Model Mis-cat-ifications

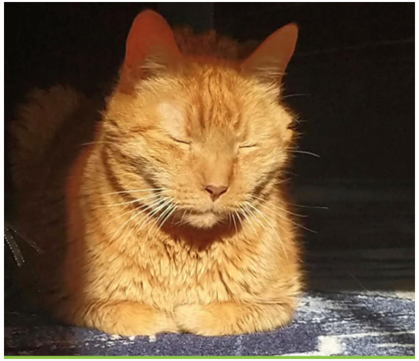
Katharine King

Problem Statement

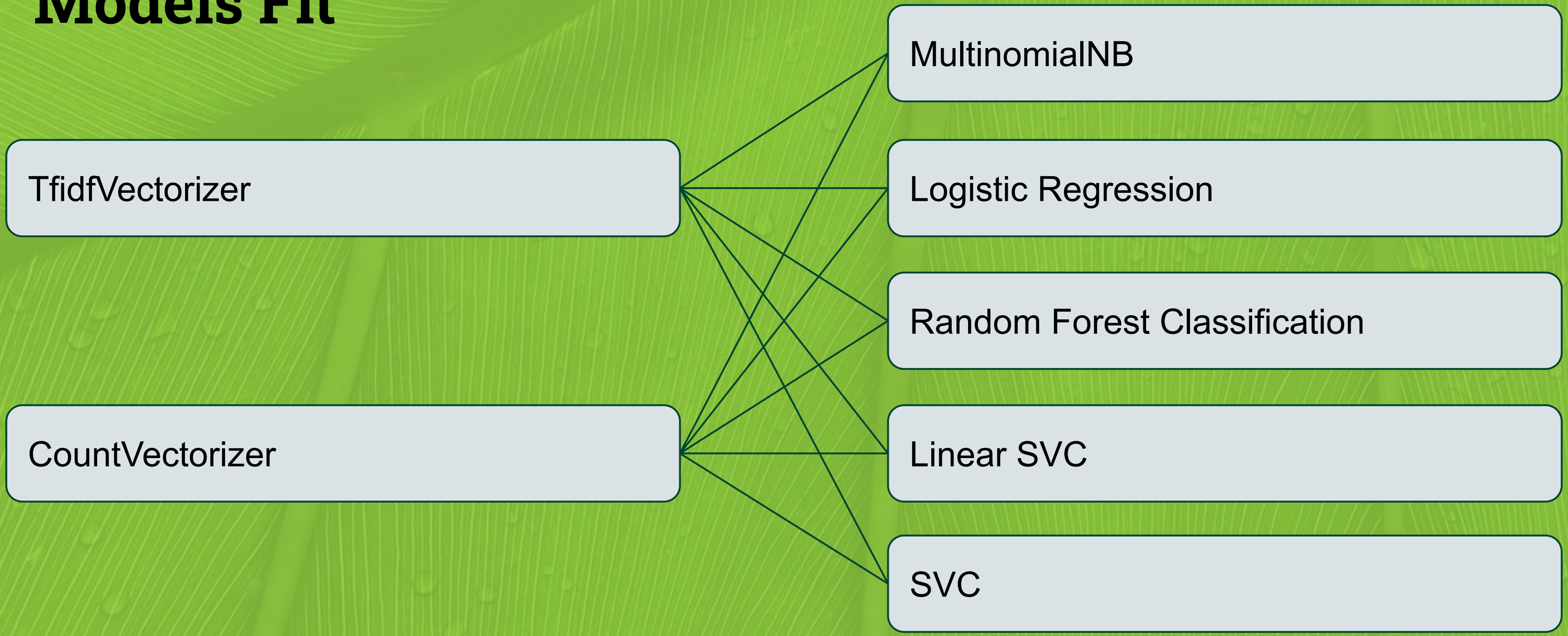
Do varying classification models misidentify similar subsets of data? Or do subsets of misidentified values vary by model?

Reddits

- Things that you take care of for no reason
  - r/Plants
  - r/Cats
- Mainly photo posts
- Pulled 5000 posts/ subreddit



Models Fit



TfidfVectorizer

CountVectorizer

MultinomialNB

Logistic Regression

Random Forest Classification

Linear SVC

SVC

Misclassification Rates

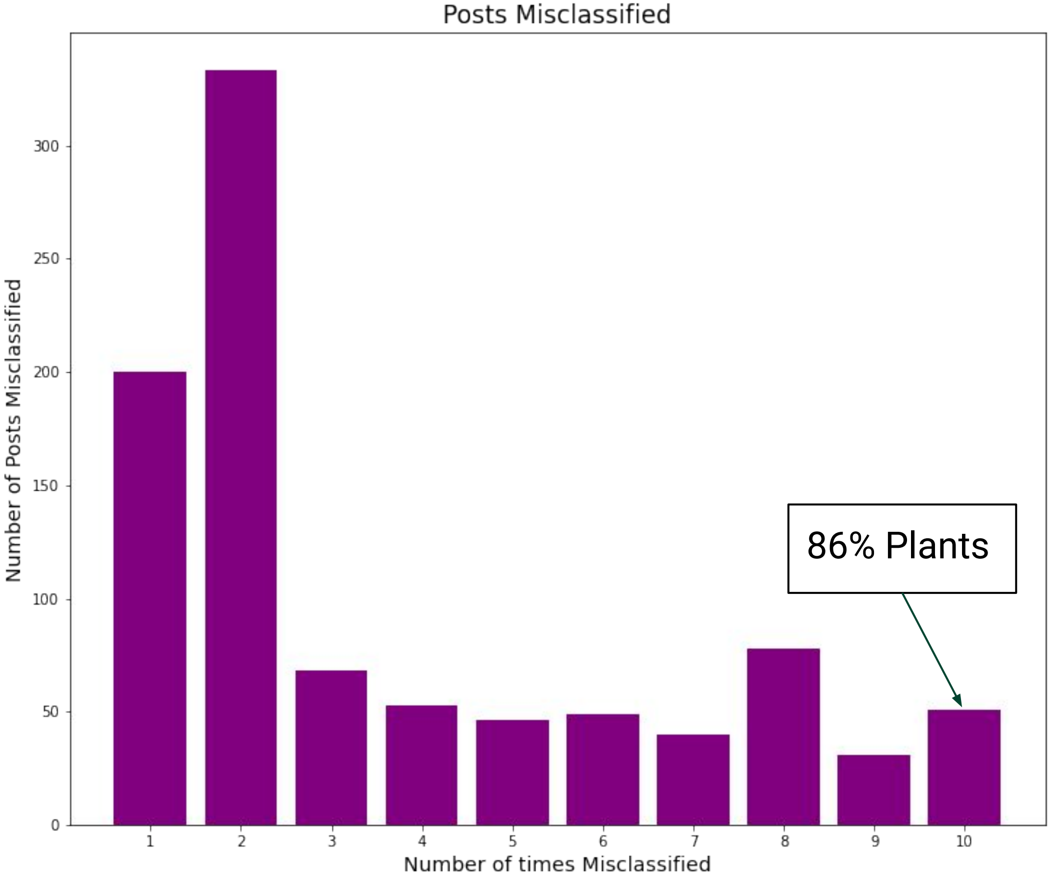
	TfidfVectorizer	CountVectorizer
MultinomialNB	9.85%	9.18%
Logistic Regression	8.88%	8.45%
Random Forest Classification	17.52%	16.48%
Linear SVC	8.82%	9.85%
SVC	7.82%	9.18%

Model Analysis

- Combination of TfidfVectorizer and Support Vector Classification
- Score .923
- Specificity = 90.36%
    - How many plant posts were correctly identified?
- Sensitivity = 94%
    - How many cat posts were correctly identified?
- Misclassification rate = 7.82%

Misclassification Frequency

- 2351 posts never misclassified
- 51 posts misclassified by every model!



86% Plants

All my favorite babies on one table 😍

Quote - 1 🌱

Suzie loves to adopt new pets and I cannot refuse.

Quote - 2 🌱

Leaf boop

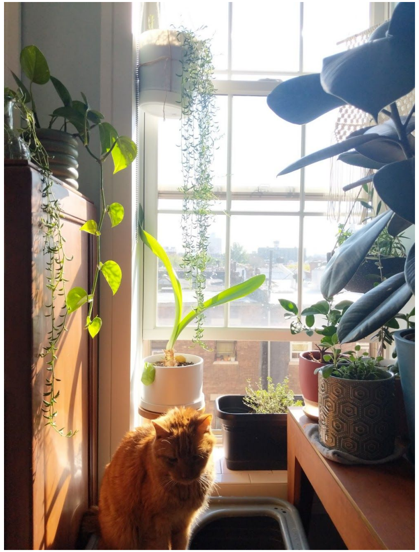
Quote - 3

A vicious jungle cat checking out her territory after a fresh rain

Quote - 4 🌱

Conclusions

- Models to not always error in the same way:
  - ~9% of the data misclassified at any one time by a model
  - 29% misclassified by at least one model



Resources

- Reddits
  - r/Plants
  - r/Cats
- API