Project 3 - Subreddits

My Name is Cam Smugereski !

I am the Marketing Analyst for Pet Supplies Plus in Bedford, NH.

My goal is help you better market dog or cat specific to the right customers.

Problem Statement

This is a binary classification problem asking if we can predict the Subreddit, cats or dogs, based on selftext or the title, or both.

The ideal outcome is to provide the marketing team with top 'Buzz Words' they can use to directly appeal to either dog or cat lovers. We also aim to be able to identify cat or dog owners based on the way they post on Reddit.

Let’s Start With Our Most Notable Discovery:

Missing Values

Dogs

```
subreddit      0
selftext       8
title          0
created_utc    0
dtype: int64
```

Cats

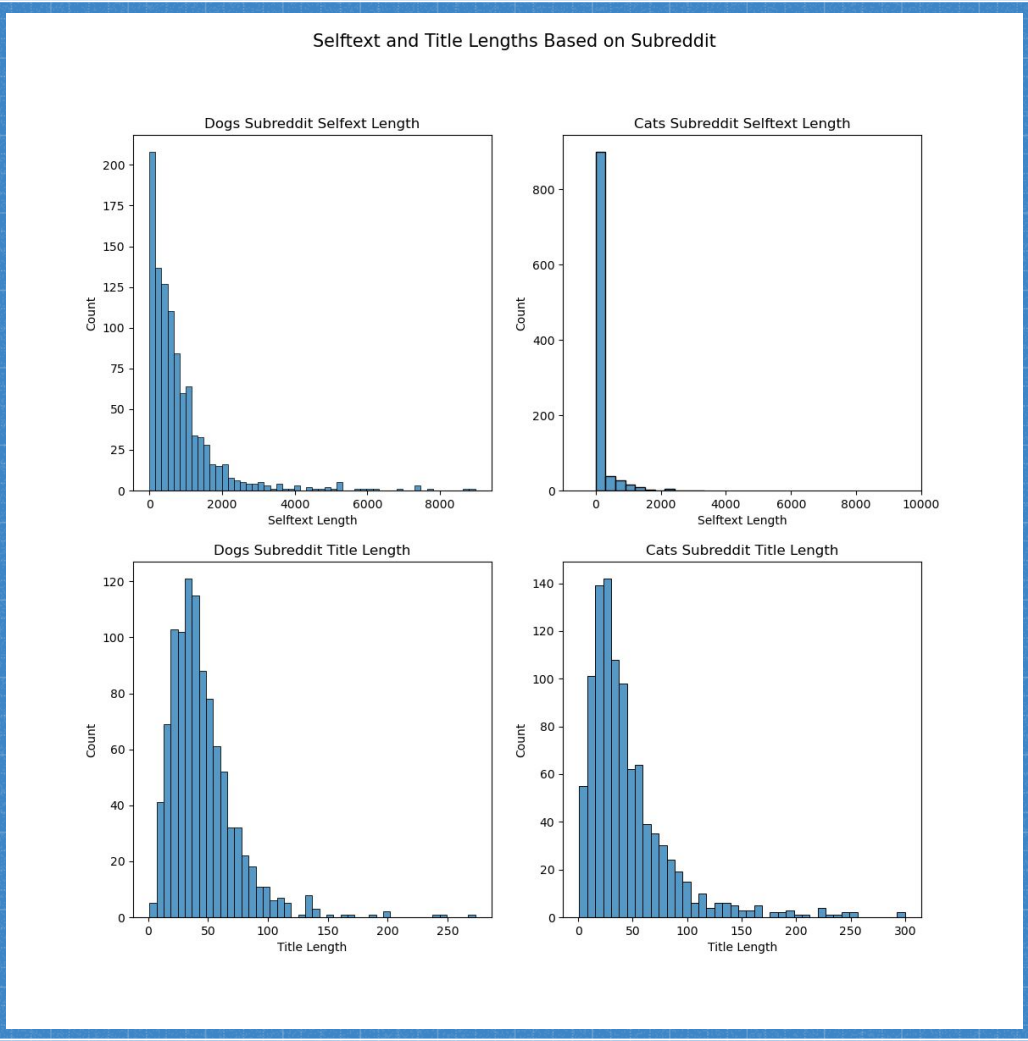
```
subreddit        0
selftext       820
title            0
created_utc      0
dtype: int64
```

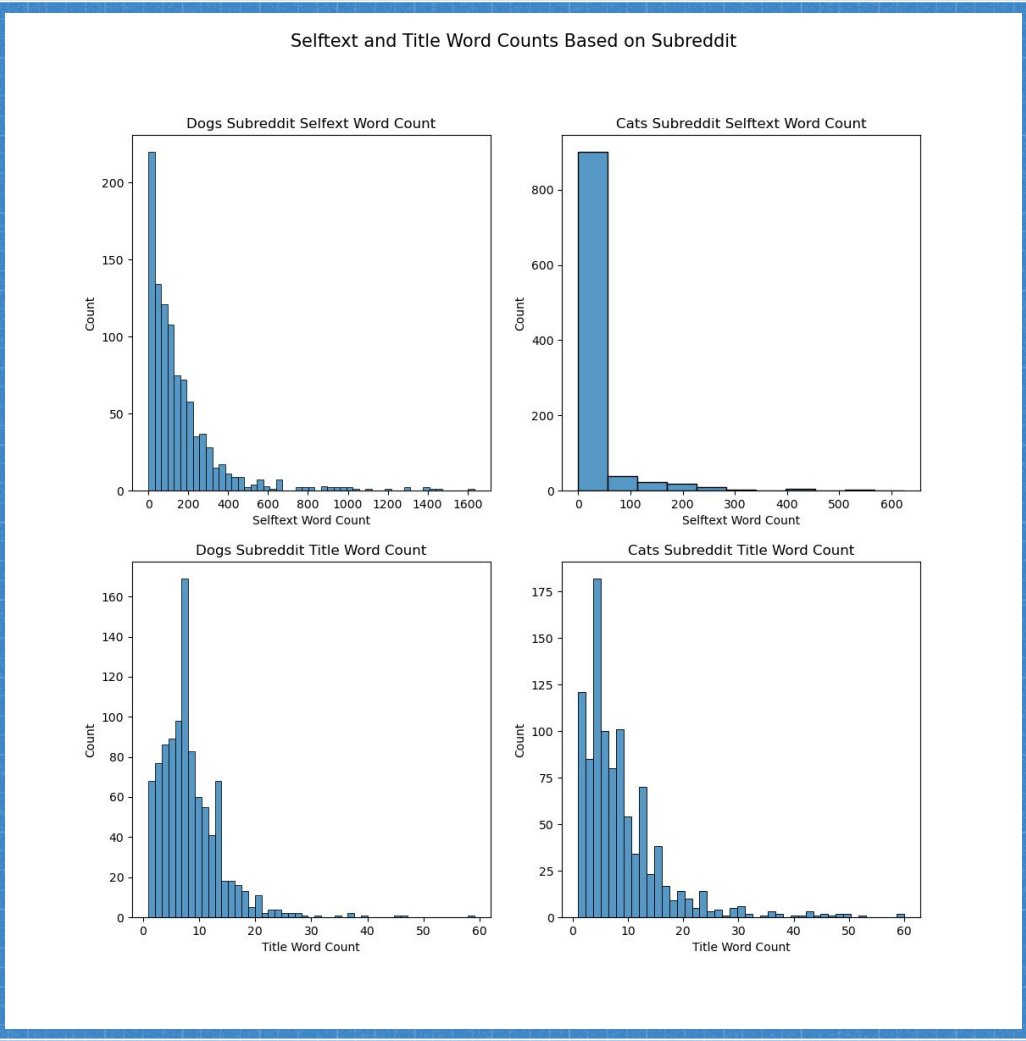
82% of our ‘cats’ Subreddit has missing values in the selftext column

Could this be an indicator of who is a cat or dog person?

What Are the Posts Telling Us?

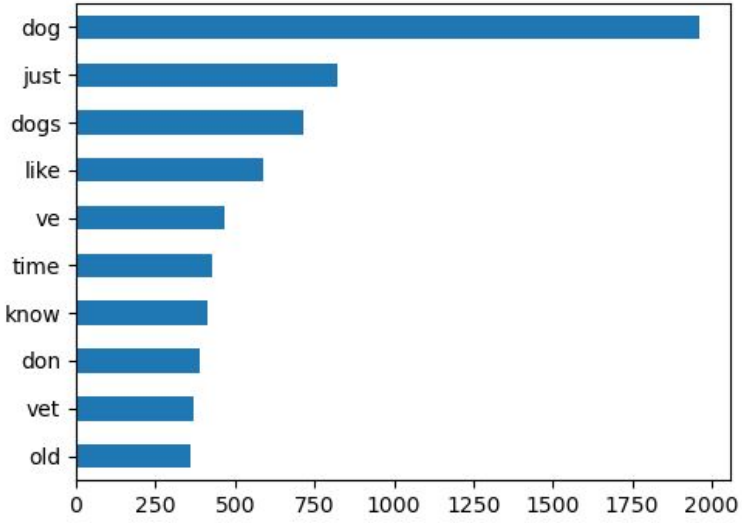
Let’s look at some simple stats!

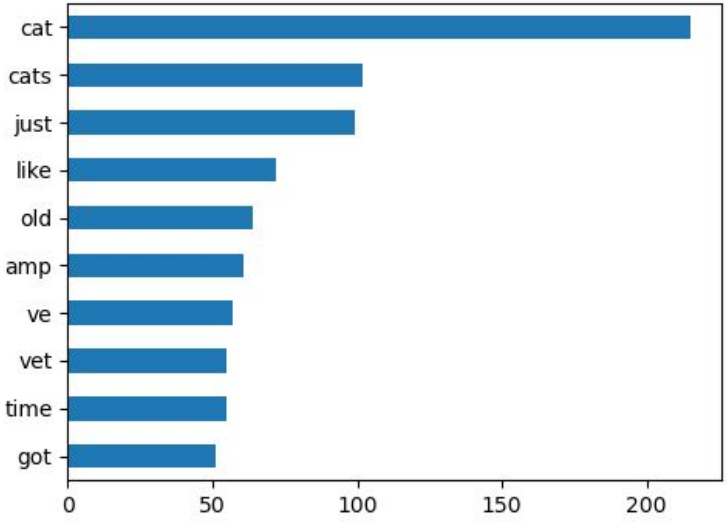

Selftext and Title Lengths Based on Subreddit
Dogs Subreddit Selfext Length
Count
Selftext Length
Cats Subreddit Selftext Length
Count
Selftext Length
Dogs Subreddit Title Length
Count
Title Length
Cats Subreddit Title Length
Count
Title Length


Selftext and Title Word Counts Based on Subreddit
Dogs Subreddit Selfext Word Count
Count
Selftext Word Count
Cats Subreddit Selftext Word Count
Count
Selftext Word Count
Dogs Subreddit Title Word Count
Count
Title Word Count
Cats Subreddit Title Word Count
Count
Title Word Count

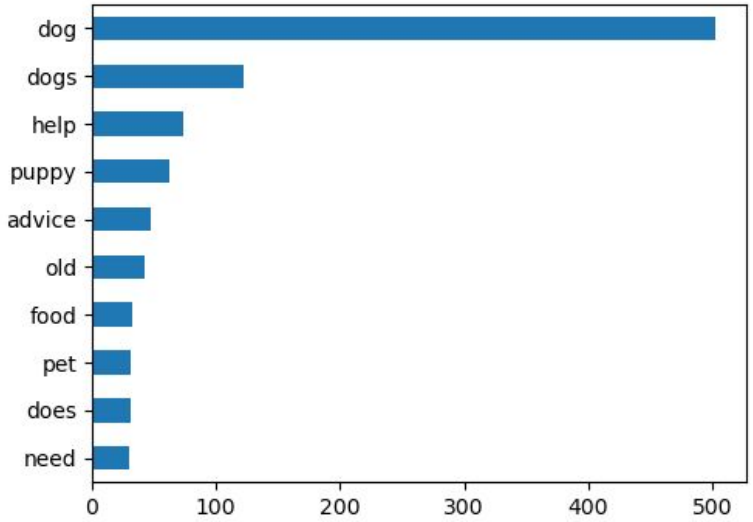
Top Words

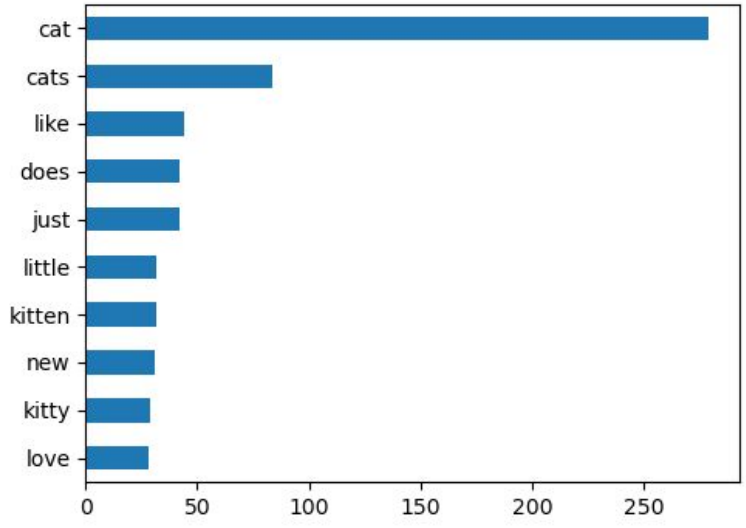
Selftext


dog
just
dogs
like
ve
time
know
don
vet
old
0
250
500
750
1000
1250
1500
1750
2000


cat
cats
just
like
old
amp
ve
vet
time
got
0
50
100
150
200

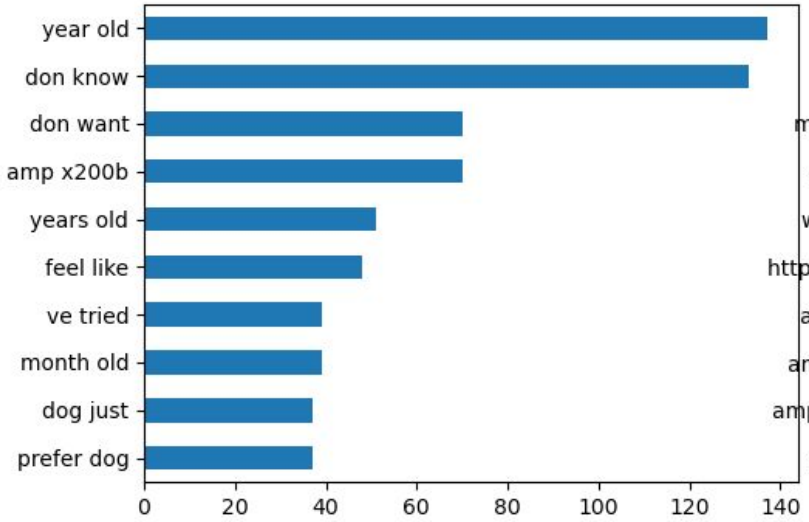
Title

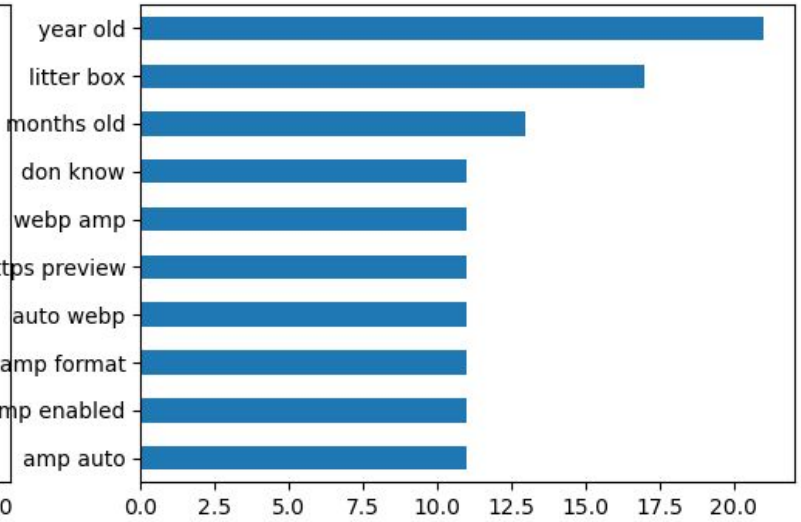

dog
dogs
help
puppy
advice
old
food
pet
does
need
0
100
200
300
400
500


cat
cats
like
does
just
little
kitten
new
kitty
love
0
50
100
150
200
250

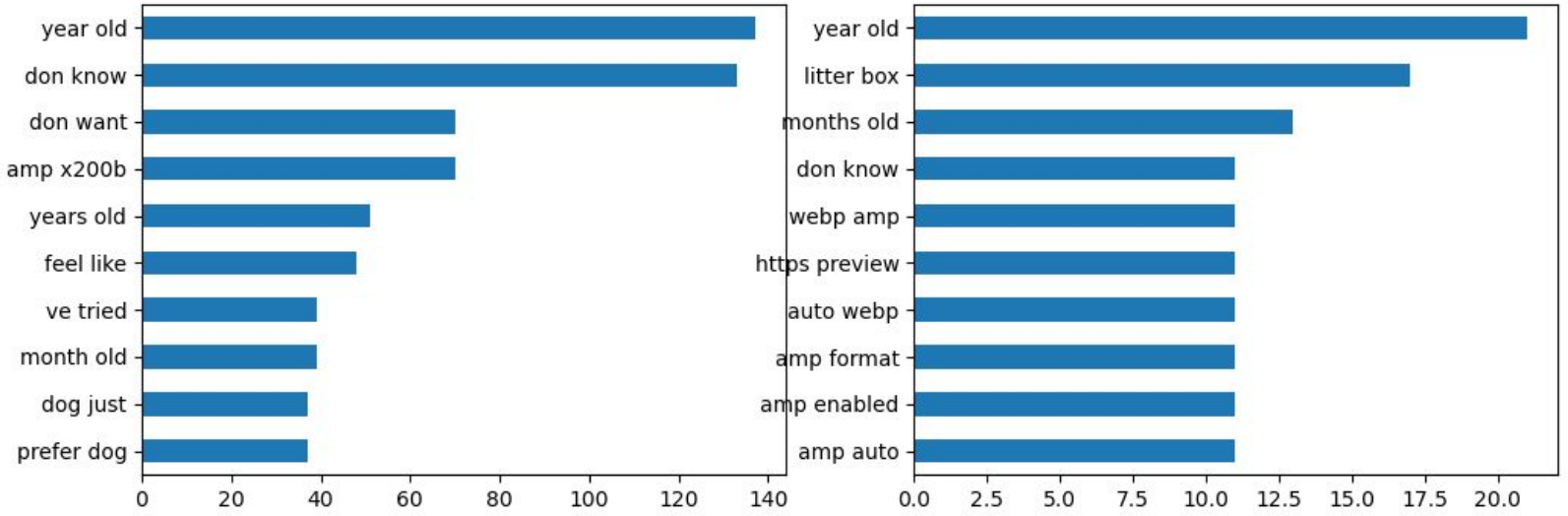
Top Bigrams

Selftext


year old
don know
don want
amp x200b
years old
feel like
ve tried
month old
dog just
prefer dog
0
20
40
60
80
100
120
140

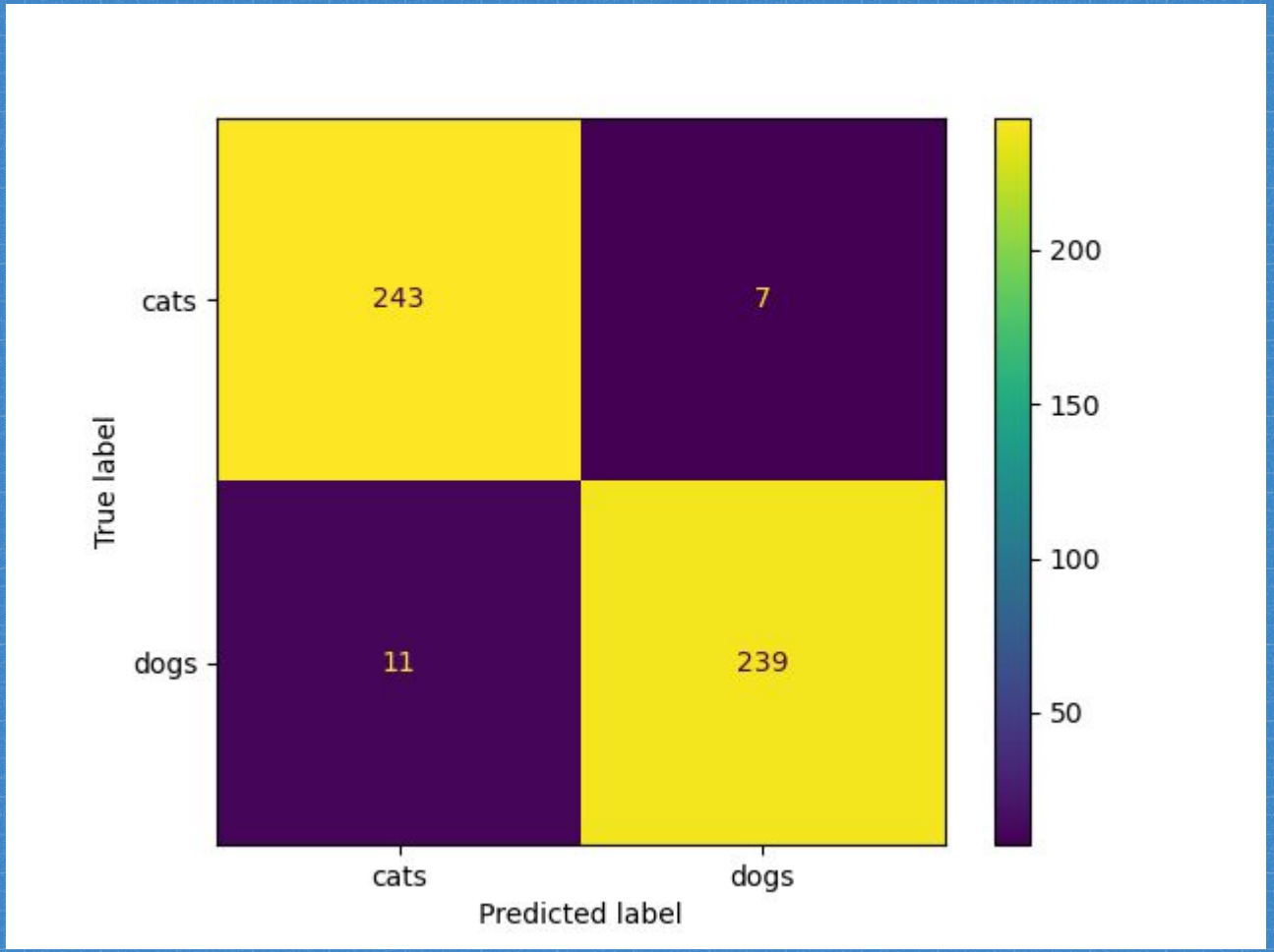

year old
litter box
months old
don know
webp amp
https preview
auto webp
amp format
amp enabled
amp auto
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0

Title


year old
don know
don want
amp x200b
years old
feel like
ve tried
month old
dog just
prefer dog
0
20
40
60
80
100
120
140
year old
litter box
months old
don know
webp amp
https preview
auto webp
amp format
amp enabled
amp auto
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0

Stacking

- Logistic Regression
- Naive Bayes
- Random Forest


243
7
11
239
cats
dogs
True label
cats
dogs
Predicted label
200
150
100
50

Training Score: 99.3%

Test Score: 96.4%

Sensitivity:95.6%

Sensitivity is our true positive rate and therefore tells us that 95% of our data that we predicted to be true, was actually true in our stacked model.

Conclusions and Recommendations

Takeaways

1. 82% of our ‘cats’ selftext is missing - What does this tell us?

2. I recommend using visuals for marketing cat products

3. Marketing for dog products can and should include some ‘Buzz Words’

4. Our model and sensitivity scores are high, but this is likely due to how different each Subreddit is

Any Questions?