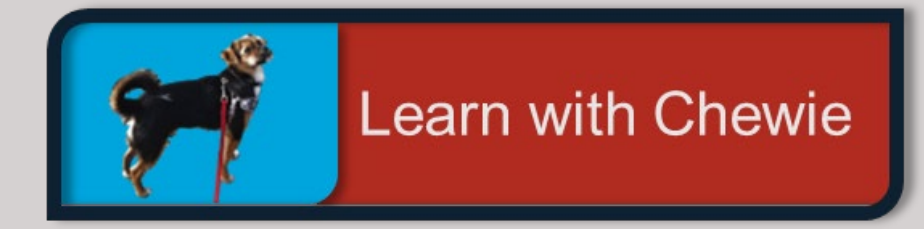


PRESENTS

WEB APIS & NLP

BRANDIE HATCH

CHEWIE

Ten-pound rescue pup who loves everyone and plans to be a therapy dog once the AKC Canine Good Citizen test is passed.

Video of Chewie learning to high five

Learning how to High Five

PROBLEM & GOALS

BUILD A CLASSIFICATION MODEL THAT CAN BE USED TO IDENTIFY SERVICE DOGS VS DOG TRAINING

Service dogs require a lot of training.

- What are the classification opportunities between the subreddits for Service Dogs and Dog Training?
- How can those classifications help to predict training opportunities for service and other dogs?

BACKGROUND

Attempts at this particular study have not been made in the past.

By the end of the presentation today, the goal is to communicate:

- investigative process
- inferences
- recommendations

DATA COLLECTION

Data Pulled from Subreddit:

1. The Dog Training Subreddit data set included 4730 observations of nine variables.
2. The Service Dogs Subreddit data set included 4717 observations of nine variables.

Variable	Notes
Subreddit	Classification tool
Title	Text to be vectorized and used in model
Self text	Text to be vectorized and used in model
Number of Comments	Used as numerical tool in model
Score	Used as numerical tool in model
Over_18	Helped to remove inappropriate content (removed)
Author	Collected in case of concerns (removed)
Created UTC	Utilized for collecting most recent data (removed)

EDA/CLEANING

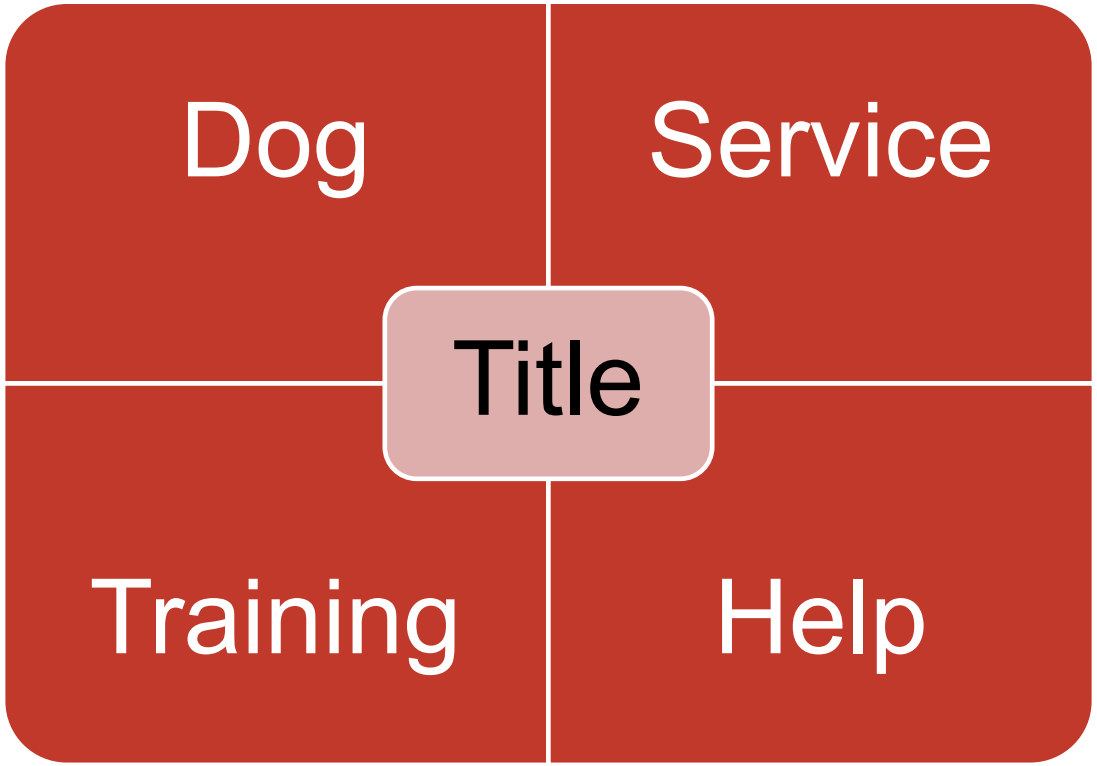
- During cleaning, no nulls that needed to be removed
- There was not much difference between the scores of the two subreddits:
  - Service Dogs mean score was 1.5
  - Dog Training mean score was 1.0
- There was a significant difference in the number of comments for each subreddit:
  - Service Dogs number of comments mean was 9.2
  - Dog Training number of comments mean was .8

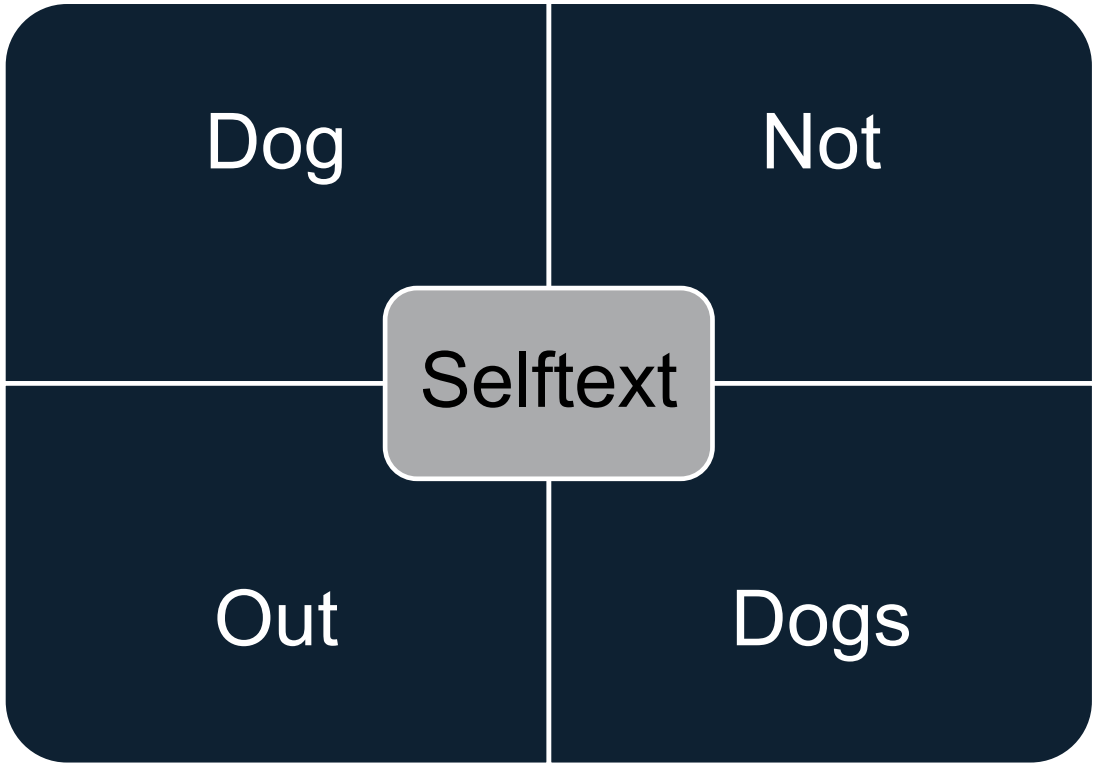
  *Number of comments will provide a good numerical predictor in the model
- Class imbalance was almost non-existant:

Service Dogs	0.499576
Dog Training	0.500424

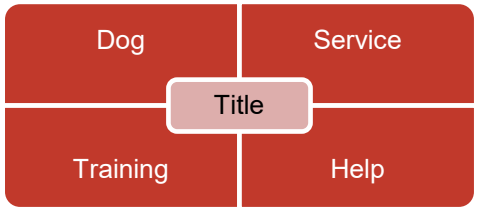
- Combined Title and Self text into a single column for Vectorization, Lemmatizing, Stemming, and Tokenizing

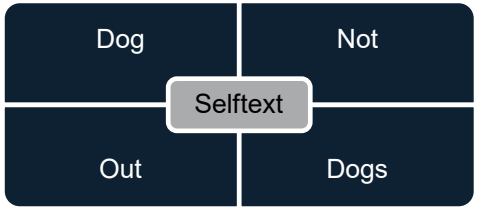
DATA EXPLORATION: MOST COMMON WORDS

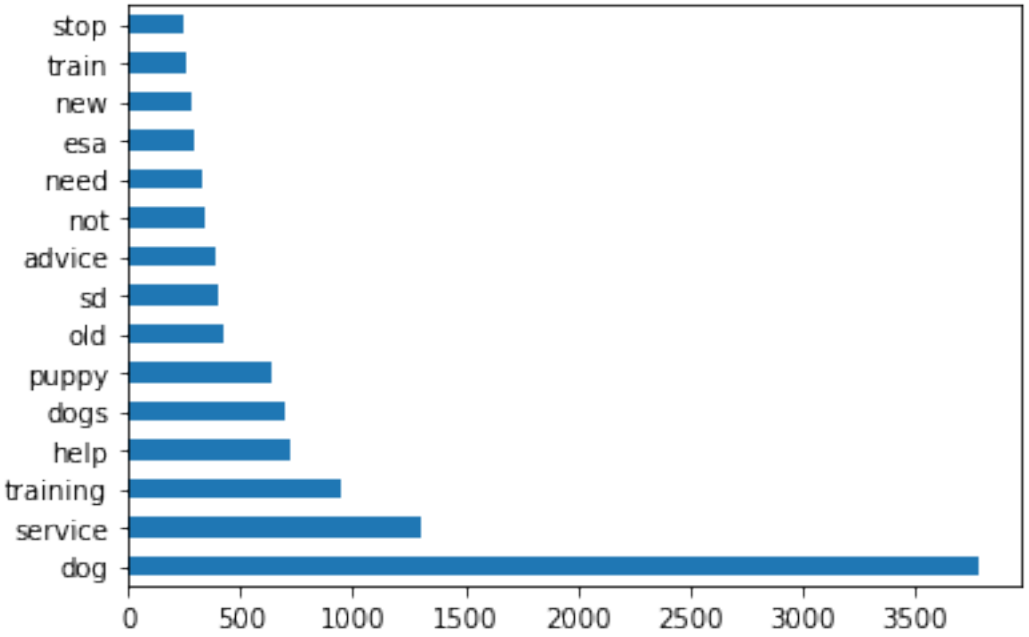


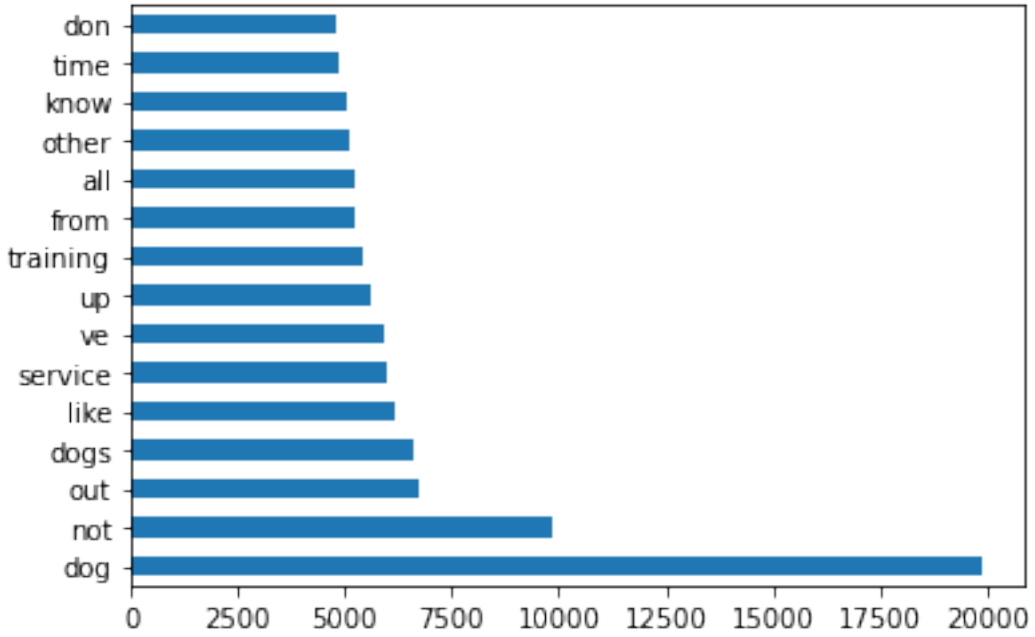


DATA EXPLORATION: MOST COMMON WORDS

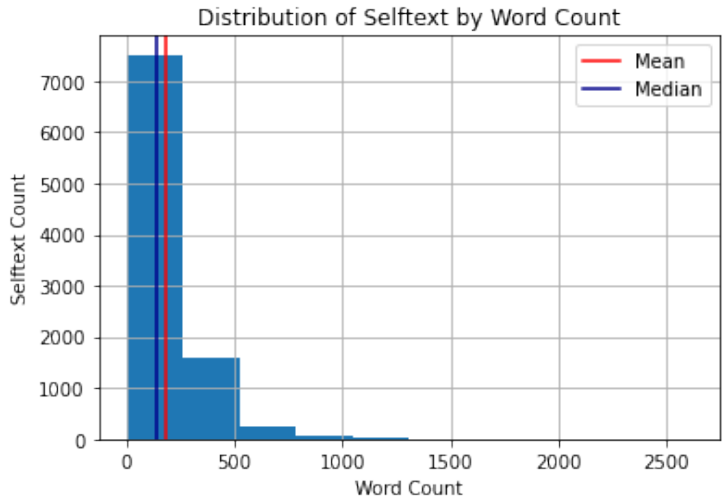


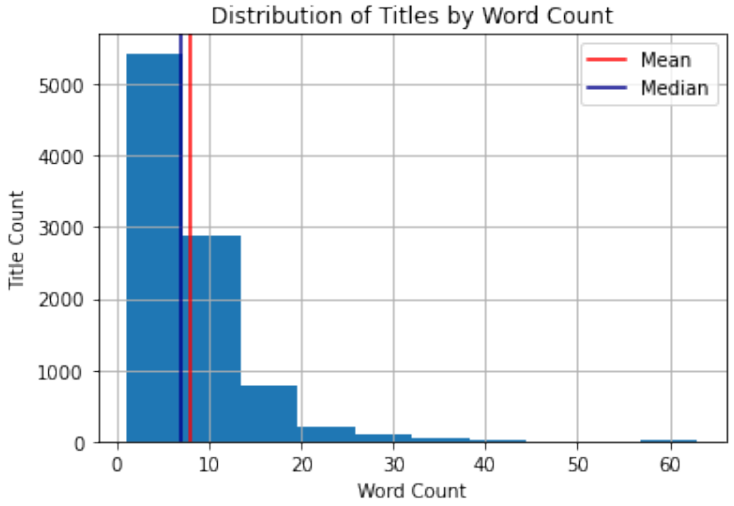


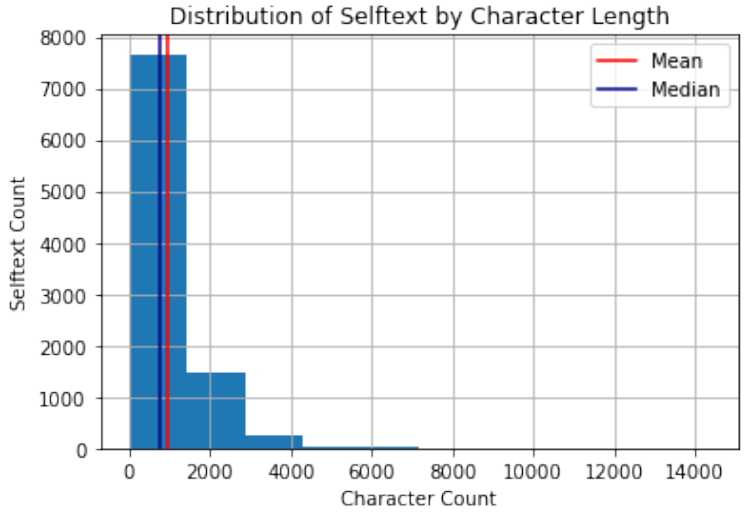


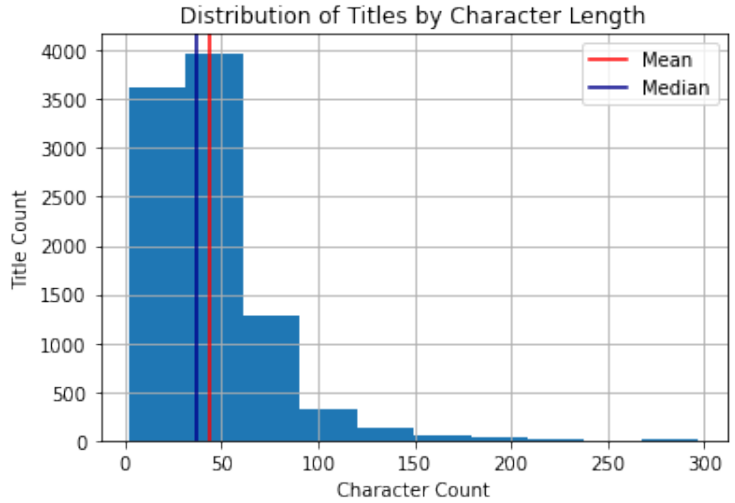


DATA EXPLORATION: LONGEST & SHORTEST



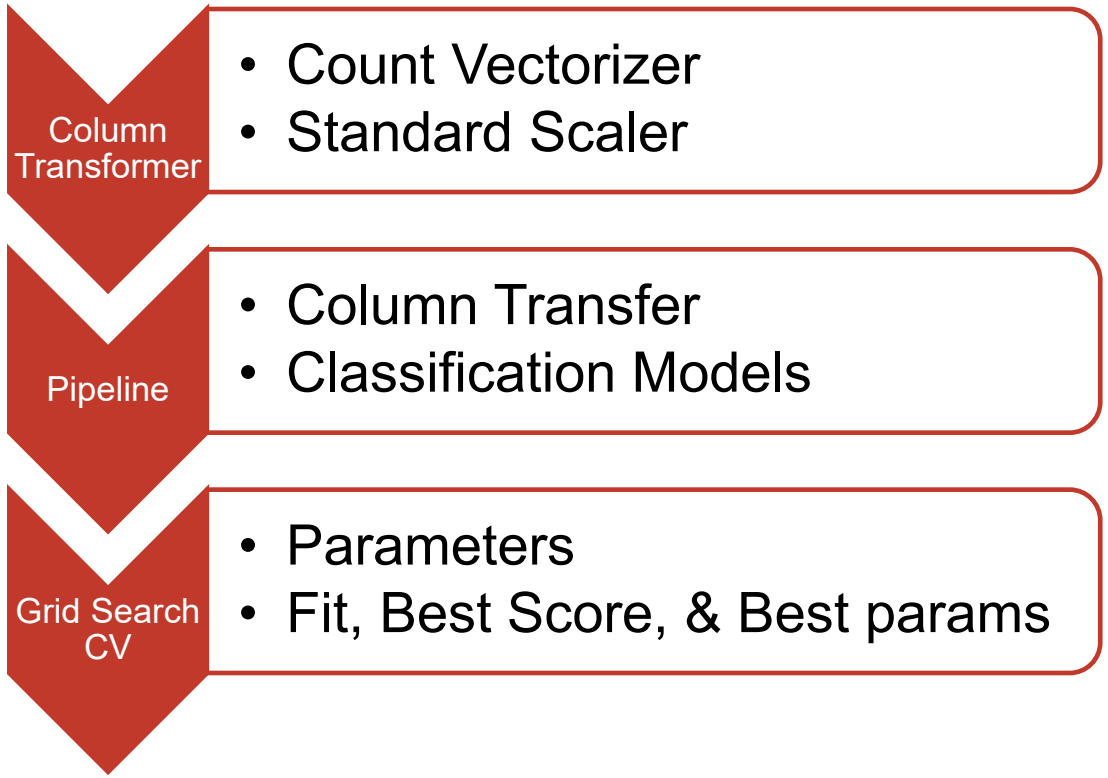






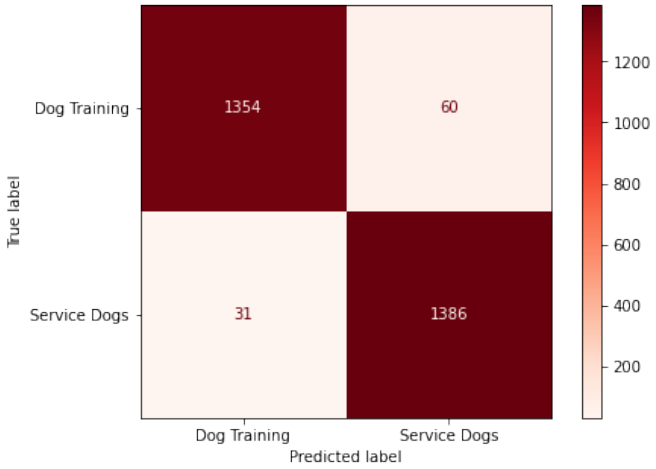
INVESTIGATIVE PROCESS

Models	Best Scores
Random Forest Classifier	0.9554882664647995
Decision Tree Classifier	0.953217259651779
Logistic Regression	0.9654806964420892
SVM: Linear SVC	0.9609386828160484
SVM: C-Support Vector Classification	0.8996214988644965

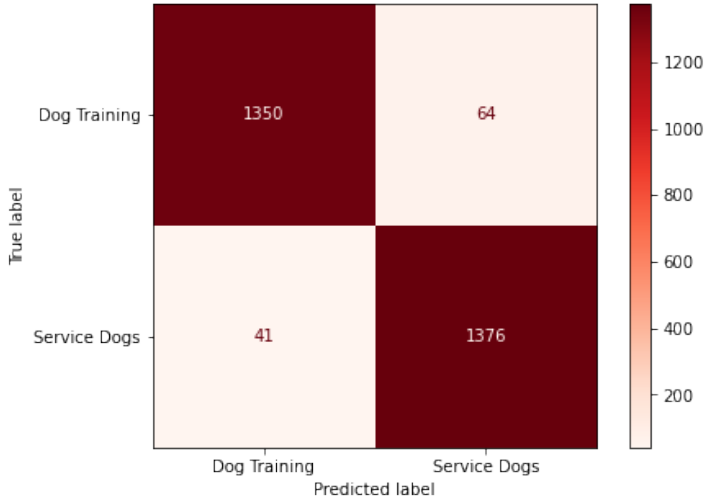


CONFUSION MATRIX DISPLAYS

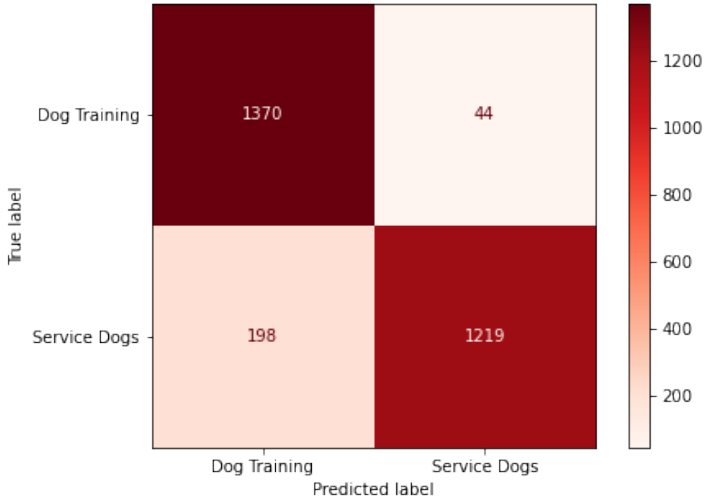
Logistic Regression



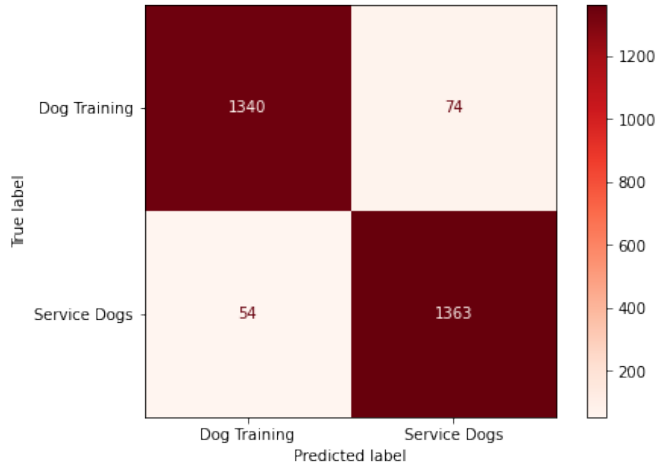
SVM: Linear SVC



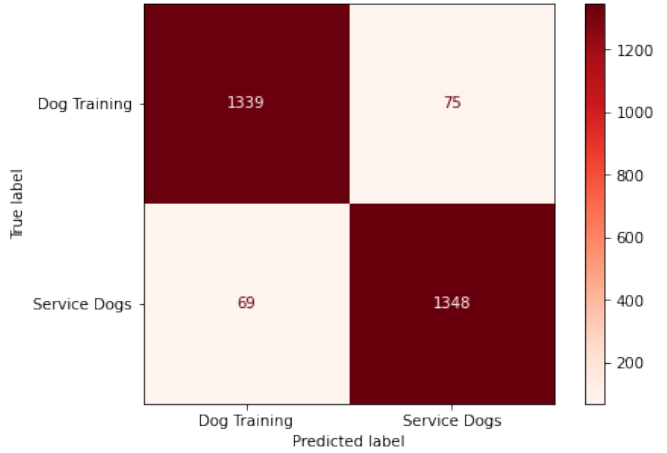
SVM: C-Support Vector Classification



Random Forest Classifier



Decision Tree Classifier



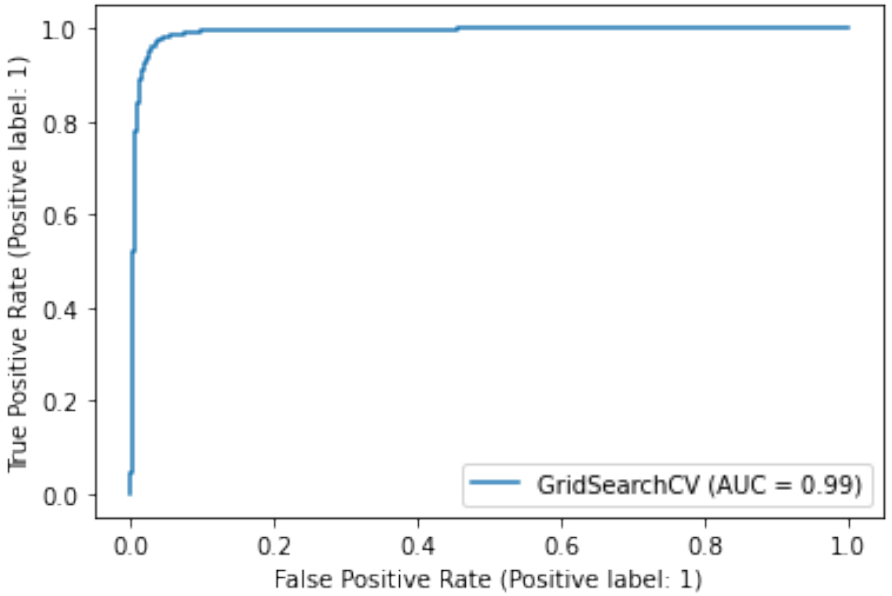
MODEL FINDINGS

- Overall, the Logistic Regression model was the best model to identify Service Dogs vs Dog Training Subreddits.
- The SVM: C-Support Vector Classification model could also be used.

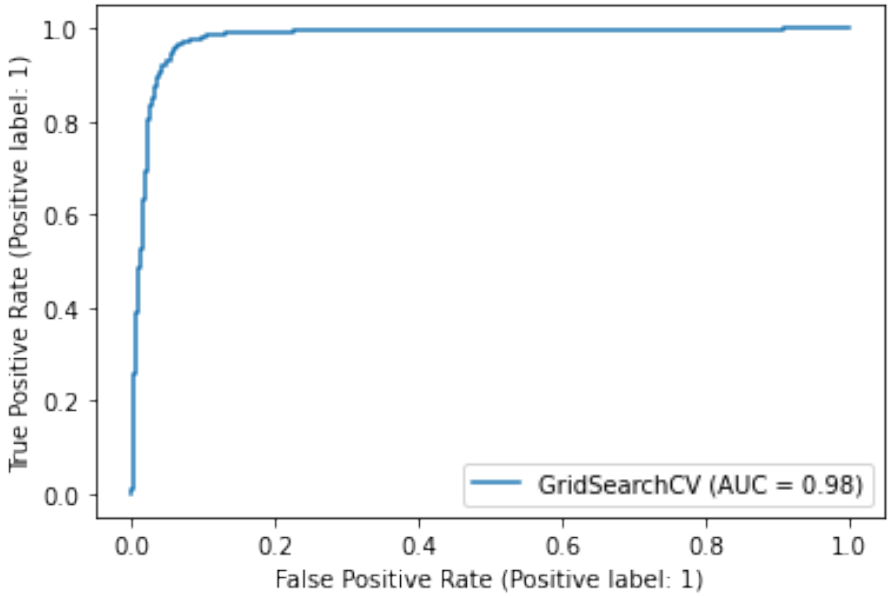
Metric	Best Model
Bias/Variance Tradeoff	Logistic Regression, SVM: C-Support Vector Classification
Accuracy	Logistic Regression
Precision	SVM: C-Support Vector Classification
Recall	Logistic Regression
Specificity	SVM: C-Support Vector Classification
F1 Score	Logistic Regression

ROC CURVES

Logistic Regression



SVM: C-Support Vector Classification



RECOMMENDATIONS

Problem Statements:

- What are the classification opportunities between the subreddits for Service Dogs and Dog Training?
- How can those classifications help to predict training opportunities for service and other dogs?

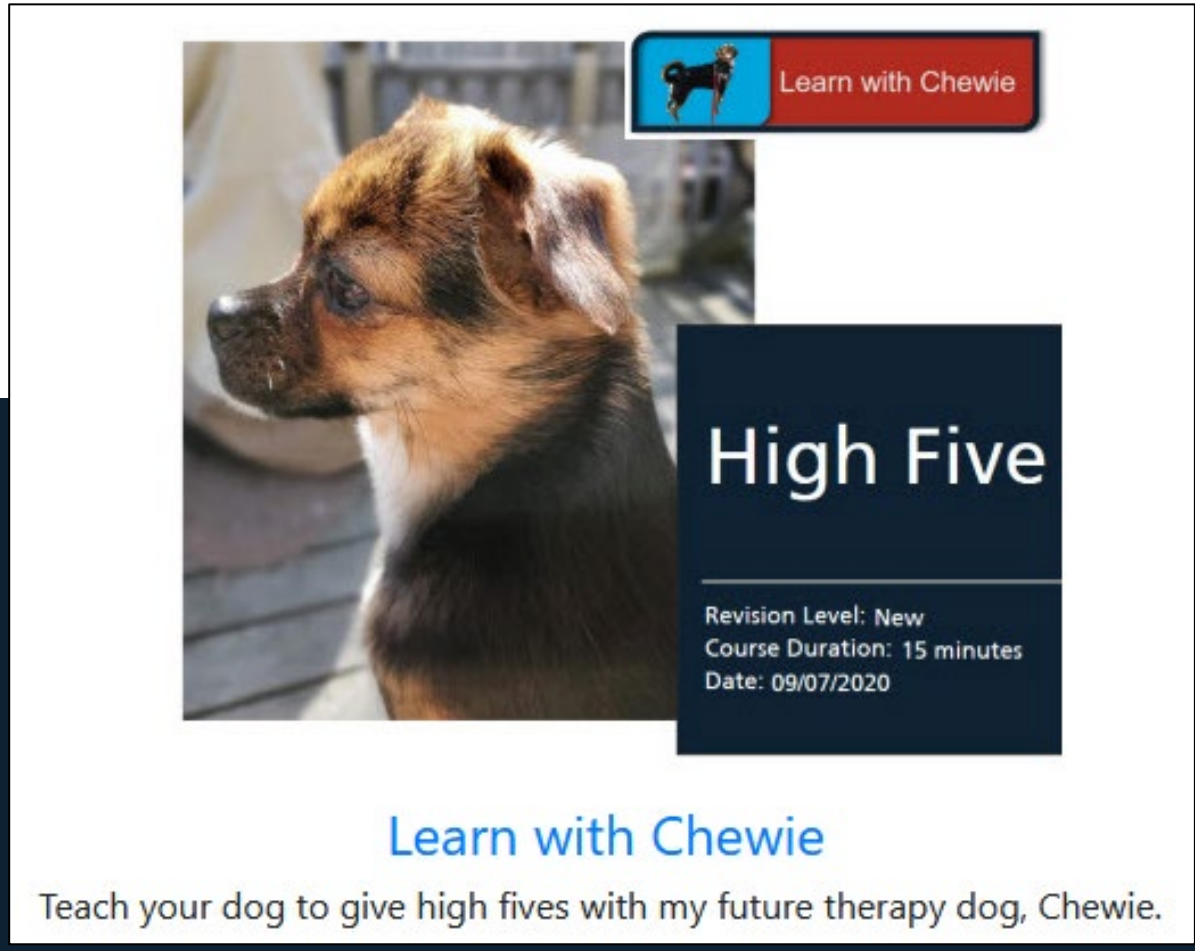
Implications and Next Steps:

- Logistic Regression model was most accurate
- Did not help predict training opportunities for service or other dogs
- Classification opportunities could be to look more into the most common words and compare the differences between each Subreddit
- Next steps would include specifically looking for training theories and pedagogy related words

QUESTIONS

Follow Chewie's Adventures:
https://www.instagram.com/chewie_honeybadger/

Connect with Brandie:
/in/brandiehatch • Github/brandiehatch



SOURCES

- https://towardsdatascience.com/20-popular-machine-learning-metrics-part-1-classification-regression-evaluation-metrics-1ca3e282a2ce
- https://www.tablesgenerator.com/markdown_tables#
- API Pushshift
- https://www.reddit.com/r/service_dogs/
- https://www.reddit.com/r/Dogtraining/

INDEX

DATA DICTIONARY

Features used listed below:

Feature	Type	Dataset	Description
subreddit	*object*	df	Subreddit Name (instance of Subreddit)
title	*object*	df	Title of submission
selftext	*object*	df	Selftext of a submission (an empty string if a link post)
author	*object*	df	Author (Redditor) of the submission
name	*object*	df	Full ID of submission, prefixed with t4_
score	*int64*	df	Total points for a submission
num_comments	*int64*	df	Number of comments on the submission

Created with: https://www.tablesgenerator.com/markdown_tables#

Variable	Model	Best Score (gridsearch)	Train Score	Test Score	Accuracy	ROC Accuracy	Precision	Recall	Specificity	F1 Score
gs	Random Forest Classifier	0.952763058289175	1.000000000000000	0.953020134228187	0.953020134228187	0.953012220770418	0.946453407510431	0.960479887085391	0.945544554455445	0.953415061295972
tree_gs	Decision Tree Classifier	0.953217259651779	1.000000000000000	0.949134581419993	0.949134581419993	0.949132278385616	0.947294448348559	0.951305575158786	0.946958981612447	0.949295774647887
log_gs	Logistic Regression	0.965480696442089	0.999848599545798	0.967855881314023	0.967855881314023	0.967844989963256	0.958506224066390	0.978122794636556	0.957567185289957	0.968215158924205
gs_svm	SVM: Linear SVC	0.960938682816048	1.000000000000000	0.962910632285411	0.962910632285411	0.962901981296022	0.955555555555555	0.971065631616090	0.954738330975954	0.963248162408120
gs_svc	SVM: C-Support Vector Classification	0.899621498864496	0.933080999242997	0.914517838219710	0.914517838219710	0.914575387370373	0.965162311955661	0.860268172194777	0.968882602545968	0.909701492537313

Variable	Model	Best Score	Train Scor	Test Score	Accuracy	ROC Accur	Precision	Recall	Specificity	F1 Score
gs	Random F	0.952763	1	0.95302	0.95302	0.953012	0.946453	0.96048	0.945545	0.953415
tree_gs	Decision T	0.953217	1	0.949135	0.949135	0.949132	0.947294	0.951306	0.946959	0.949296
log_gs	Logistic Re	0.965481	0.999849	0.967856	0.967856	0.967845	0.958506	0.978123	0.957567	0.968215
gs_svm	SVM: Line	0.960939	1	0.962911	0.962911	0.962902	0.955556	0.971066	0.954738	0.963248
gs_svc	SVM: C-Su	0.899621	0.933081	0.914518	0.914518	0.914575	0.965162	0.860268	0.968883	0.909701