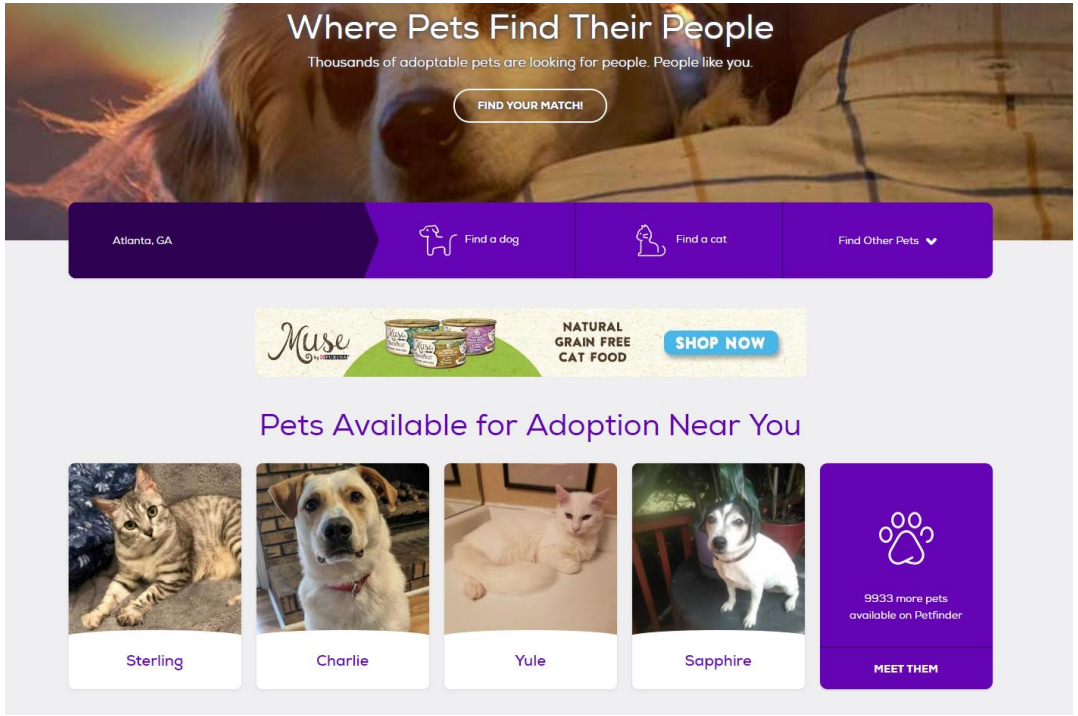
Can an animal's adoption be predicted from its PetFinder profile?

Thinkful Unit 3 Capstone

Julia Raykin

The Problem

- Millions of stray animals are in shelters in dangers of being euthanized worldwide (World Health Organization)
- Petfinder – possible solution?
  - Brings together data on animals in local animal shelters
- Research Question – how do we improve Petfinder profiles and increase adoption rates in shelters?



Data Source

- Data from PetFinder.com will be analyzed to determine how an animal's PetFinder profile affects the rates at which animals get adopted.
- The dataset was obtained from Kaggle (https://www.kaggle.com/c/petfinder-adoption-prediction/data)

Goals

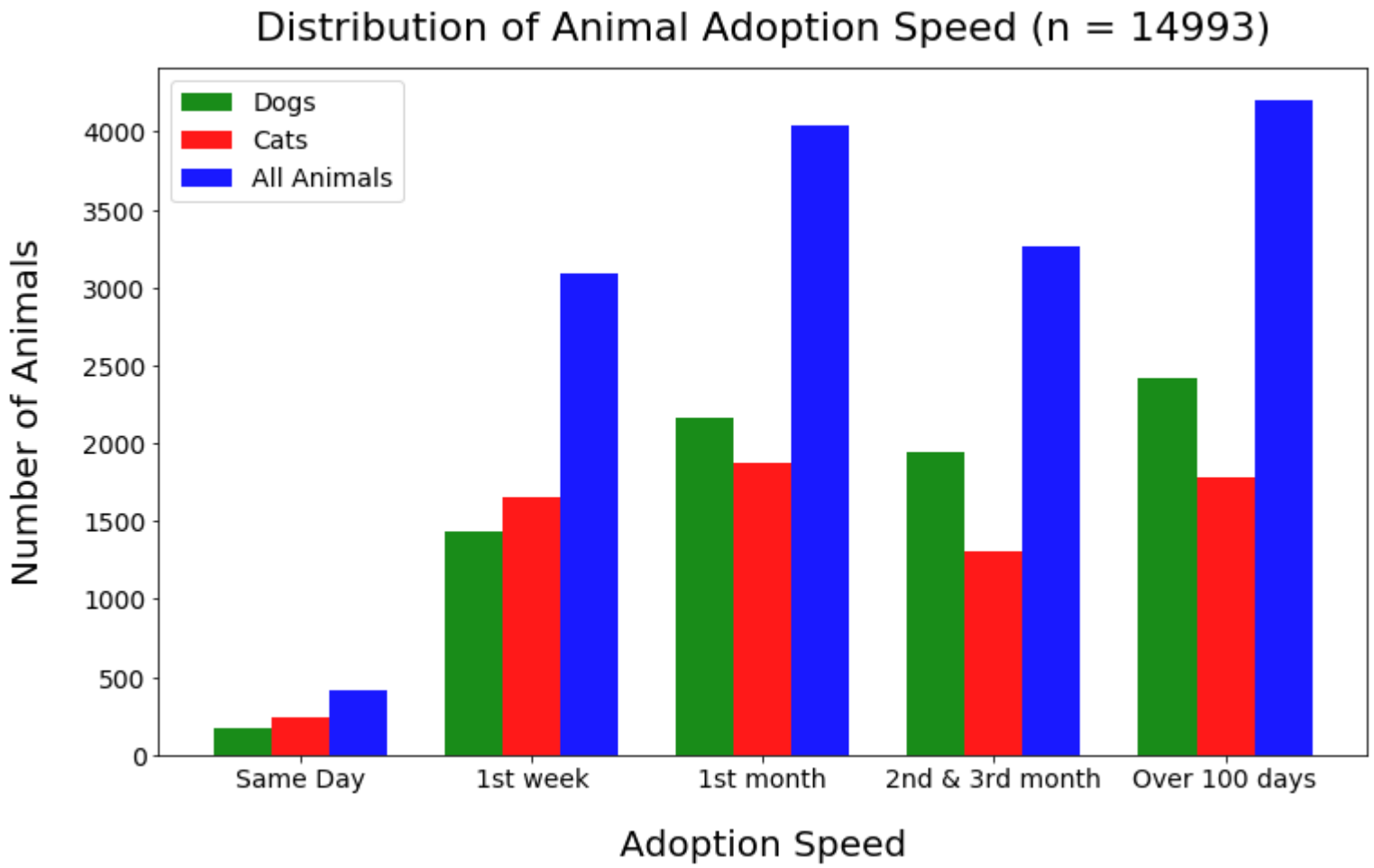
- Determine which animals get adopted fastest
  - Identify important features in Petfinder profiles
  - Develop a model to predict an animal's adoptability using supervised machine learning
  - Tune the model features to improve its capabilities

Available Data

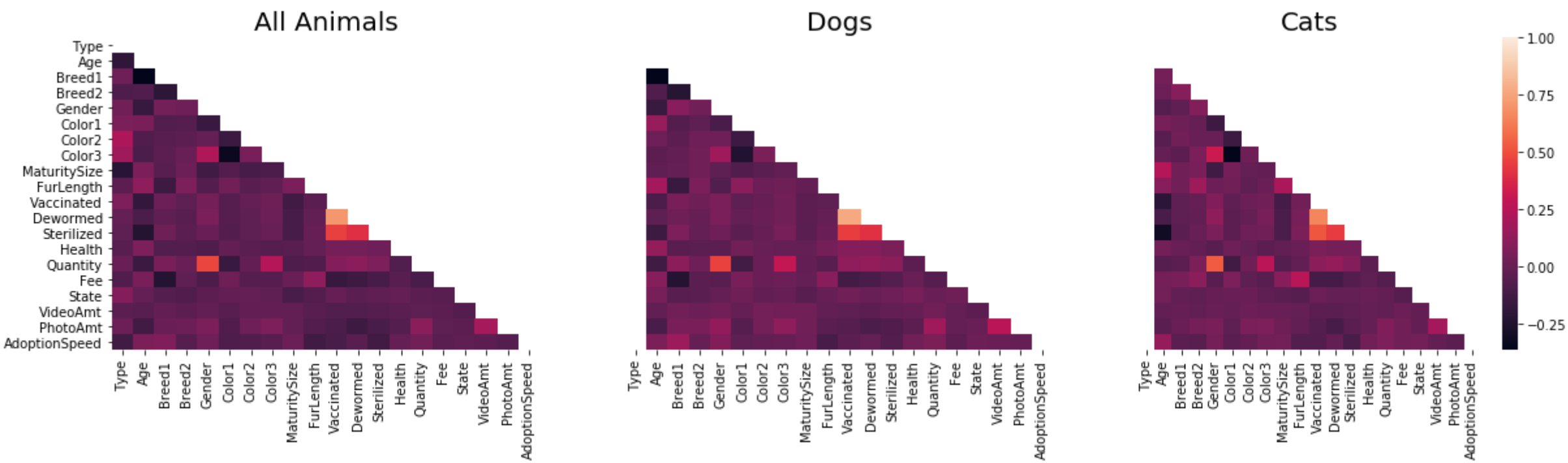
- Numerical features – age, quantity, fee, number of uploaded videos, number of uploaded photo, maturity size, fur length, health

- Categorical features – dewormed, vaccinated, sterilized, breed, state, gender, color

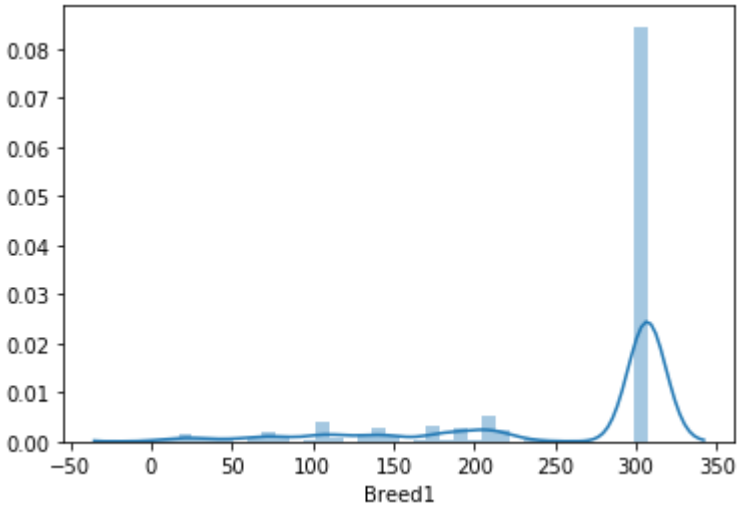
Adoption Speed

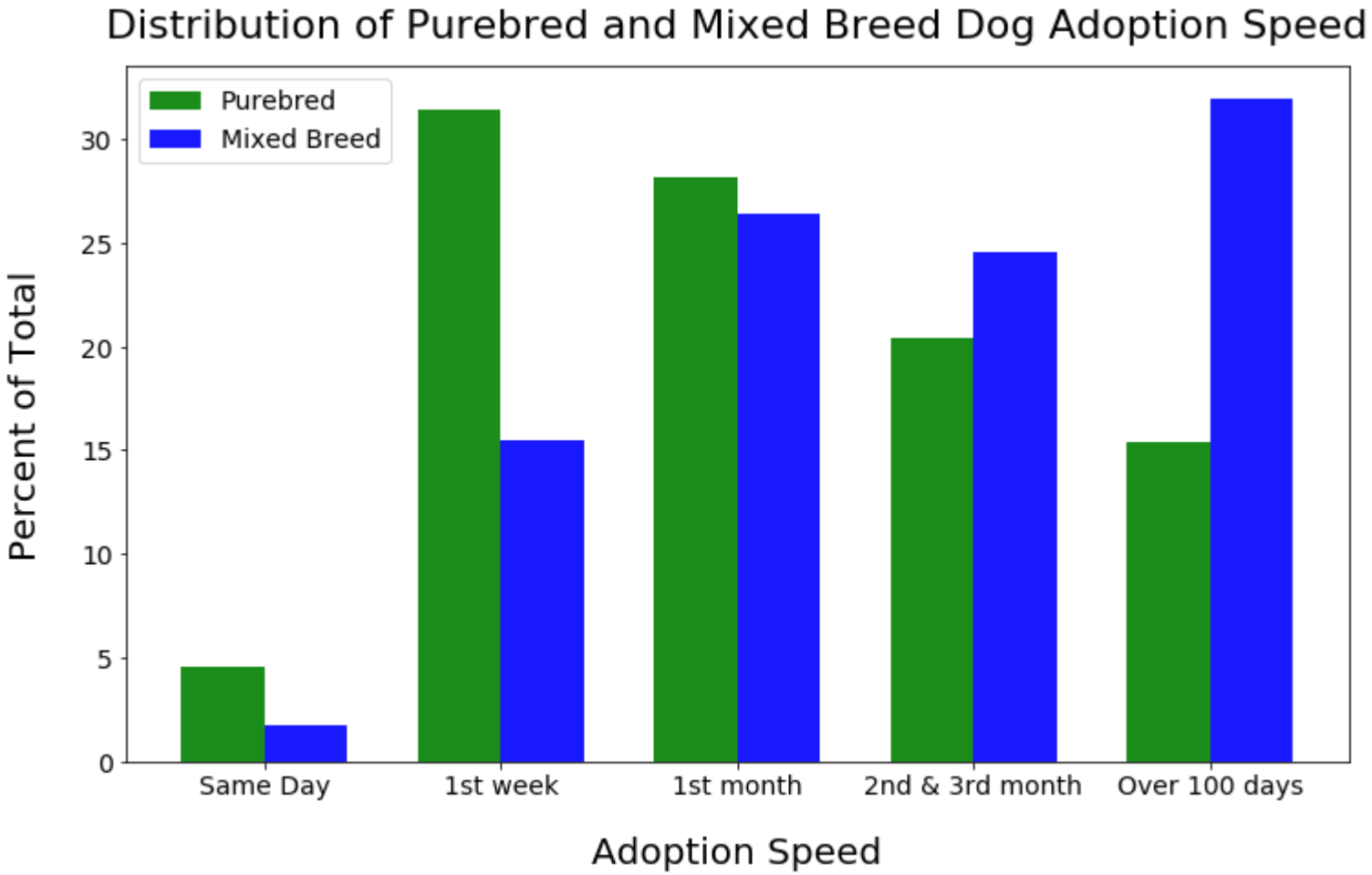


Which factors adoption speed?

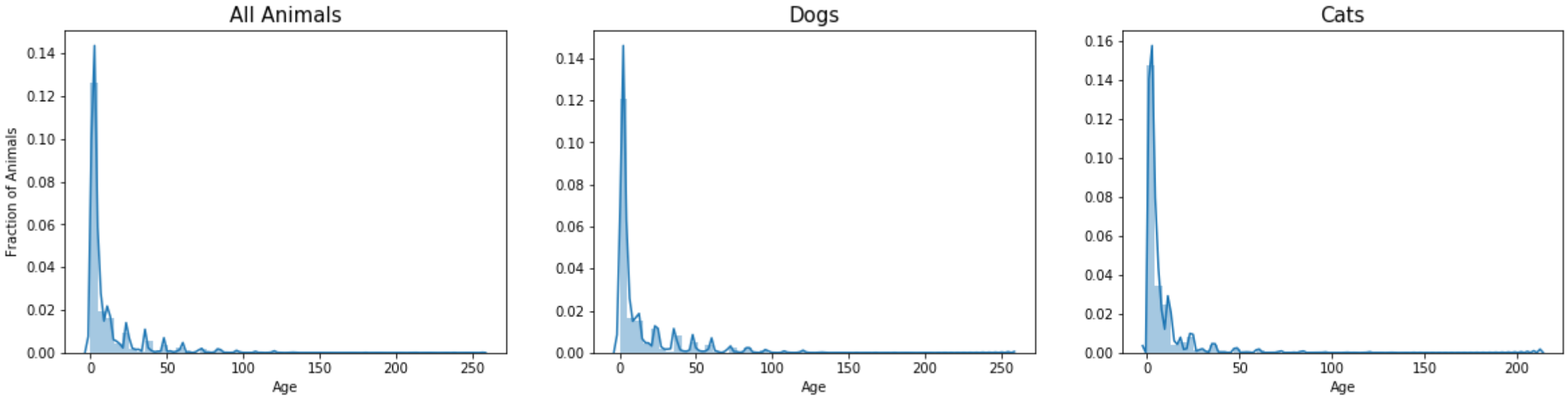


Purebred dogs are adopted earlier than mixed breed dogs



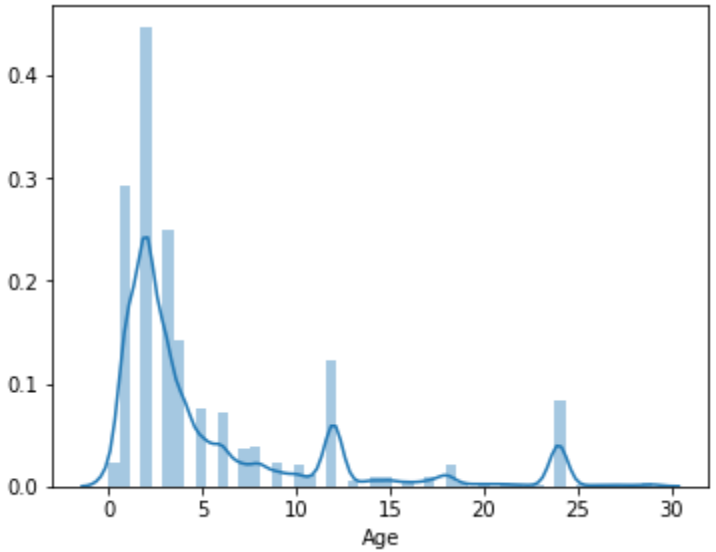


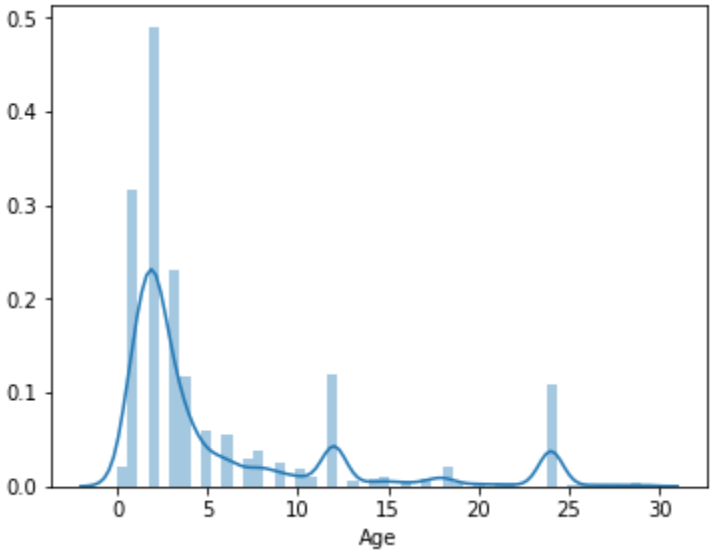
Distribution by Age

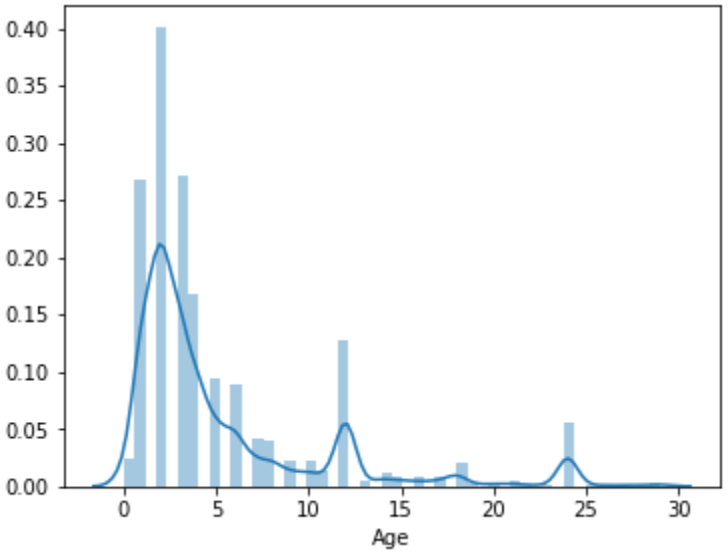


Since it is unlikely that any animals are above the age of 30, all of these animals were removed

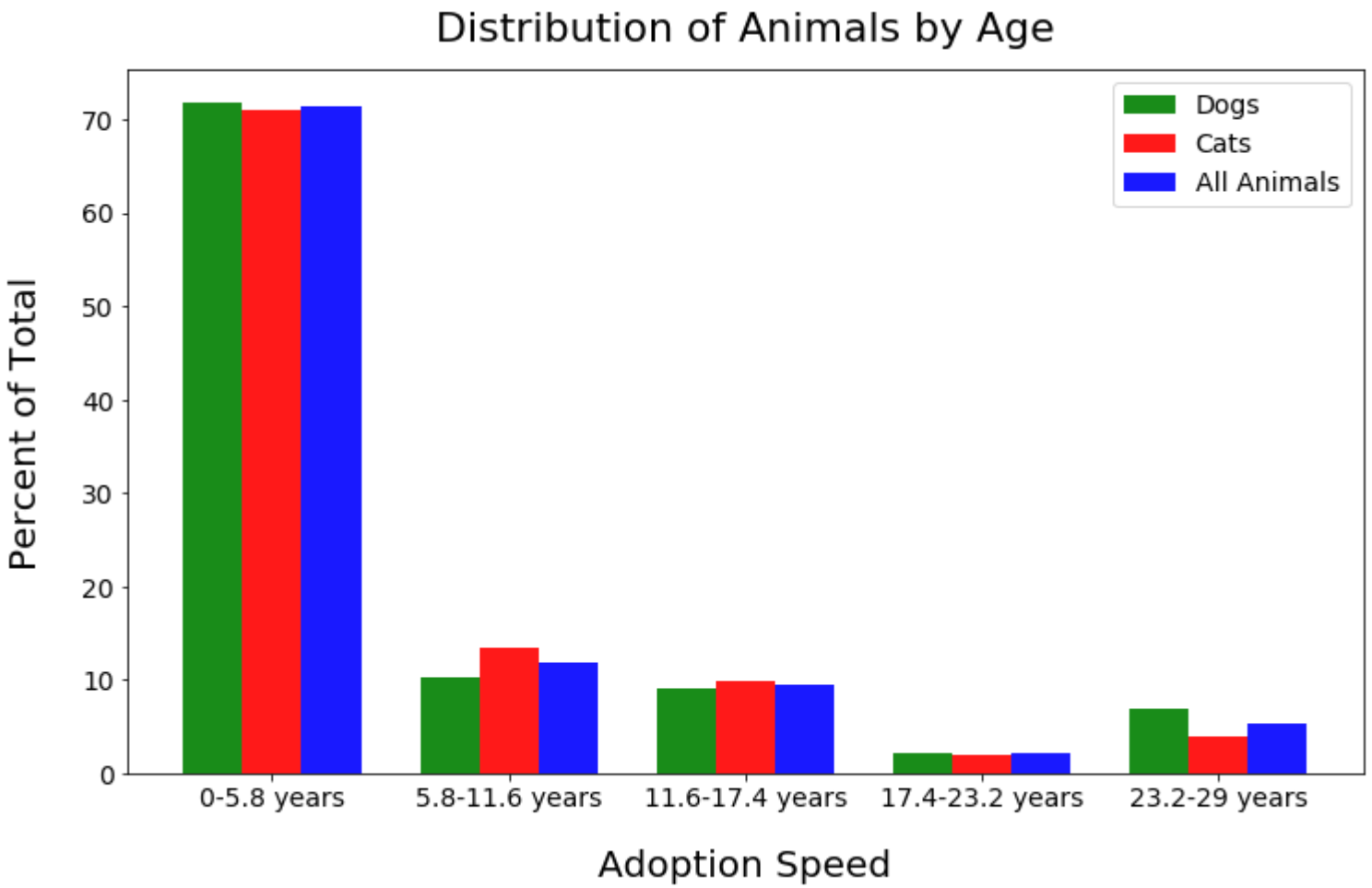
Distribution by Age







Most animals are between 1 and 3



Most animals were under 5.8 years old. To look at age distributions more evenly animals were binned by the following age range:

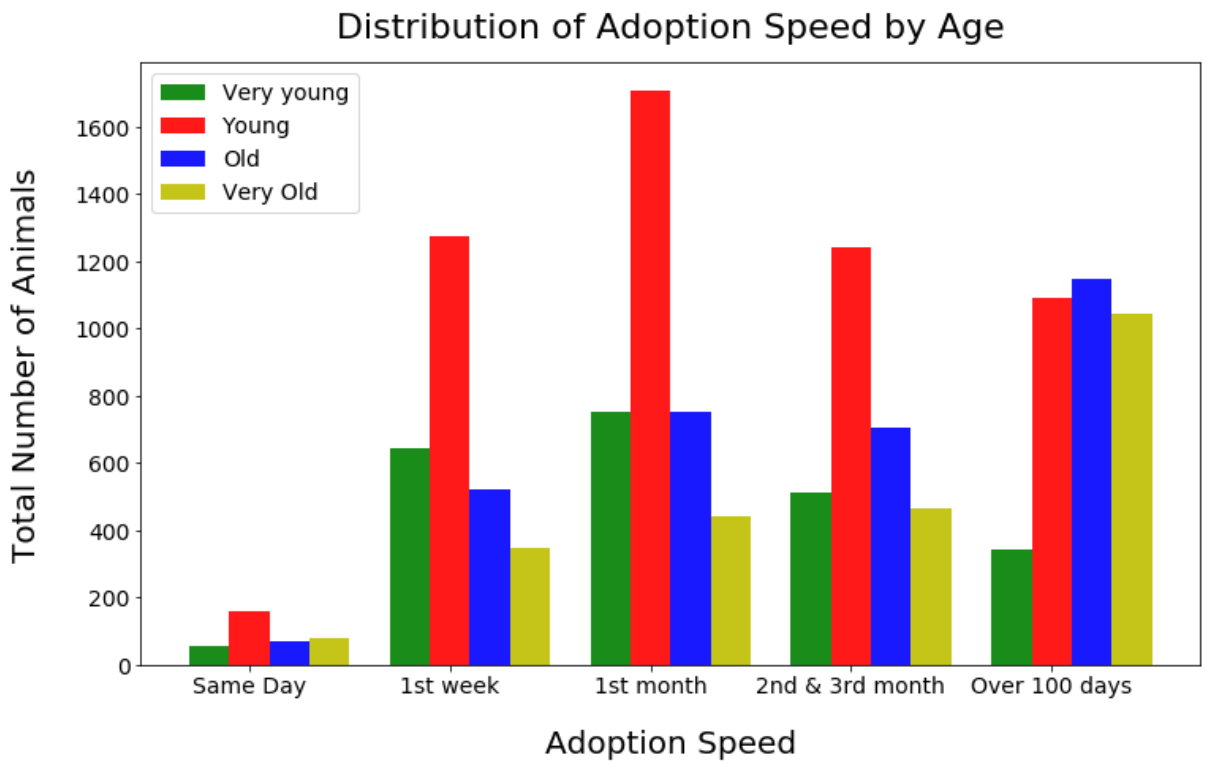
Very young: Under 1

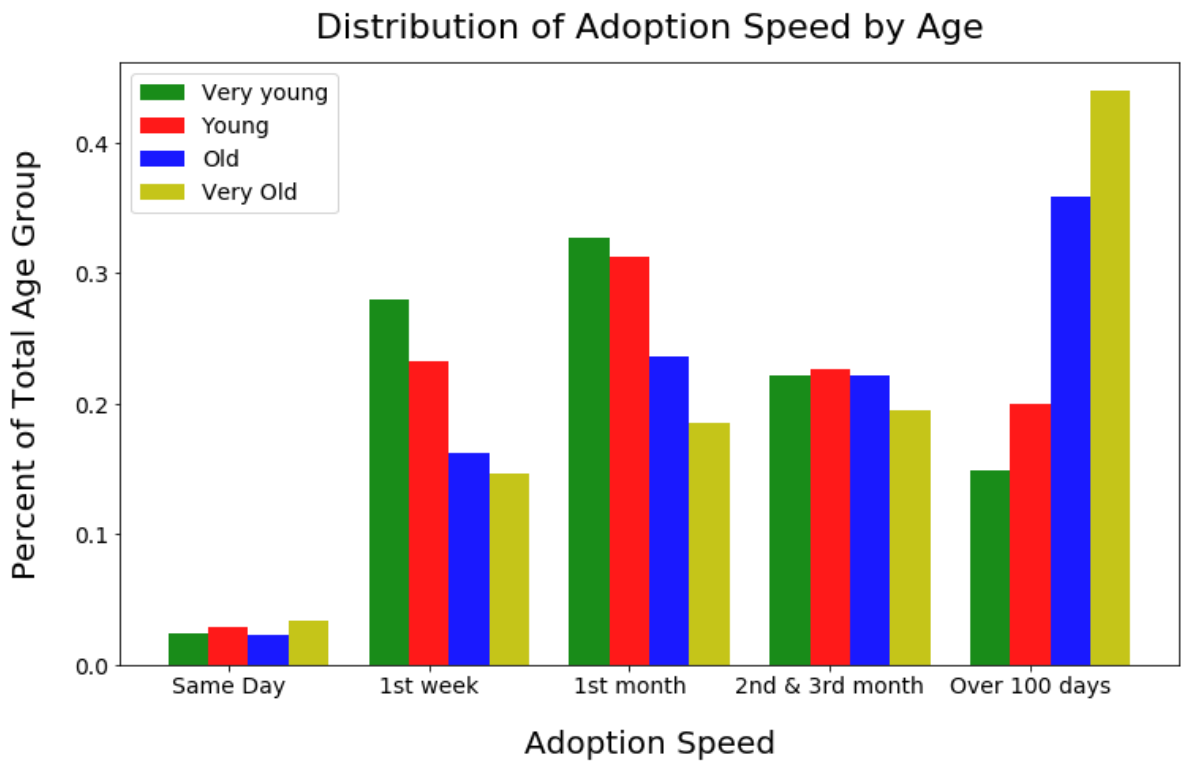
Young: 2-4

Old: 4-10

Very old: older than 10

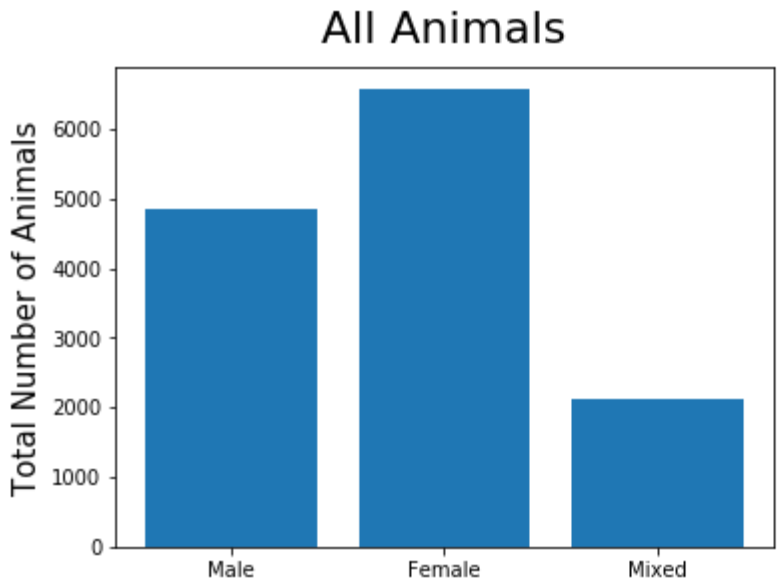
How does age affect rate of adoption?

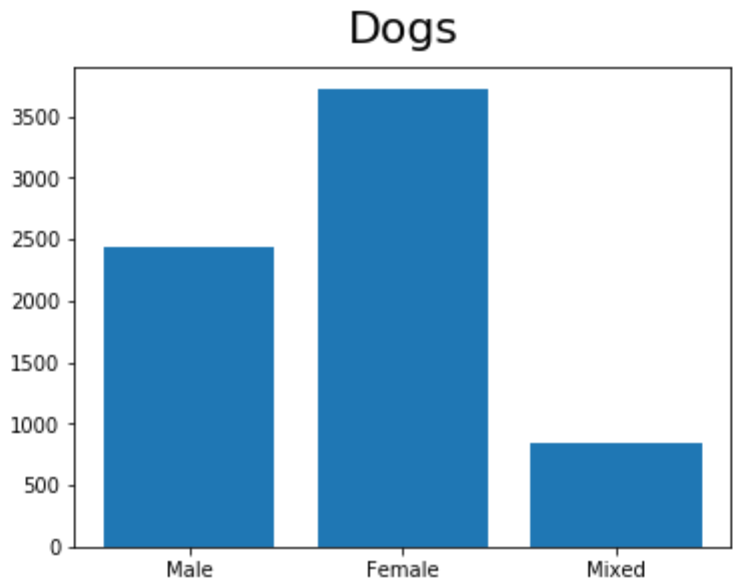


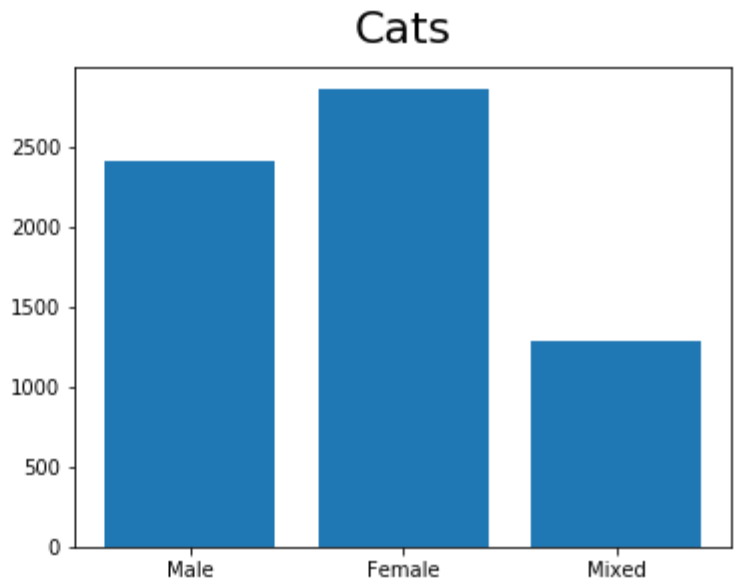


A higher percentage of older animals remain unadopted after 100 days.

Does gender affect adoption rates?





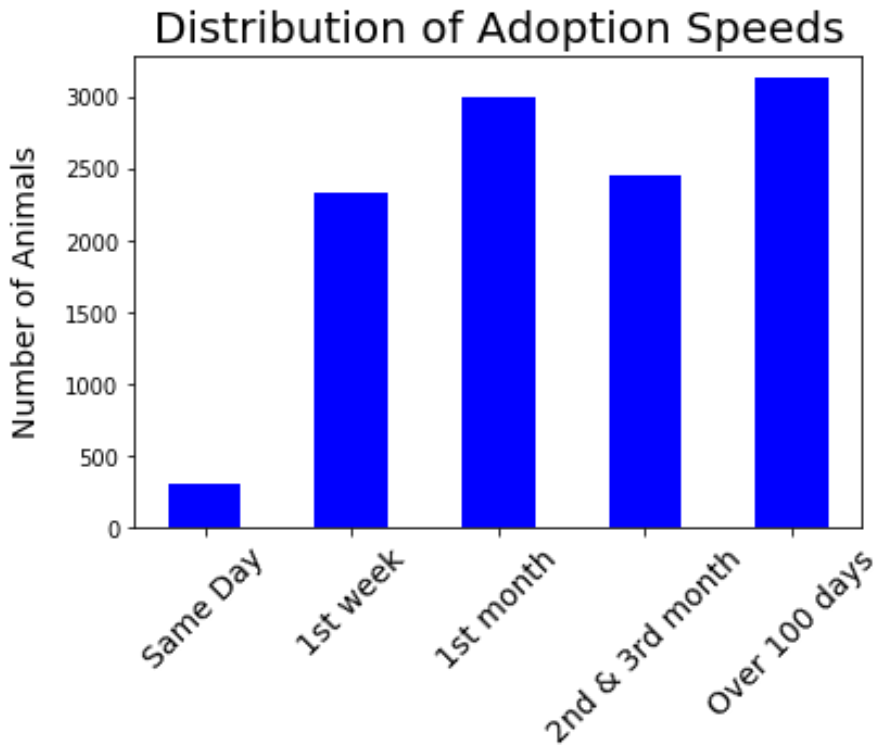


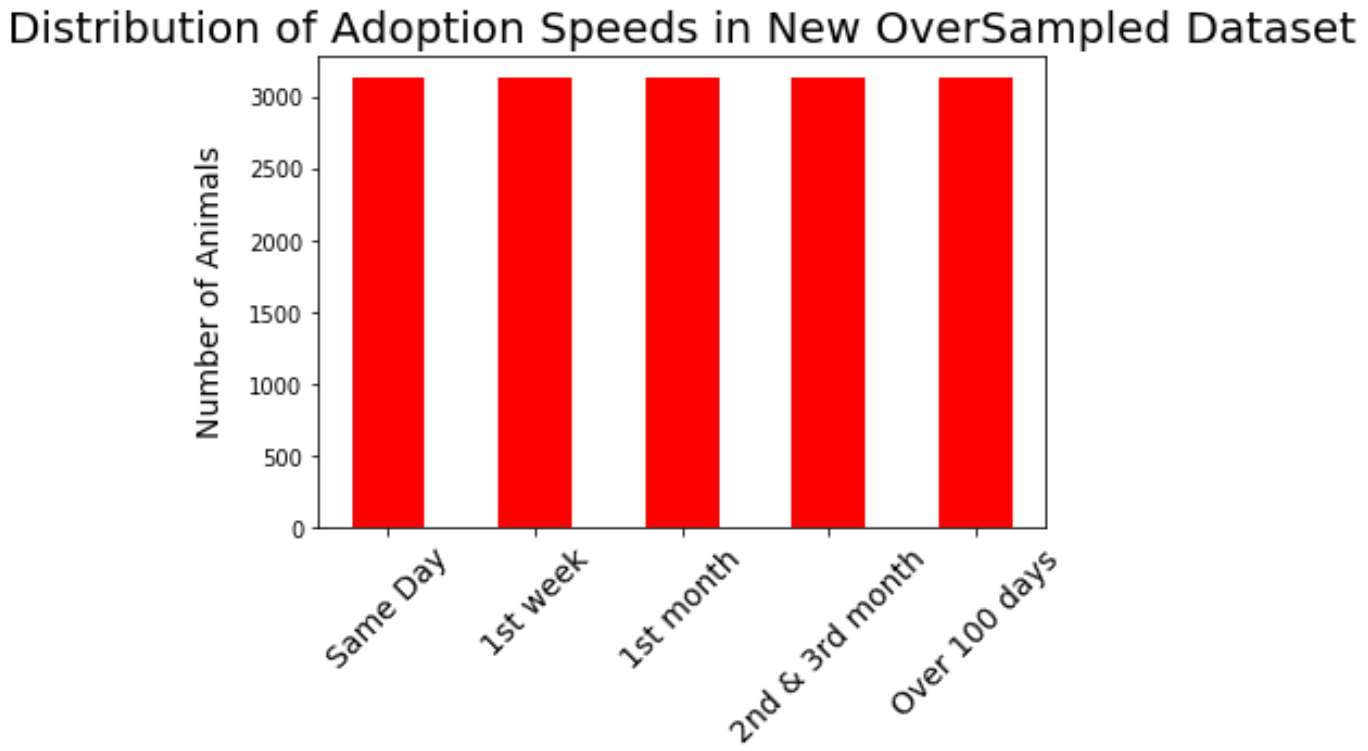
Take out mixed from model, since it is not related to gender

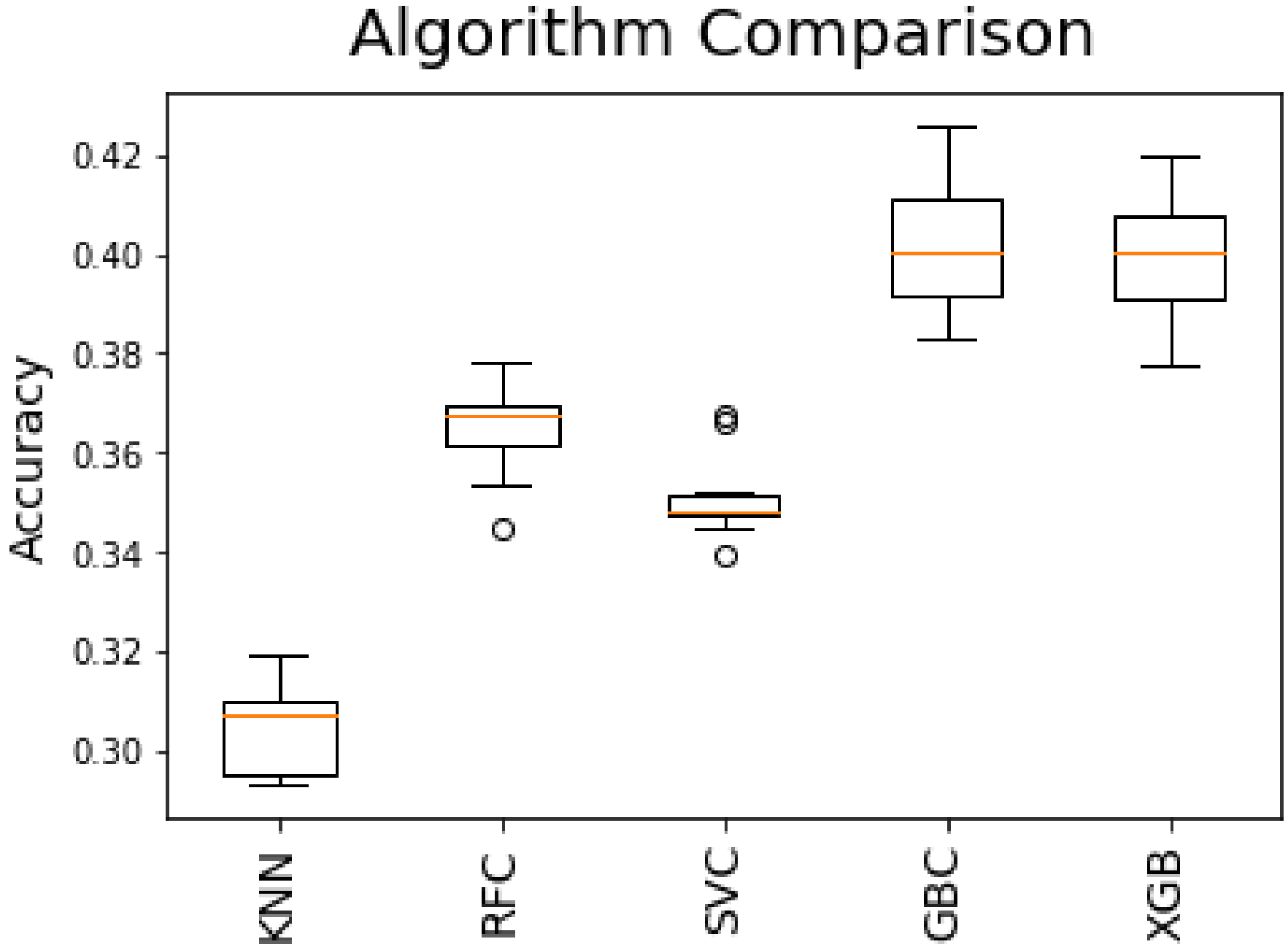
Model 1

- One hot encode all categorical features, standardize all continuous features
- Split the data into training and testing datasets
- Try models:
  - KNN Classification
  - Random Forest Classification
  - Support Vector Classification
  - Gradient Boost Classification
  - Xgboost Classification

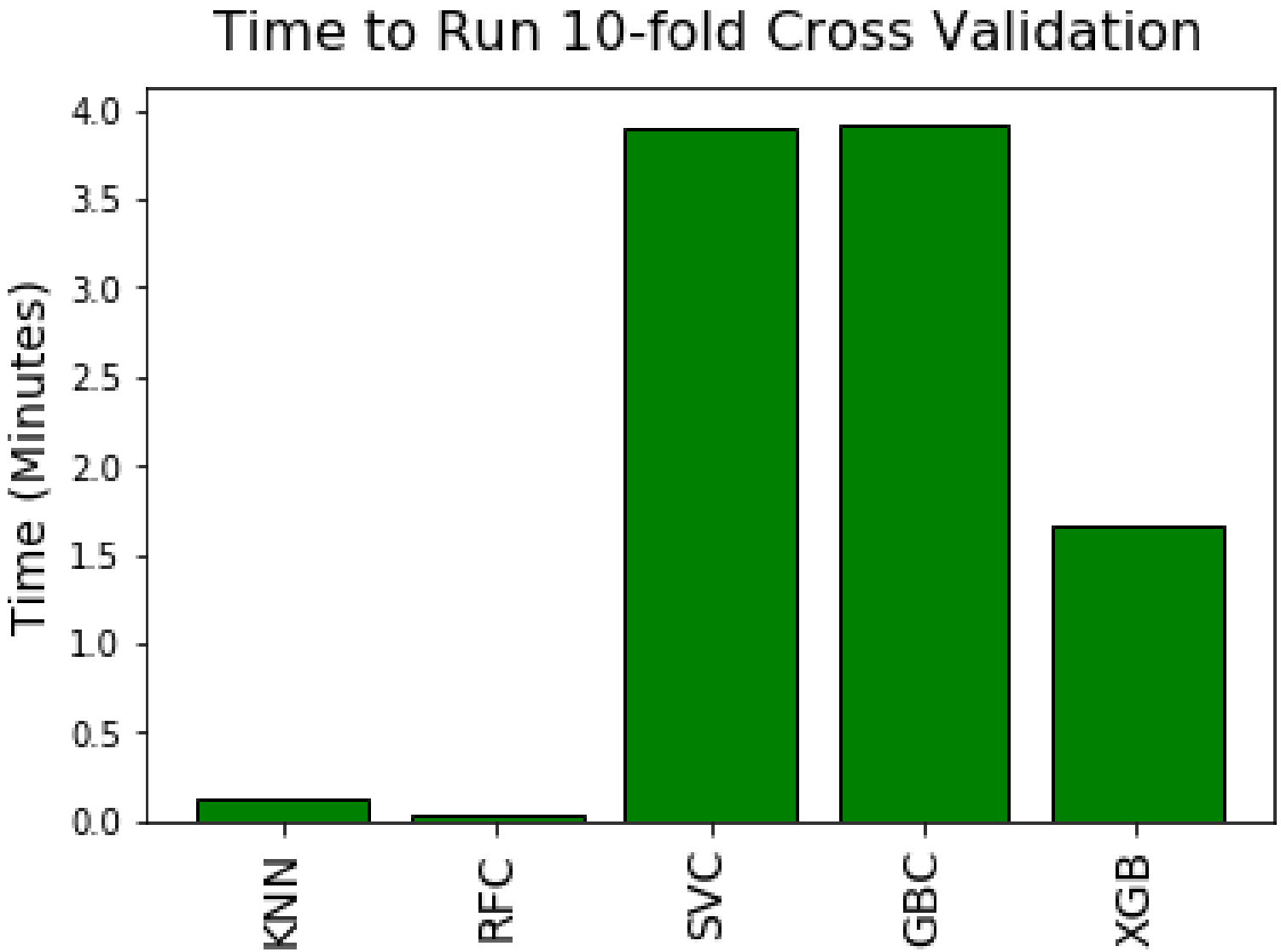
SMOTE was used to oversample the training set



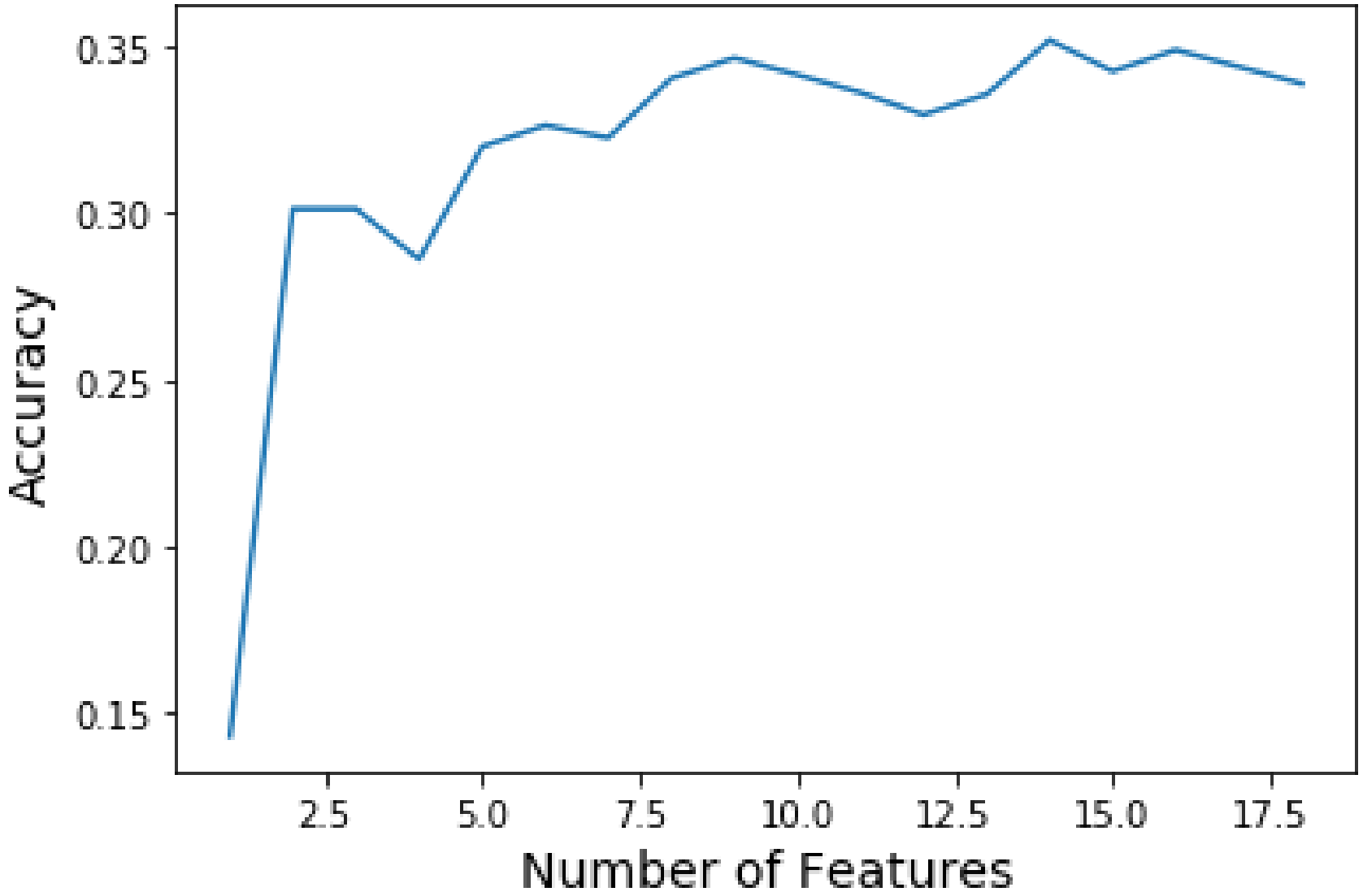



Algorithm Comparison
Accuracy
0.42
0.40
0.38
0.36
0.34
0.32
0.30
KNN
RFC
SVC
GBC
XGB

Algorithm Comparison



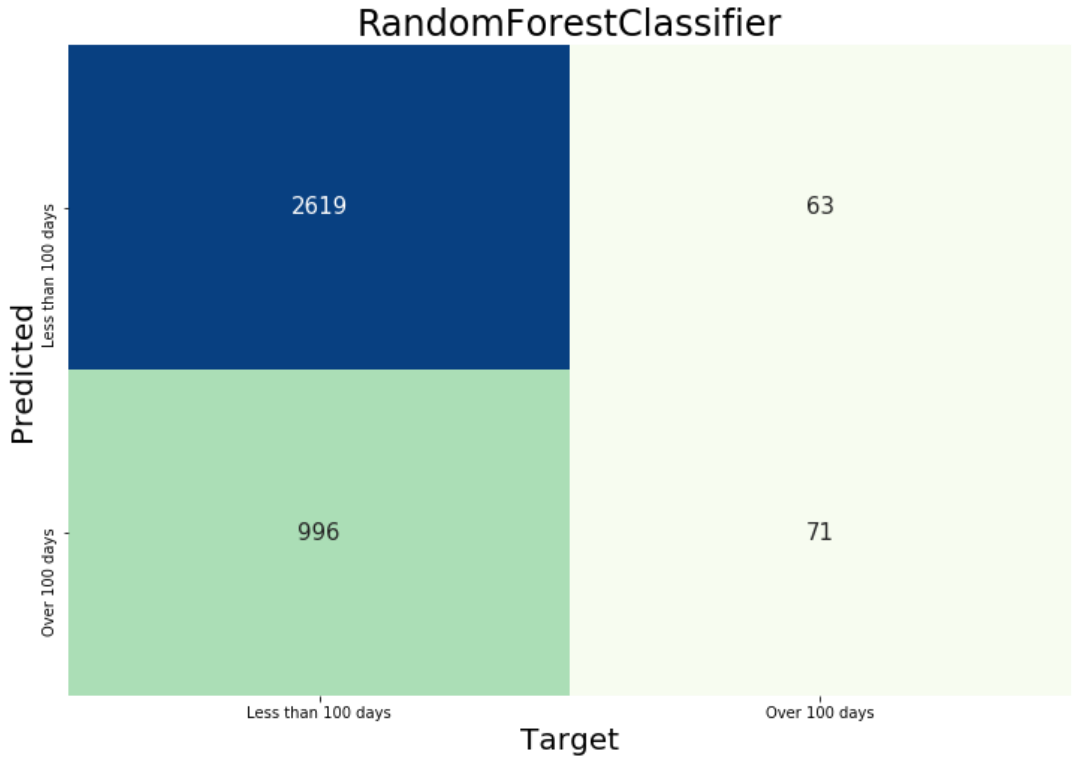
SelectKBest Features

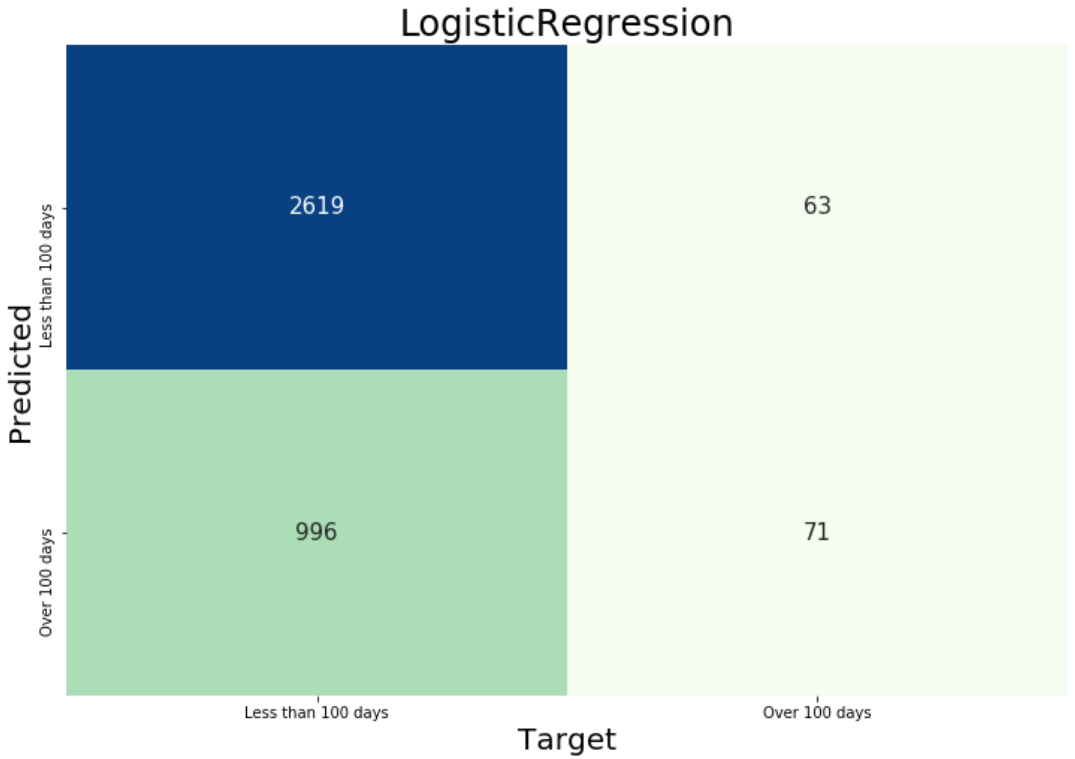


Model 2

- Try Converting Adoption Rate to Binary
  - 0 if adopted faster than 100 days
  - 1 if longer than 100 days
- Logistic Regression and Random Forest

Results





Training set accuracy: 0.98

Test set accuracy: 0.74

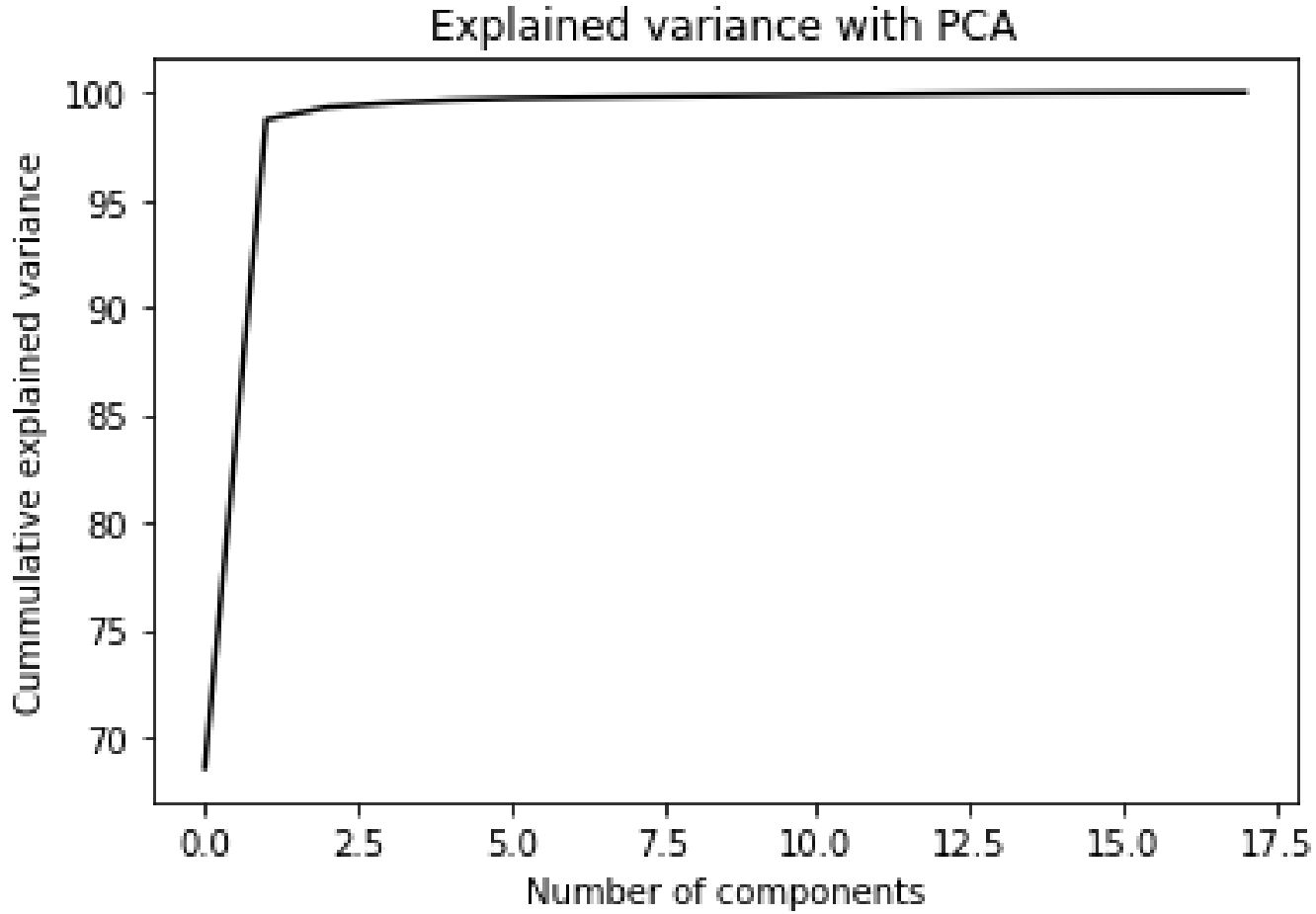
Cross validation results: 0.739+/-0.025

Training set accuracy: 0.73

Test set accuracy: 0.72

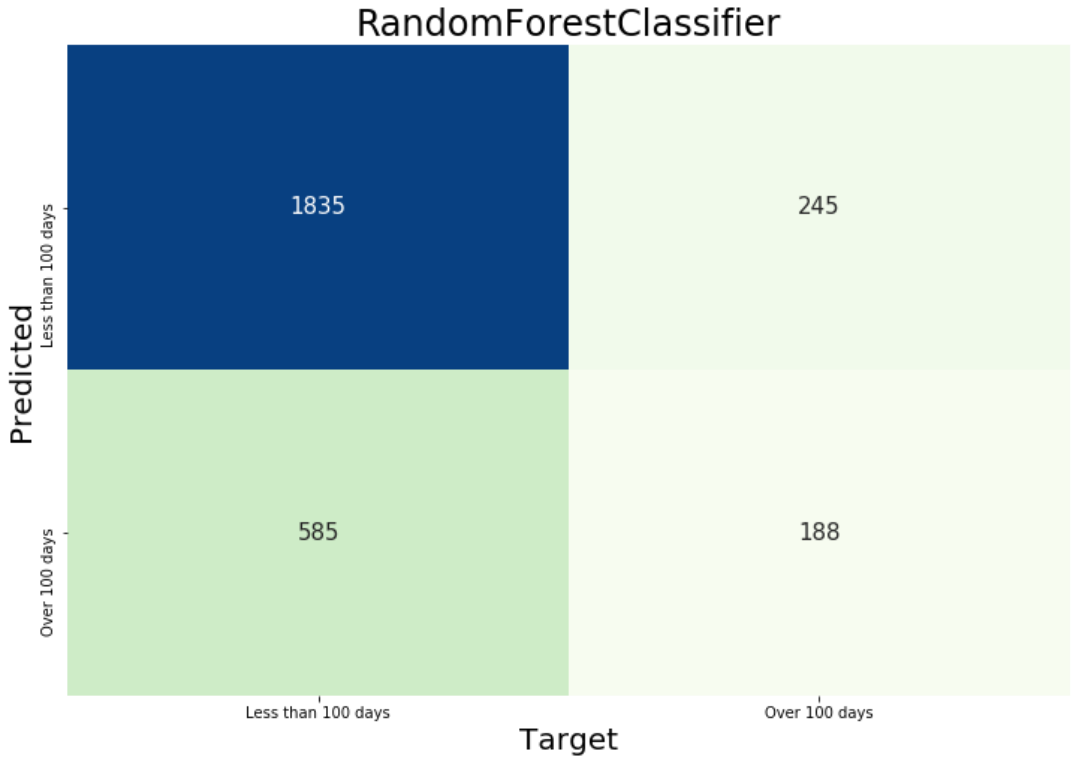
Cross validation results: 0.727+/-0.00831

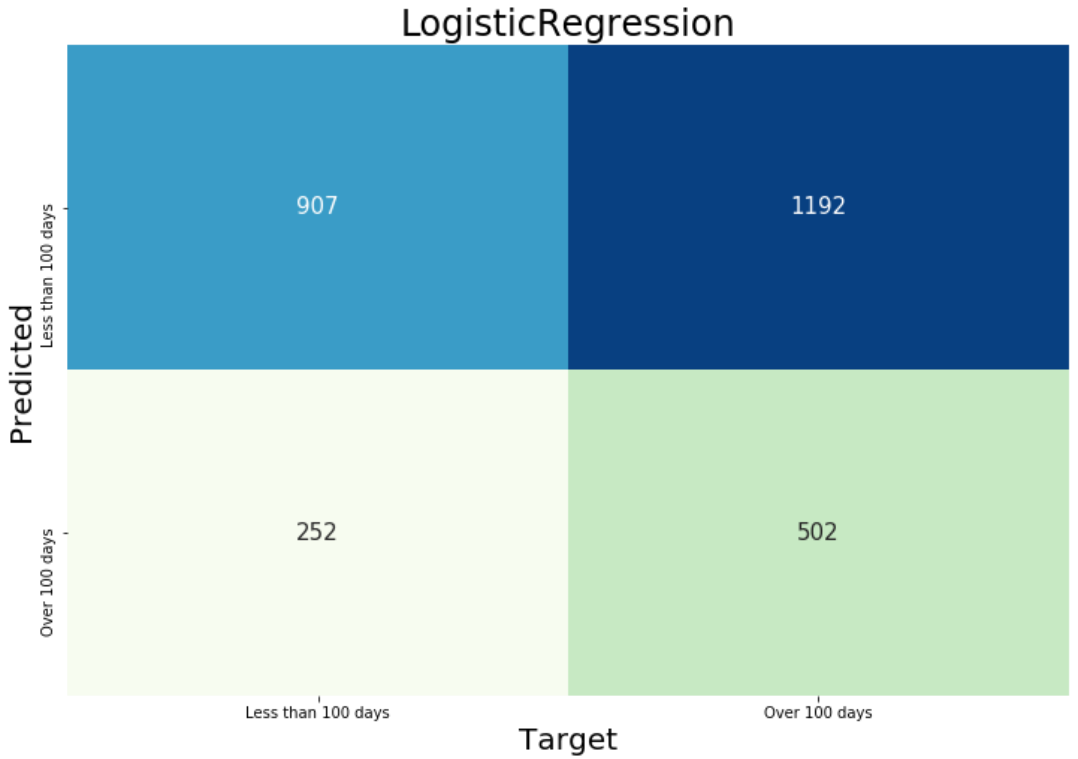
Model 3 - PCA



Choose 3 components – explains 99.3% of variance

Model 3 - PCA





Training set accuracy: 0.98

Test set accuracy: 0.66

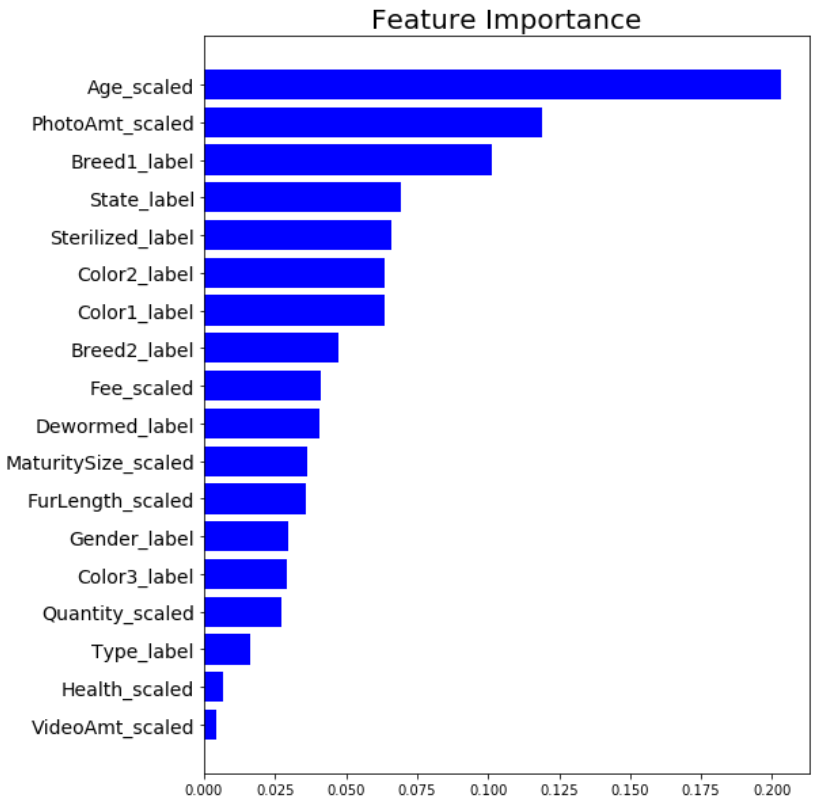
Cross validation results: 0.72+/-0.0905

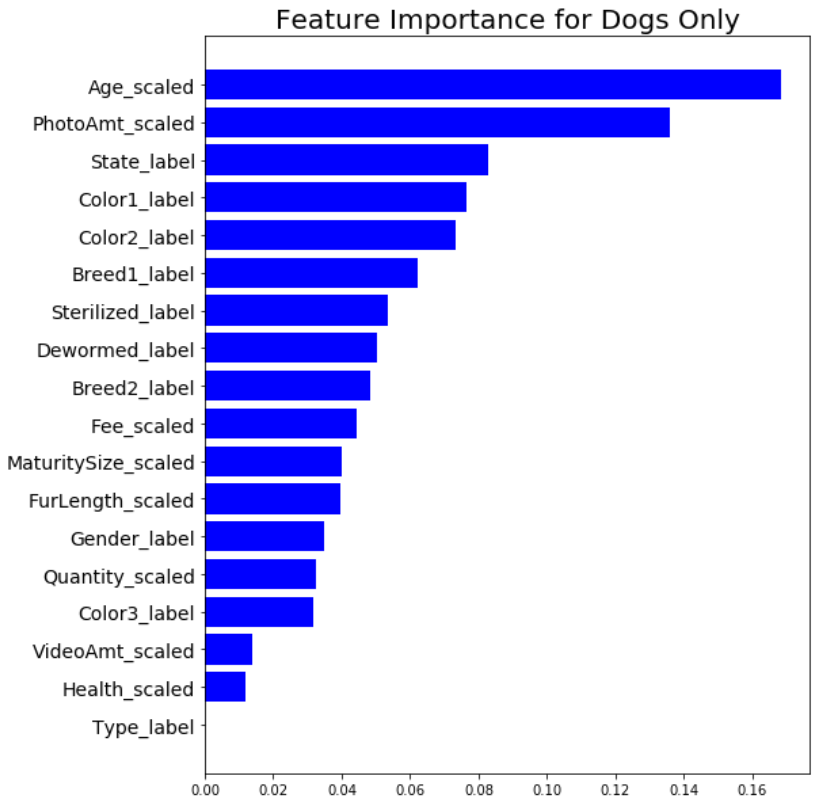
Training set accuracy: 0.56

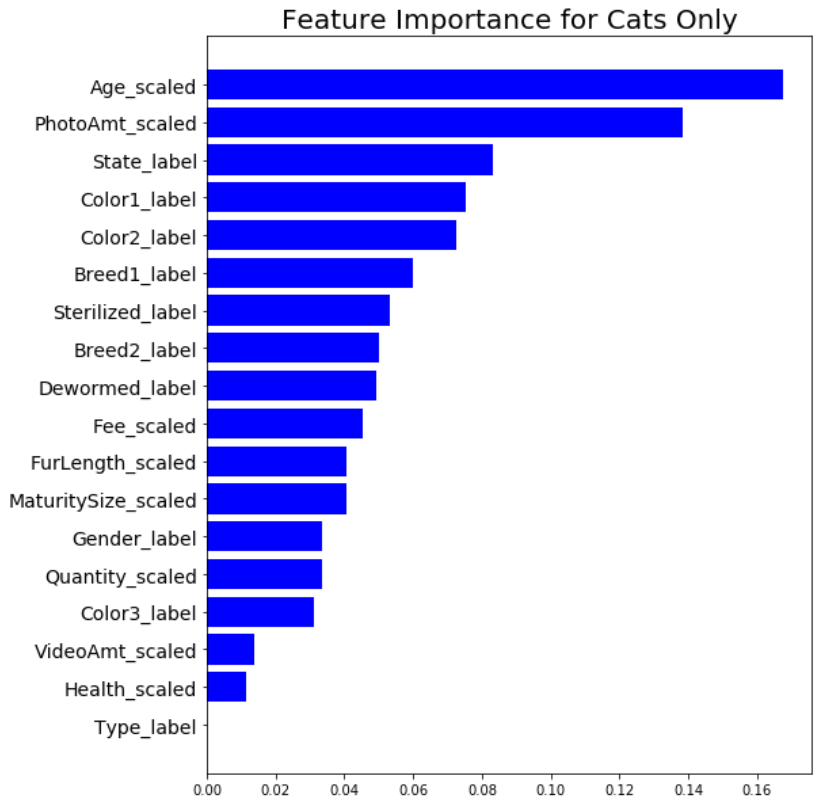
Test set accuracy: 0.49

Cross validation results: 0.56+/-0.0141

Feature Importance







Conclusions

- Random forest classification performed best overall
  - Had overfitting issues however
- Ridge regression performed worse, but had less overfitting
- PCA did not improve performance
- Age, number of provided photos, and breed of animal were the most important feature

Proposed Further Research

- Expand model to include data for other countries
  - This dataset just looks at adoptions in Malaysia
- Look at trends in U.S. by county and state
- Include sentiment analysis of the animal descriptions in the models
- Include images and videos in the models to determine if it improves predictability