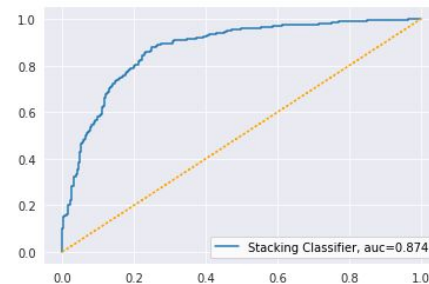
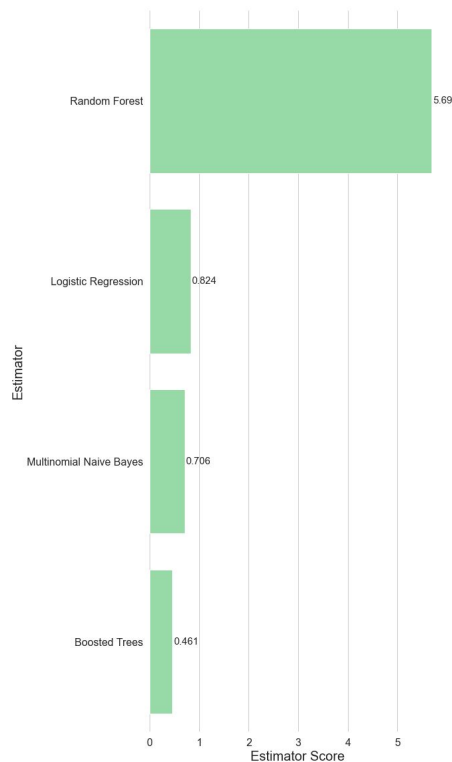
- Random Forest
 - criterion : gini
 - max_features : log2
 - n_estimators : 1,000
- Logistic Regression
 - max_iter : 100
 - penalty : l2
 - solver : liblinear
- Naive Bayes
 - alpha : 0.11
 - fit_prior : True
- Boosted Trees
 - n_estimators : 250

}

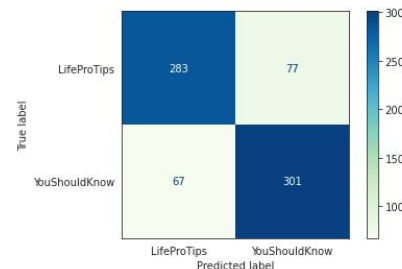
Accuracy of 80%

Recall of 82%

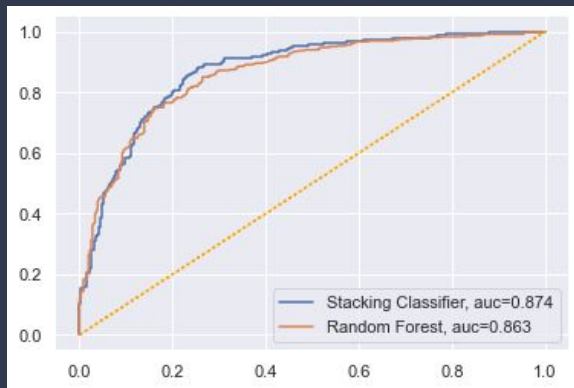
Specificity of 79%



	precision	recall	f1-score	support
0	0.81	0.79	0.80	360
1	0.80	0.82	0.81	368
accuracy			0.80	728
macro avg	0.80	0.80	0.80	728
weighted avg	0.80	0.80	0.80	728



Summary



Model	Random Forest	Stacking Classifier
Accuracy	79%	80%
Recall	85%	82%
Specificity	72%	79%

- **Accuracy** of both models are very **similar**.
- **Random Forest** if you are predicting **LifeProTips**
- **Stacking Classifier** if you are predicting **YouShouldKnow**