which subreddit??

r/travel or not r/travel?
That is the question.

The challenge:

Help reddit decide if it is worth investing in a data science team to perform text analytics.

Project Scope:

Proof-of-concept: Build a predictive model to classify subbredit posts into the correct category: r/travel, or not r/travel.

Goal: 90% Accuracy or Higher

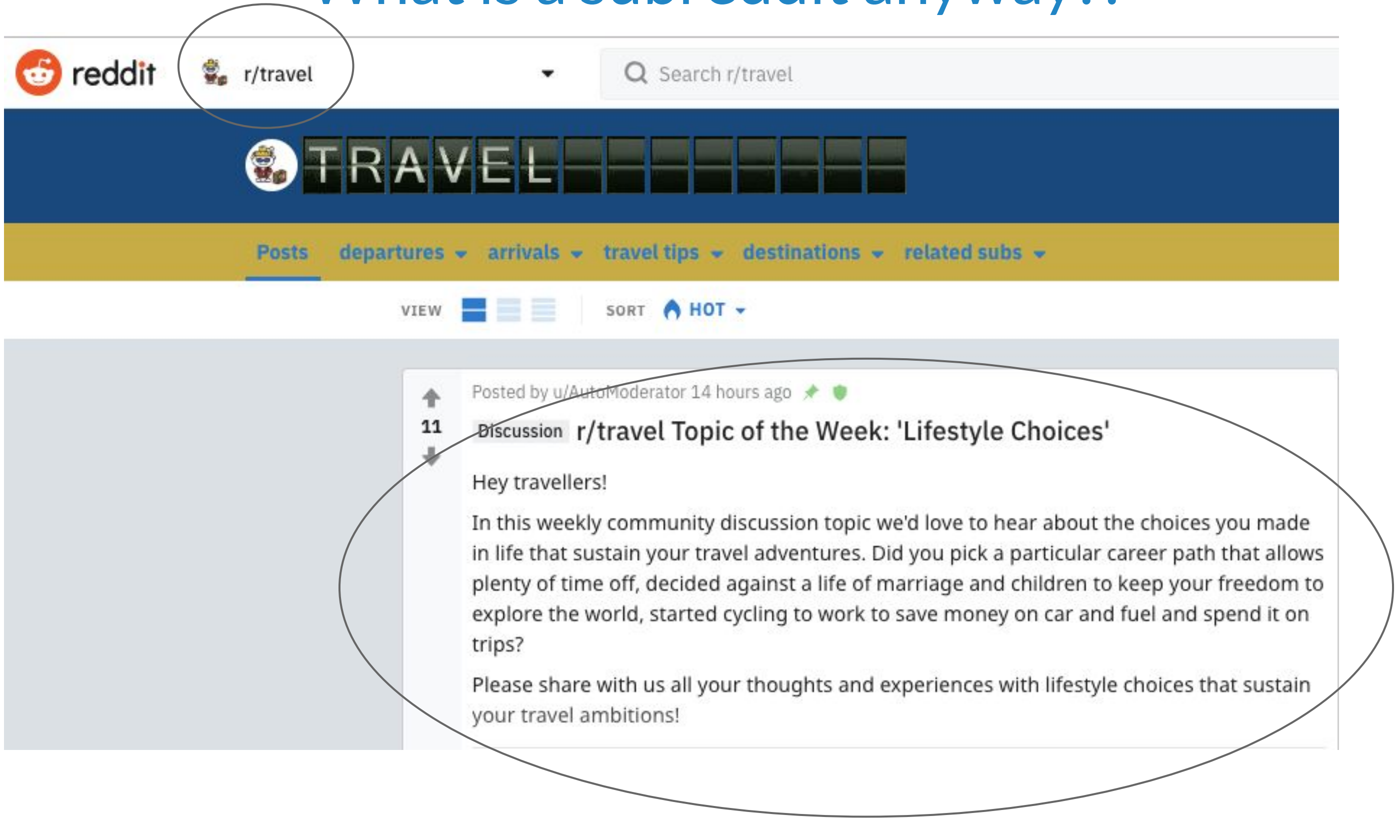
Subreddits for comparison:

- **r/Fitness**
- **r/gardening**
- **r/wine**

Overview of the Process

- ▷ Getting data
- ▷ Exploring and cleaning data
- ▷ Building and analyzing models
- ▷ Drawing conclusions

What is a subreddit anyway??



Posted by u/AutoModerator 14 hours ago

Discussion r/travel Topic of the Week: 'Lifestyle Choices'

Hey travellers!

In this weekly community discussion topic we'd love to hear about the choices you made in life that sustain your travel adventures. Did you pick a particular career path that allows plenty of time off, decided against a life of marriage and children to keep your freedom to explore the world, started cycling to work to save money on car and fuel and spend it on trips?

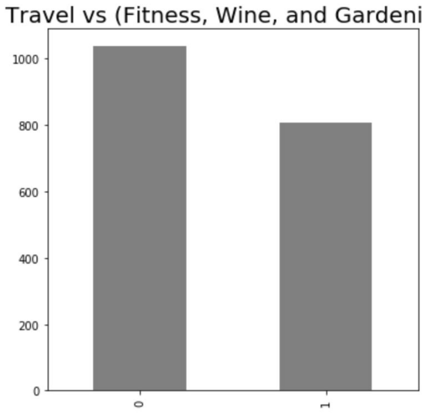
Please share with us all your thoughts and experiences with lifestyle choices that sustain your travel ambitions!

Getting Data - The Reddit API

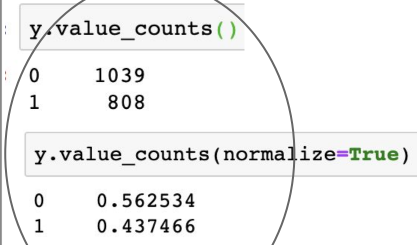
- Requests.get()
  - python code to fetch data from reddit in 25-post batches:
  - requests.get(self.url, params=params, headers= self.header)
- RedditPostReader class to manage interaction with the reddit site
  - gather_posts(url,n=100) method
  - Configured to save these fields:

    ['subreddit', 'id', 'selftext', 'title', 'author', 'created', 'ups', 'downs']
  - Skips posts with empty string subreddits (images or videos)
- Pre-processing fetched posts
  - Drop duplicates - lost more than half the rows
  - Saved files

Exploring and Cleaning Data

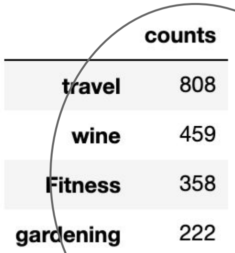


Total posts: 1,847
Baseline Accuracy: 56%

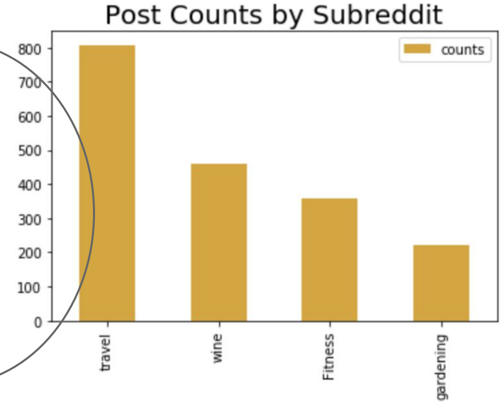


```
y.value_counts()

0     1039
1      808
```

```
y.value_counts(normalize=True)

0     0.562534
1     0.437466
```



	counts
travel	808
wine	459
Fitness	358
gardening	222



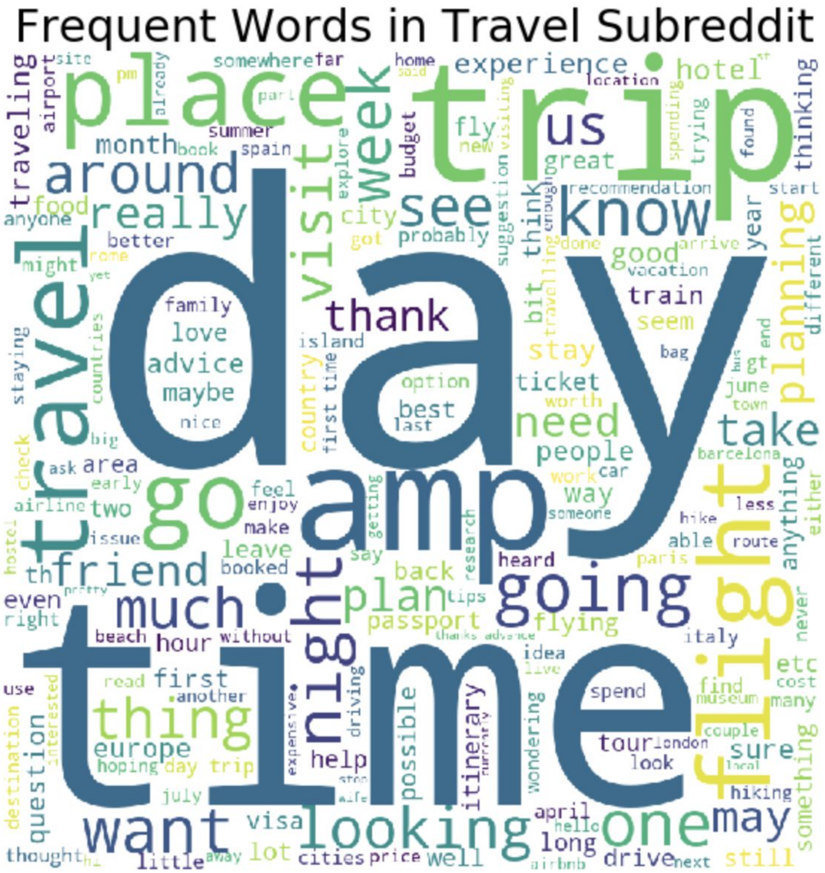
Use a RegexpTokenizer(r'[a-z]+') to go from this:

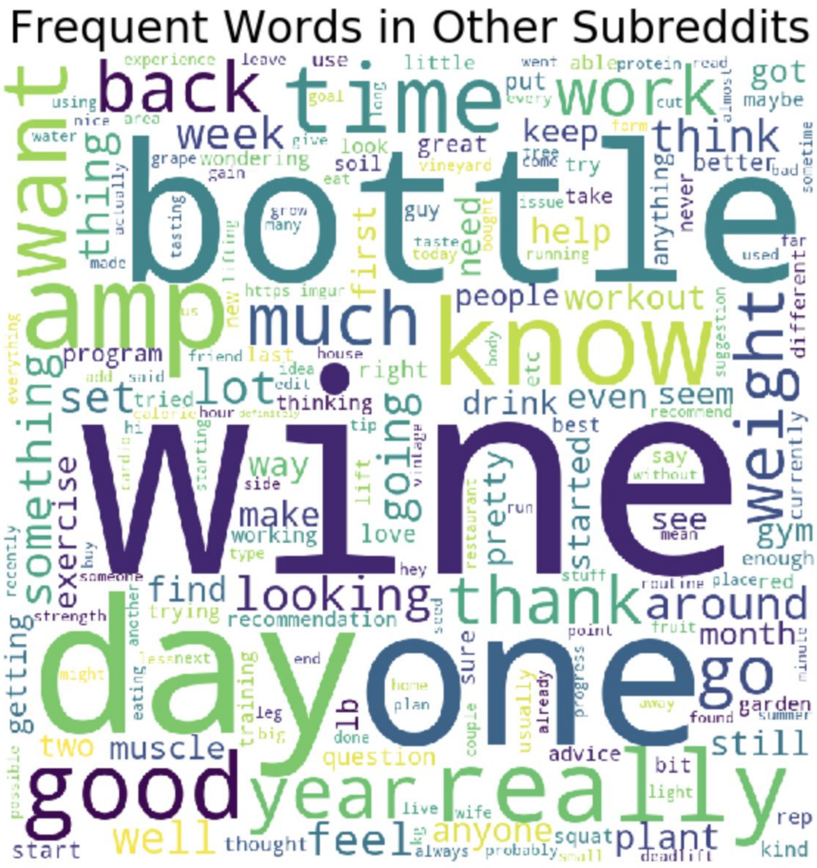
```
"Hi Reddit! My friends and I will be going on our post-graduation trip to Korea, and would like to travel to Seoul, J
eju, and Busan in May.\n\nWhat is the best/cheapest way to travel? Plane or train?\n\nAlso, when looking at flights,
a lot of airlines have different types of fares (special, discount, normal, event), which fare usually includes check
ed baggage (we saw that Eastar only had two out of three fare types that included baggage, but Asiana airlines does
n't have this information explicitly on their site)?\n\nThanks for your help in advance! :)\n\n\\*Also has been poste
d to r/koreatravel"
```

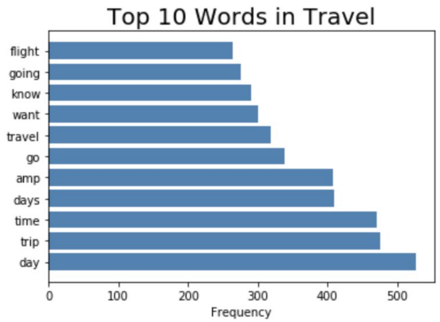
to this:

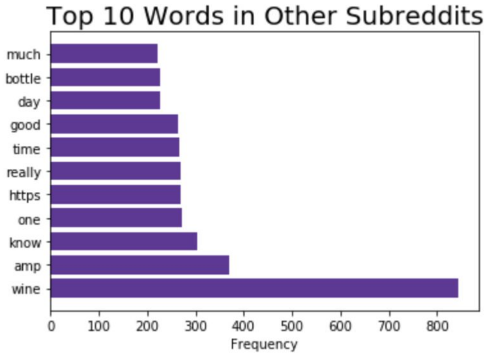
```
'hi reddit my friends and i will be going on our post graduation trip to korea and would like to travel to seoul jeju
and busan in may what is the best cheapest way to travel plane or train also when looking at flights a lot of airline
s have different types of fares special discount normal event which fare usually includes checked baggage we saw that
eastar only had two out of three fare types that included baggage but asiana airlines doesn t have this information e
xplicitly on their site thanks for your help in advance also has been posted to r koreatravel'
```

Exploring and Cleaning Data





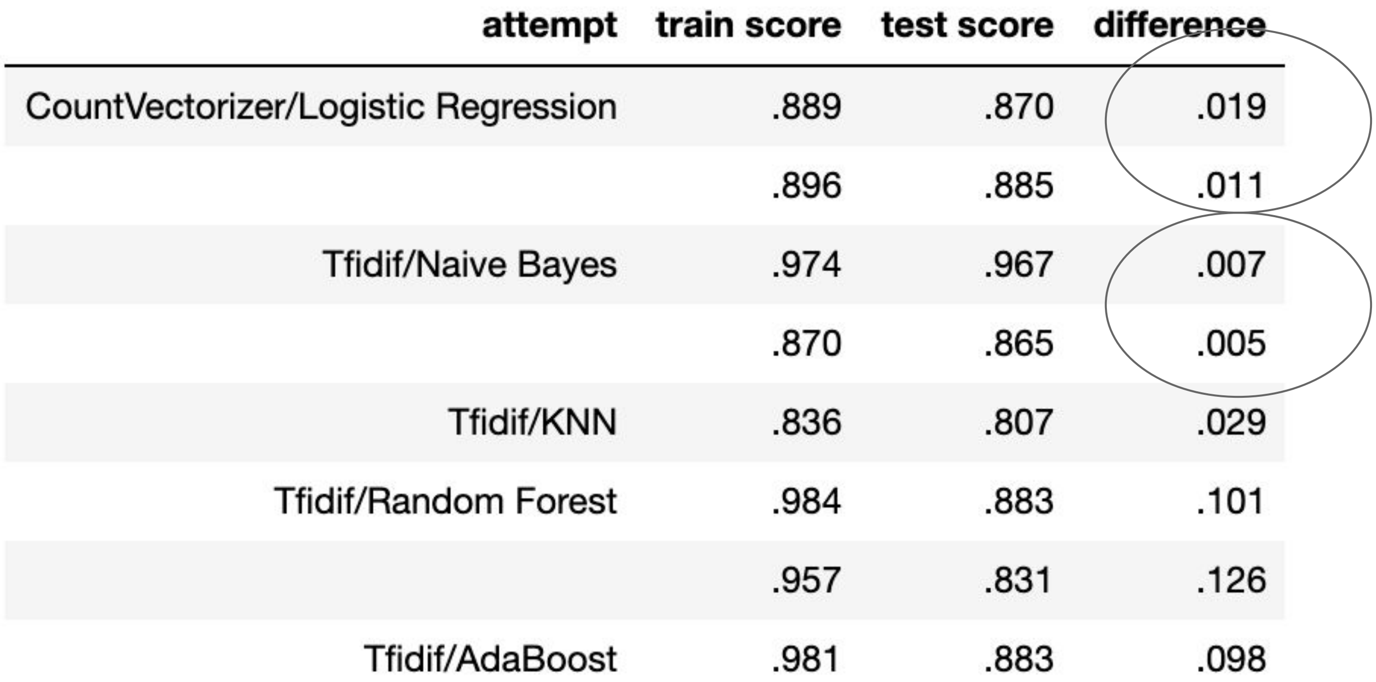




Building and Analyzing Models

After cleaning the data we were left with a dataset with 1,847 posts. We split the data into two sets for the purpose of training and testing the models:
Size of training data: 1385, Size of test data: 462

Use a GridSearch Technique to let the computer try a variety different tokenizers, estimators, and tuning parameters to find the best performing model.



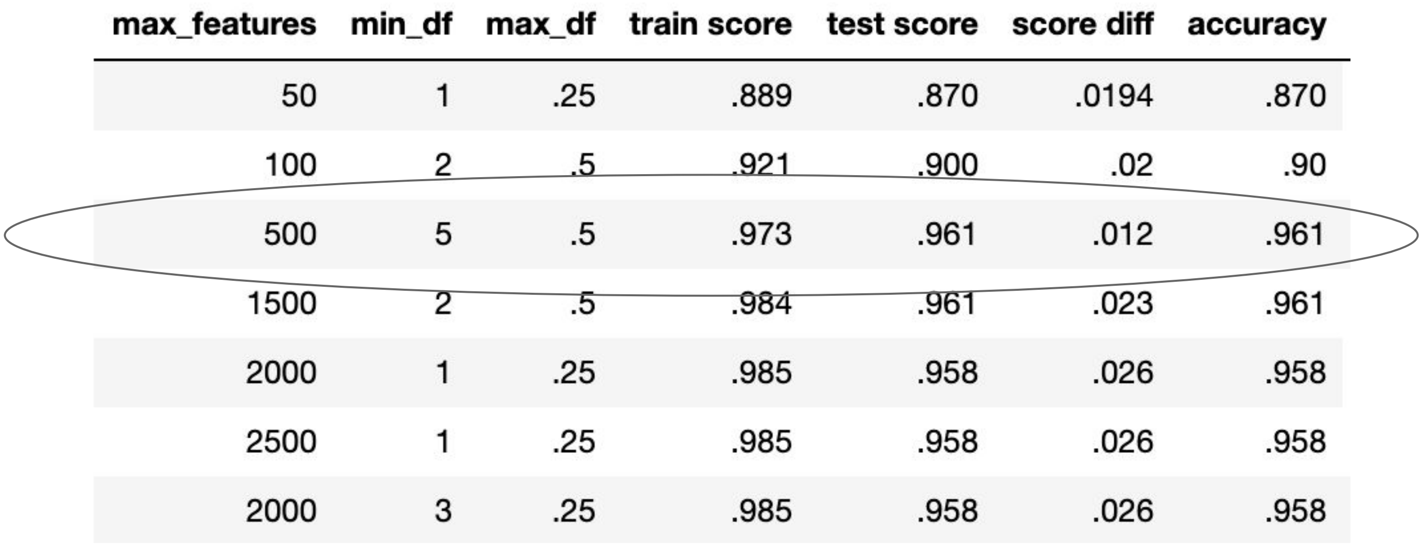
attempt	train score	test score	difference
CountVectorizer/Logistic Regression	.889	.870	.019
	.896	.885	.011
Tfidif/Naive Bayes	.974	.967	.007
	.870	.865	.005
Tfidif/KNN	.836	.807	.029
Tfidif/Random Forest	.984	.883	.101
	.957	.831	.126
Tfidif/AdaBoost	.981	.883	.098

Building and Analyzing Models

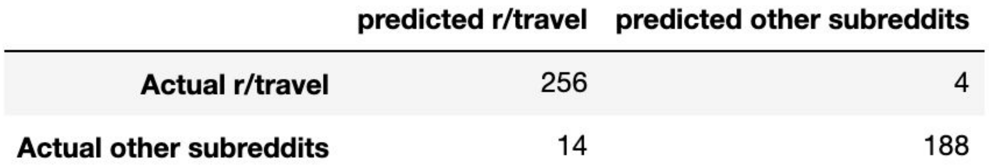
Performed additional GridSearch testing CountVectorizor with Naive Bayes and Tfidif with Logistic Regression. Common english words, known as ‘stop words’ were removed in all cases.

Best Model was Tfidif with Logistic Regression, with 500 word features.

Beat the 90% Accuracy goal!!



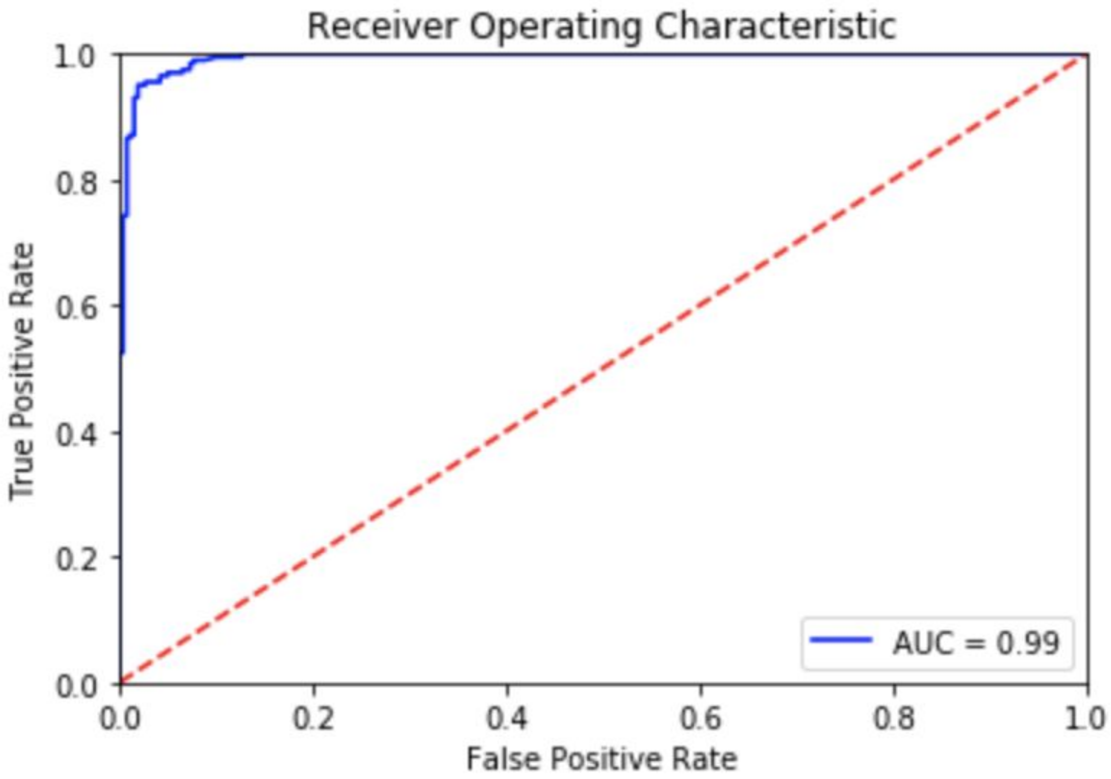
max_features	min_df	max_df	train score	test score	score diff	accuracy
50	1	.25	.889	.870	.0194	.870
100	2	.5	.921	.900	.02	.90
500	5	.5	.973	.961	.012	.961
1500	2	.5	.984	.961	.023	.961
2000	1	.25	.985	.958	.026	.958
2500	1	.25	.985	.958	.026	.958
2000	3	.25	.985	.958	.026	.958



	predicted r/travel	predicted other subreddits
Actual r/travel	256	4
Actual other subreddits	14	188

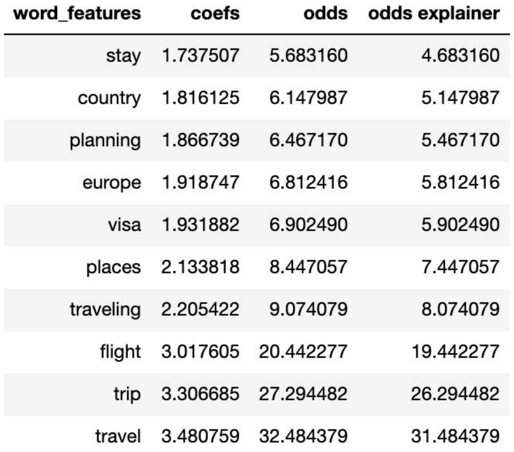
Building and Analyzing Models

The ROC Curve plots sensitivity (True Positive) and 1-Specificity (False Positive.) The area Area Under the Curve (AUC) indicates the probability that the model will score a positive value as positive.

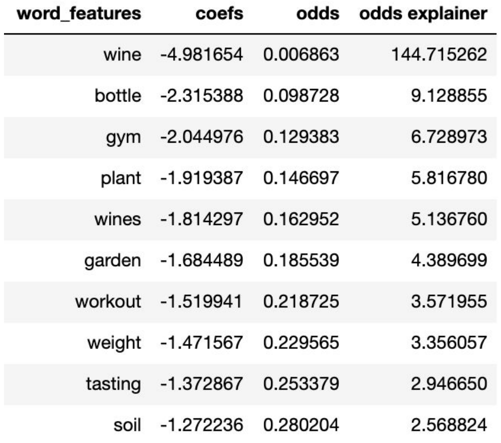


Building and Analyzing Models

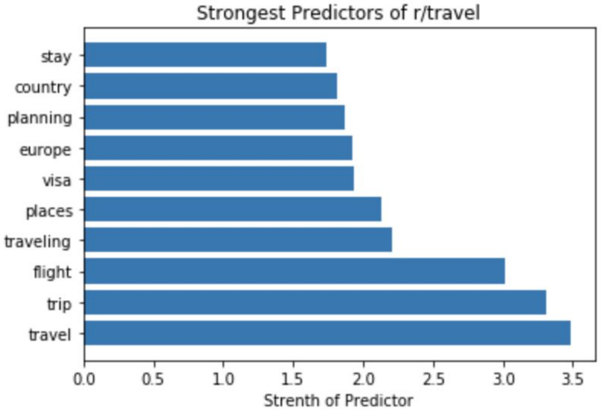
Of the 500 words used in the model, the strongest predictors for a r/travel post are shown on the left, and the strongest predictors for a non-travel post are on the right.

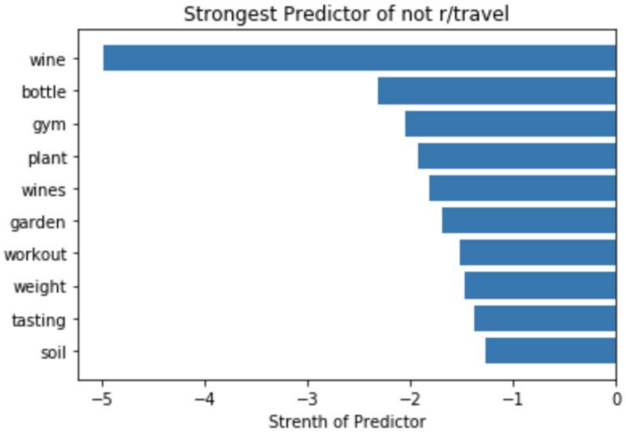


word_features	coefs	odds	odds explainer
stay	1.737507	5.683160	4.683160
country	1.816125	6.147987	5.147987
planning	1.866739	6.467170	5.467170
europe	1.918747	6.812416	5.812416
visa	1.931882	6.902490	5.902490
places	2.133818	8.447057	7.447057
traveling	2.205422	9.074079	8.074079
flight	3.017605	20.442277	19.442277
trip	3.306685	27.294482	26.294482
travel	3.480759	32.484379	31.484379



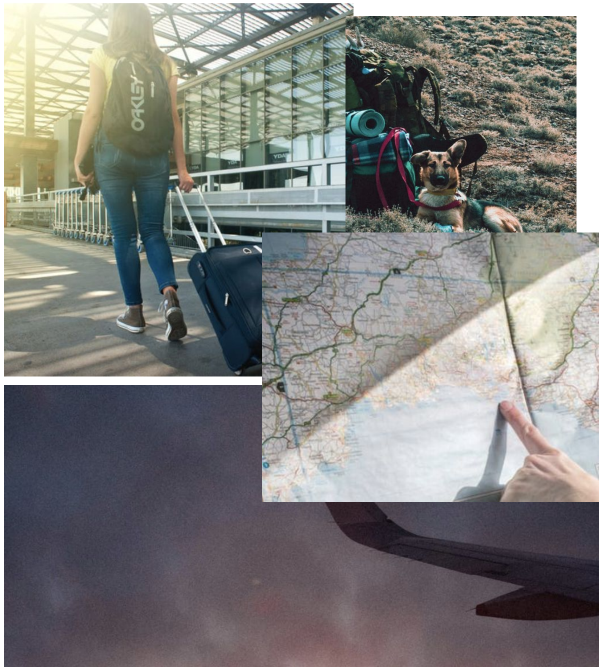
word_features	coefs	odds	odds explainer
wine	-4.981654	0.006863	144.715262
bottle	-2.315388	0.098728	9.128855
gym	-2.044976	0.129383	6.728973
plant	-1.919387	0.146697	5.816780
wines	-1.814297	0.162952	5.136760
garden	-1.684489	0.185539	4.389699
workout	-1.519941	0.218725	3.571955
weight	-1.471567	0.229565	3.356057
tasting	-1.372867	0.253379	2.946650
soil	-1.272236	0.280204	2.568824





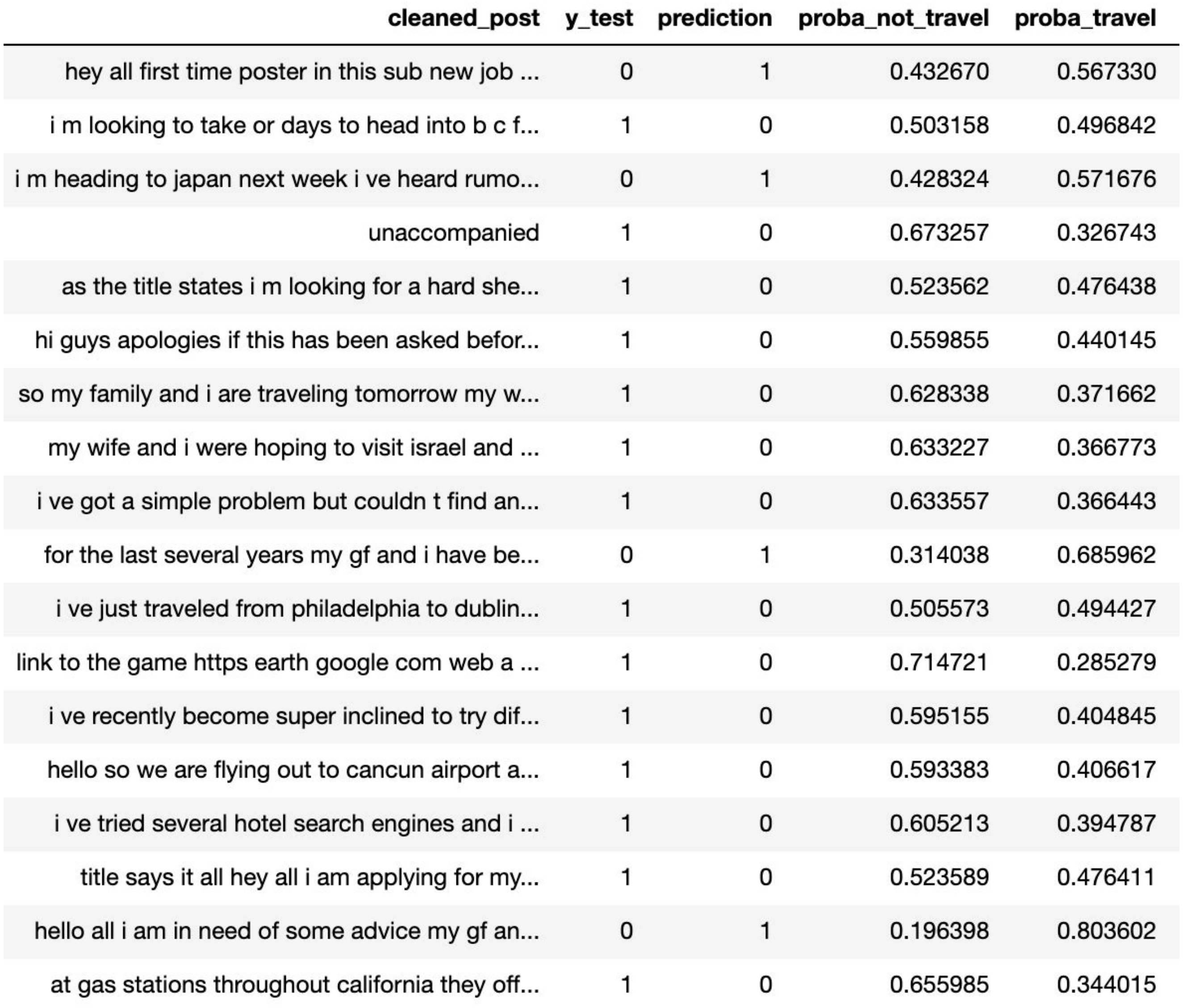
Conclusion - Success!!

- The model successfully classifies r/travel, and not r/travel, with a 96% of the time!!!
- This is not surprising, though. These topics have very distinguishing words. It may have been much harder with more similar threads.
- It DOES do a good job of illustrating the classification techniques and the modeling process.
- Recommendation: DO HIRE DATA SCIENTISTS!!!!



Just for Fun........

Misclassified Posts



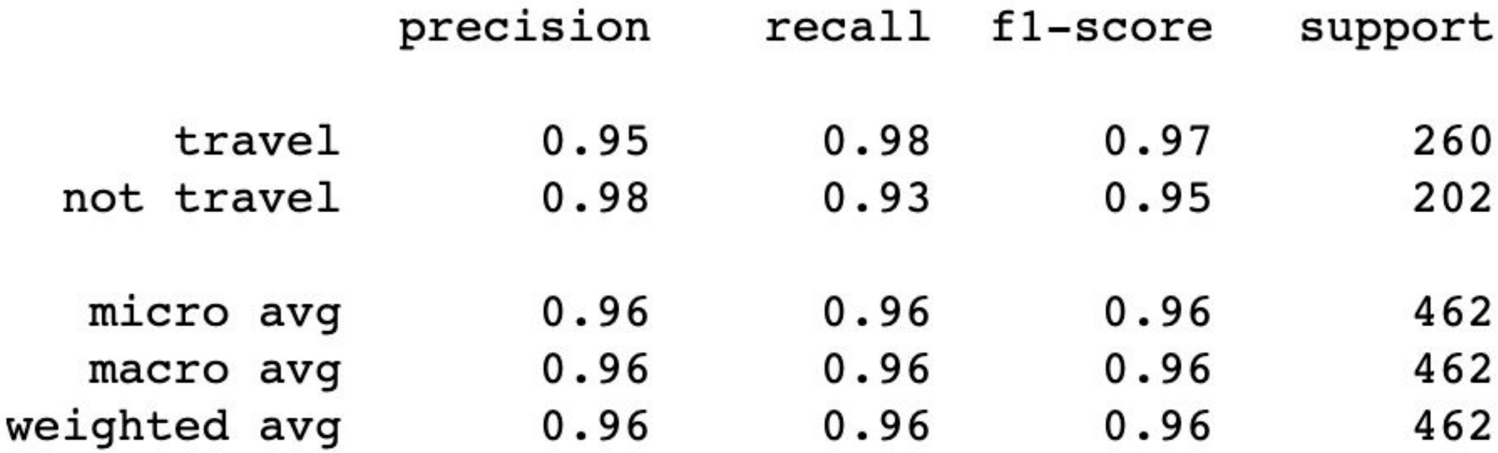
cleaned_post	y_test	prediction	proba_not_travel	proba_travel
hey all first time poster in this sub new job ...	0	1	0.432670	0.567330
i m looking to take or days to head into b c f...	1	0	0.503158	0.496842
i m heading to japan next week i ve heard rumo...	0	1	0.428324	0.571676
unaccompanied	1	0	0.673257	0.326743
as the title states i m looking for a hard she...	1	0	0.523562	0.476438
hi guys apologies if this has been asked befor...	1	0	0.559855	0.440145
so my family and i are traveling tomorrow my w...	1	0	0.628338	0.371662
my wife and i were hoping to visit israel and ...	1	0	0.633227	0.366773
i ve got a simple problem but couldn t find an...	1	0	0.633557	0.366443
for the last several years my gf and i have be...	0	1	0.314038	0.685962
i ve just traveled from philadelphia to dublin...	1	0	0.505573	0.494427
link to the game https earth google com web a ...	1	0	0.714721	0.285279
i ve recently become super inclined to try dif...	1	0	0.595155	0.404845
hello so we are flying out to cancun airport a...	1	0	0.593383	0.406617
i ve tried several hotel search engines and i ...	1	0	0.605213	0.394787
title says it all hey all i am applying for my...	1	0	0.523589	0.476411
hello all i am in need of some advice my gf an...	0	1	0.196398	0.803602
at gas stations throughout california they off...	1	0	0.655985	0.344015

Credits

- Presentation template by SlidesCarnival

Building and Analyzing Models

Sklearn provides a classification report Of the 500 words used in the model, the strongest predictors for a r/travel post are shown on the left, and the strongest predictors for a non-travel post are on the right.



```
              precision    recall  f1-score   support

      travel       0.95      0.98      0.97       260
  not travel       0.98      0.93      0.95       202

   micro avg       0.96      0.96      0.96       462
   macro avg       0.96      0.96      0.96       462
weighted avg       0.96      0.96      0.96       462
```