A Home for All

Utilizing the Kaggle PetFinder Adoption Competition to Strategize Animal Intake

Caitlin Janson

December 14, 2020

Agenda

- Problem Statement
- Dataset
- Exploratory Data Analysis
- Statistical Analysis
- Model Development and Results
- Recommendations
- Dataset limitations
- Opportunities for future work

Problem Statement

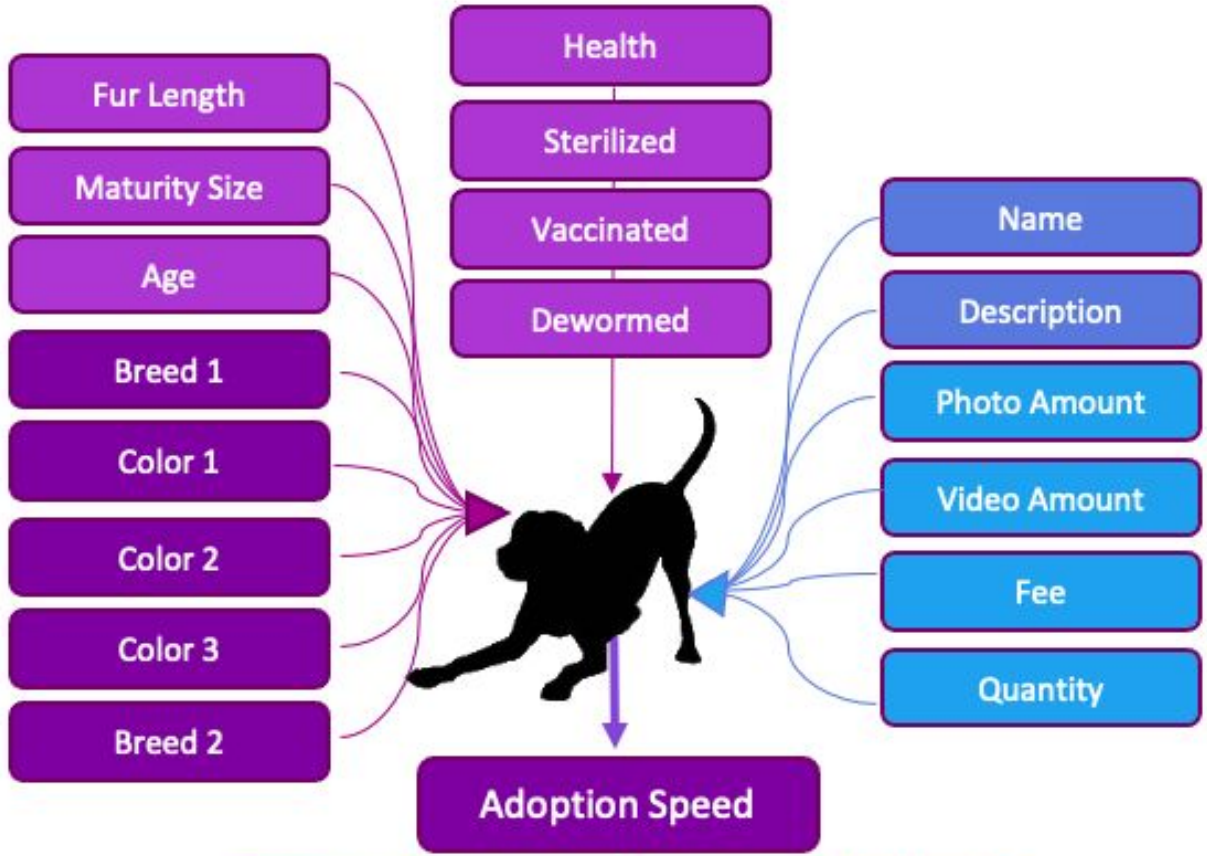
The shelters are overpopulated and it seems nothing is working! The client needs to know how can they capitalized on trends to develop a more strategic business plan for adopting out a pet without:

- Bothering with “Marketed” descriptions resulting in disappointment and pet returns
- Taking the wrong pets to events when they’ll be adopted even with minimal exposure
- Long term foster fees when a pet could be transferred to a different rescue and be adopted in under 90 days

Dataset Overview

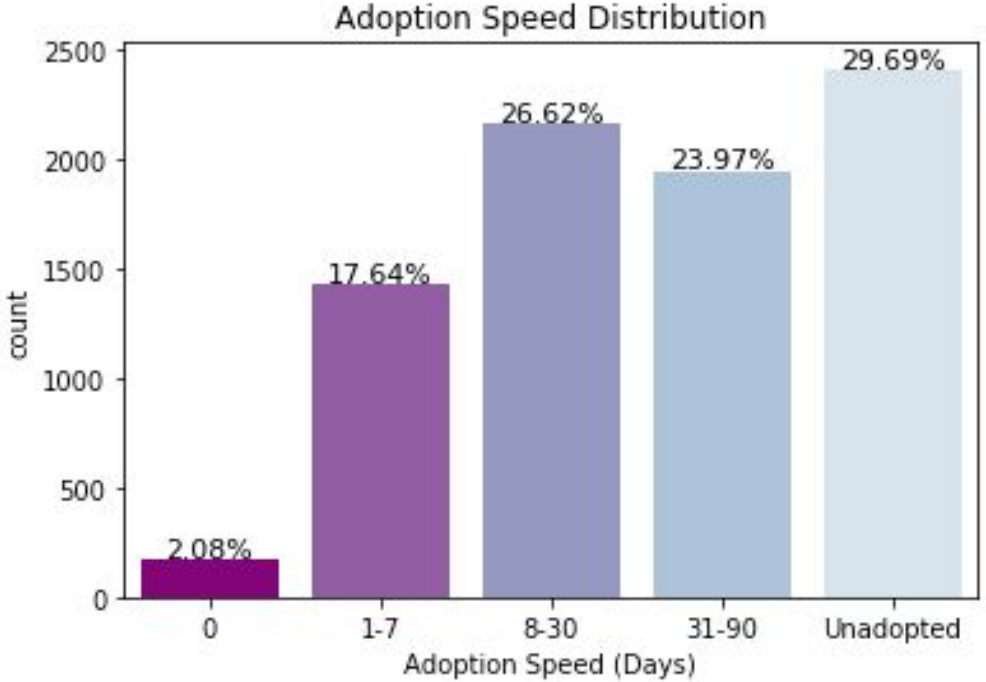
This project utilizes the [Kaggle PetFinder Dataset]

- Targeted at increasing adoption speeds for homeless pets
- Contains over 10,000 animals, including cats and dogs
- Sourced from the Malaysian region
- Is primarily ordinal data (0,1,2,3) as placeholders for descriptive data (puppy, adolescent, adult, senior)
-

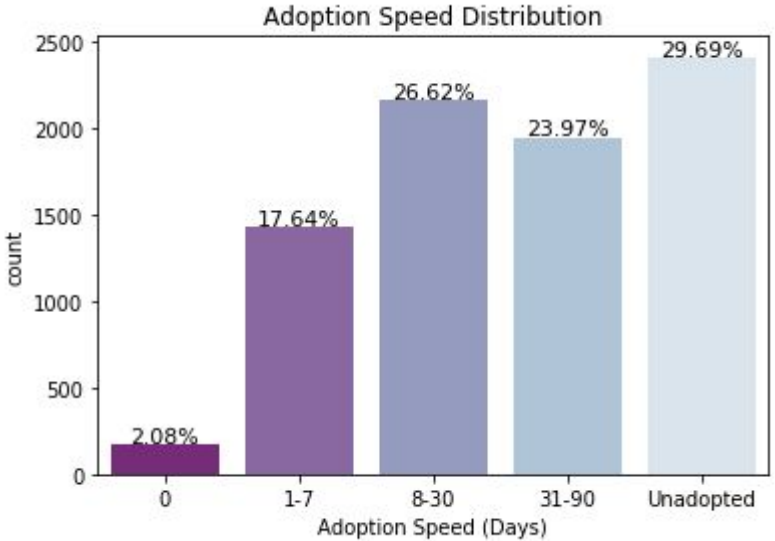


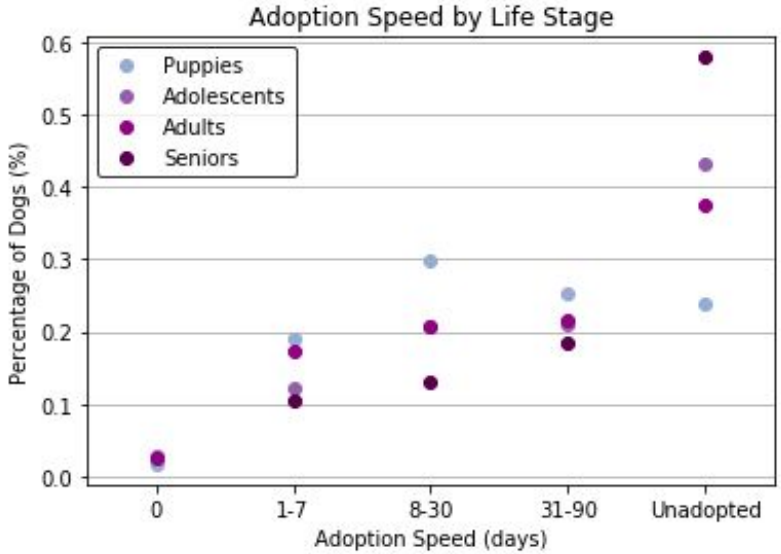
Dataset Overview

Exploratory Data Analysis- Adoption Speed

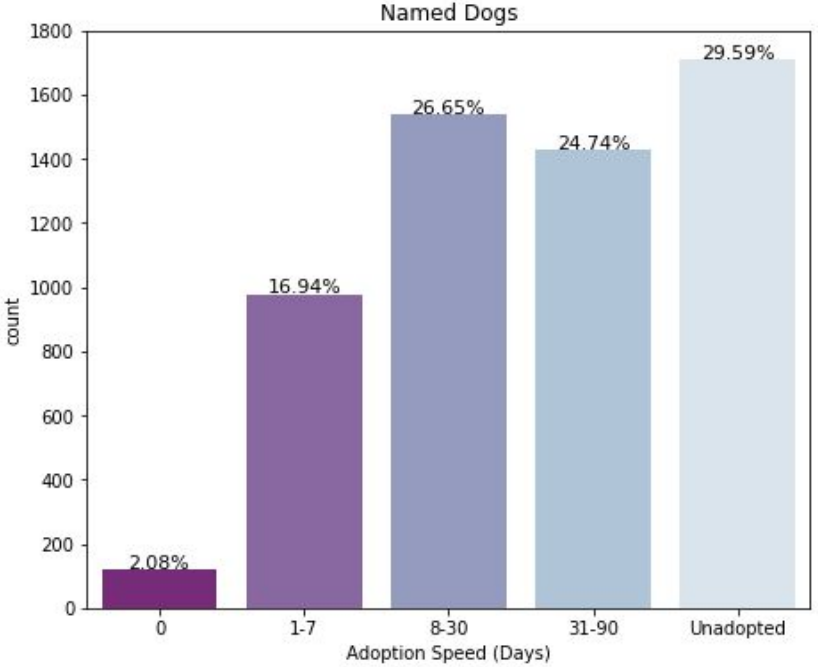


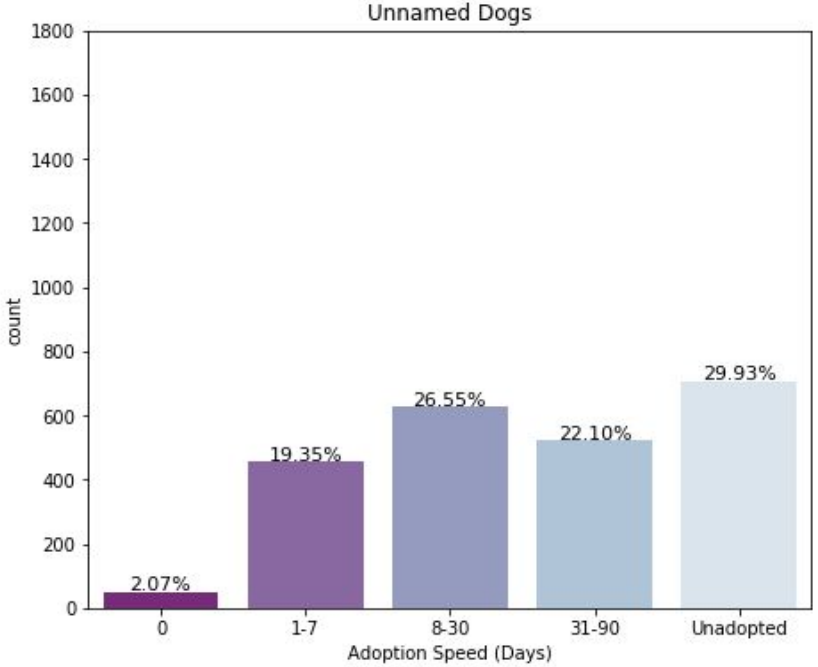
Exploratory Data Analysis- Age



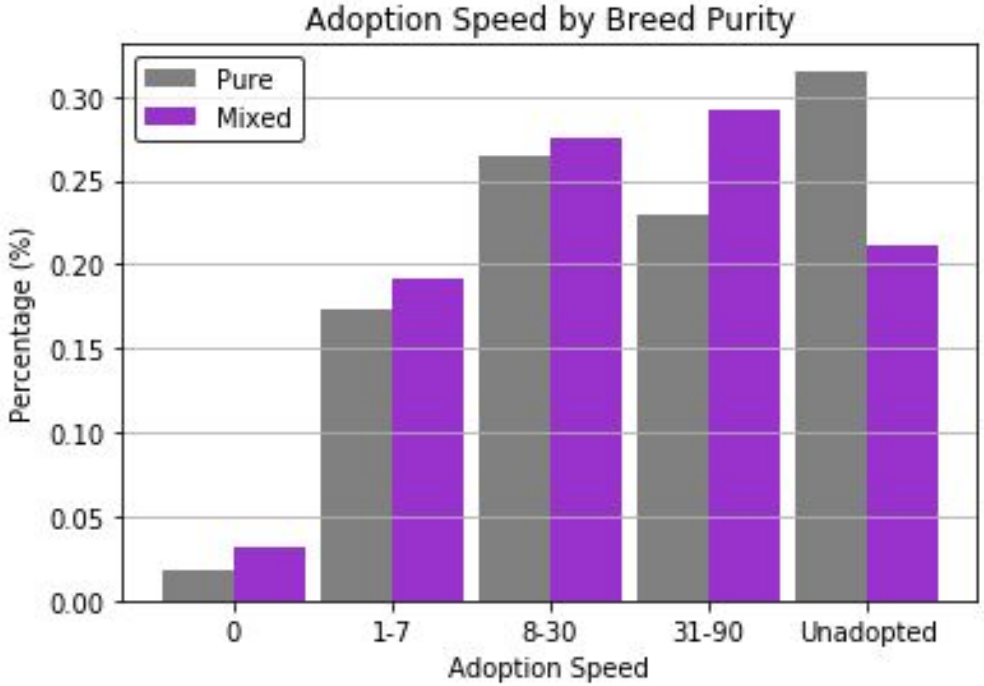


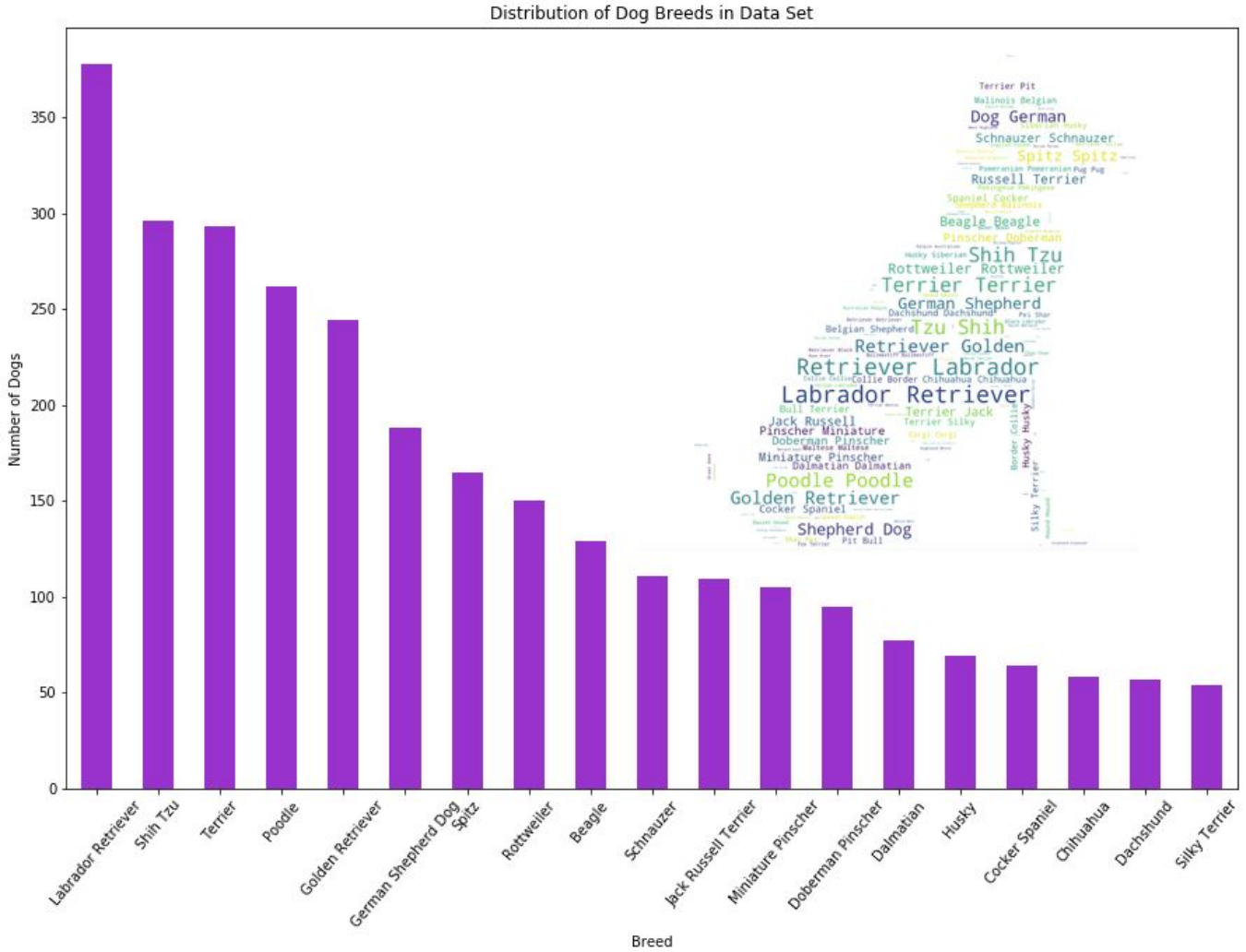
Exploratory Data Analysis - Name





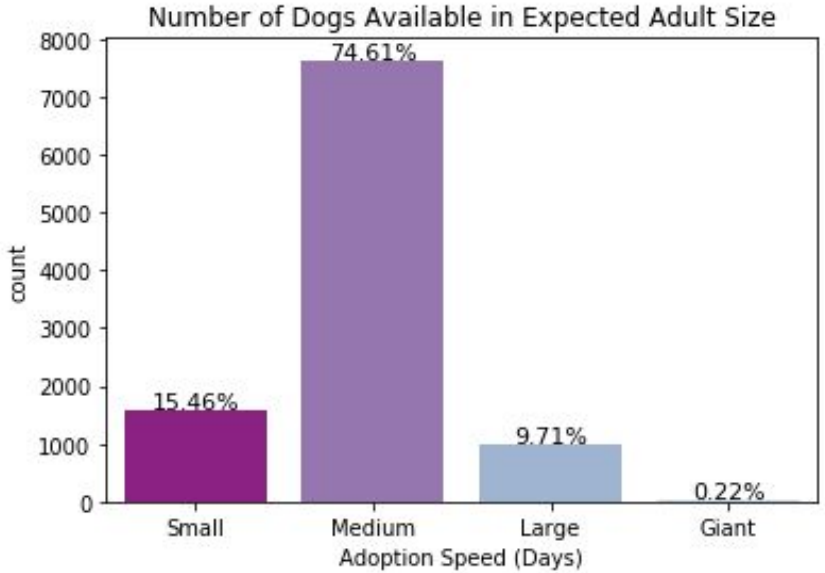
Exploratory Data Analysis- Breed

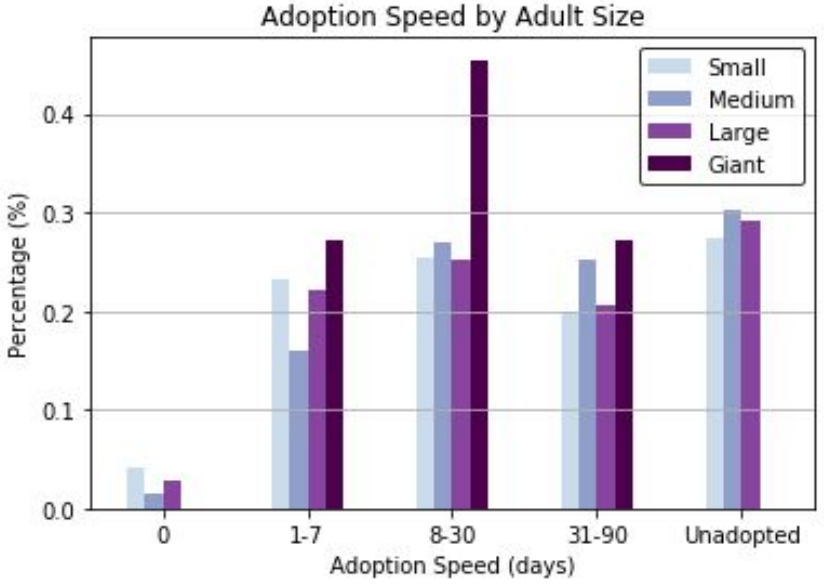



Distribution of Dog Breeds in Data Set
Number of Dogs
0
50
100
150
200
250
300
350
Labrador Retriever
Shih Tzu
Terrier
Poodle
Golden Retriever
German Shepherd Dog
Spitz
Rottweiler
Beagle
Schnauzer
Jack Russell Terrier
Miniature Pinscher
Doberman Pinscher
Dalmatian
Husky
Cocker Spaniel
Chihuahua
Dachshund
Silky Terrier
Breed

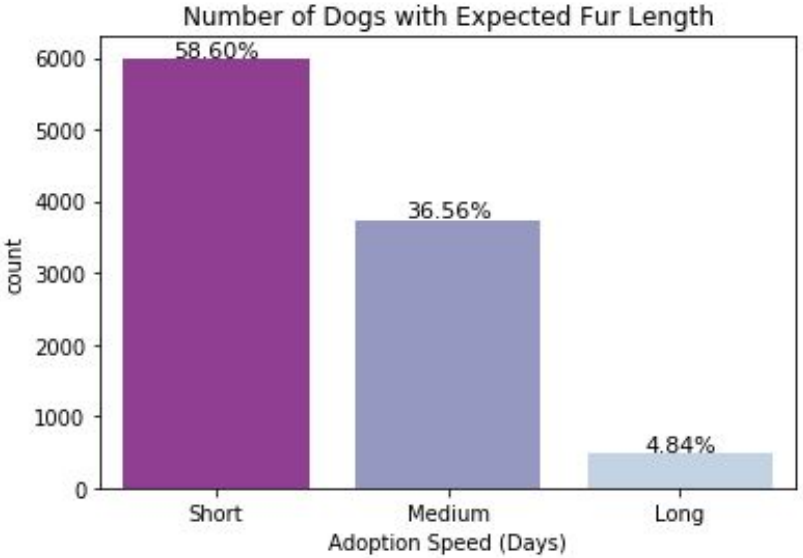
Breed

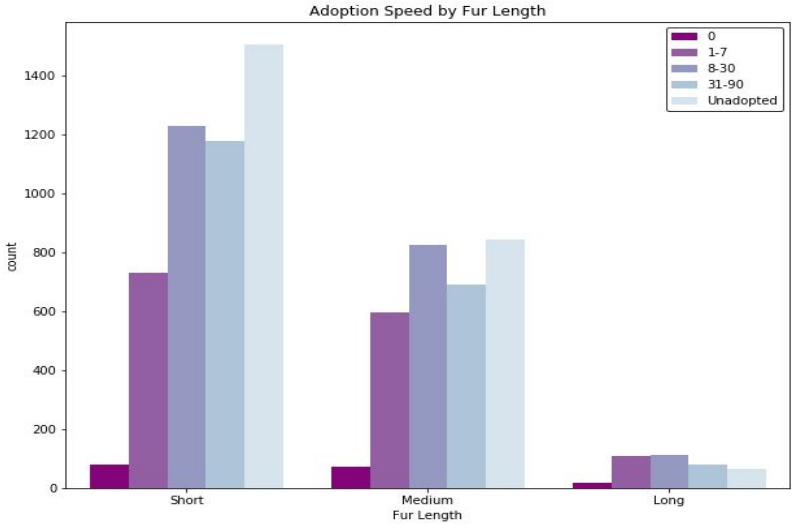
Exploratory Data Analysis - Adult Size



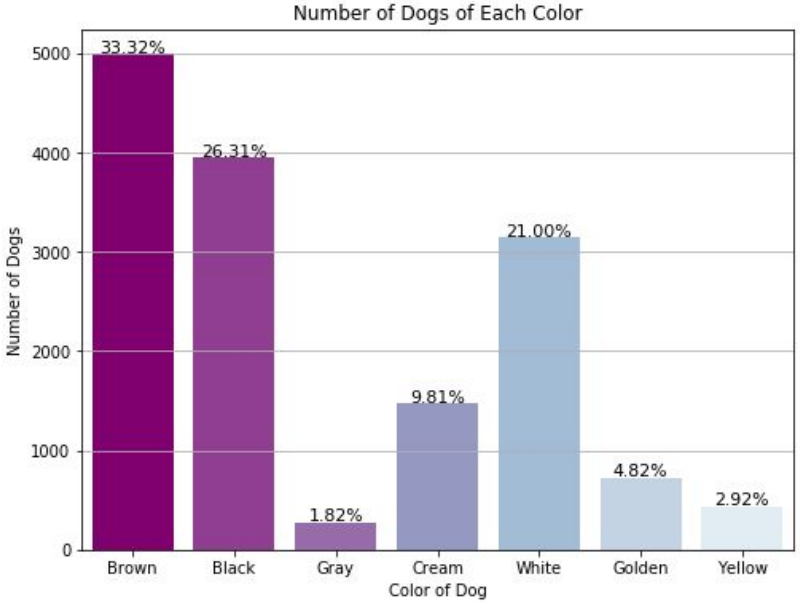


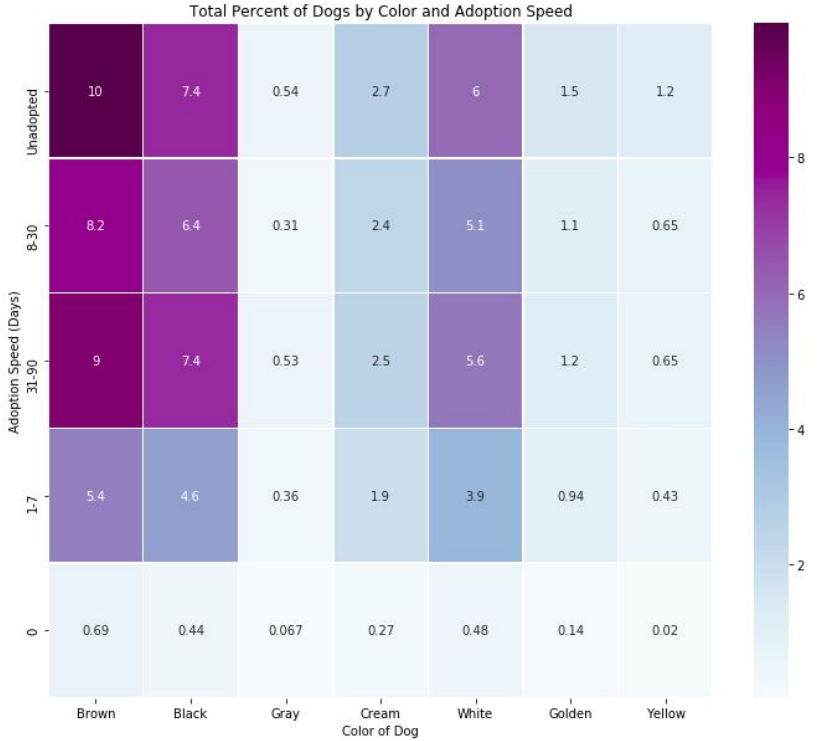
Exploratory Data Analysis - Fur Length



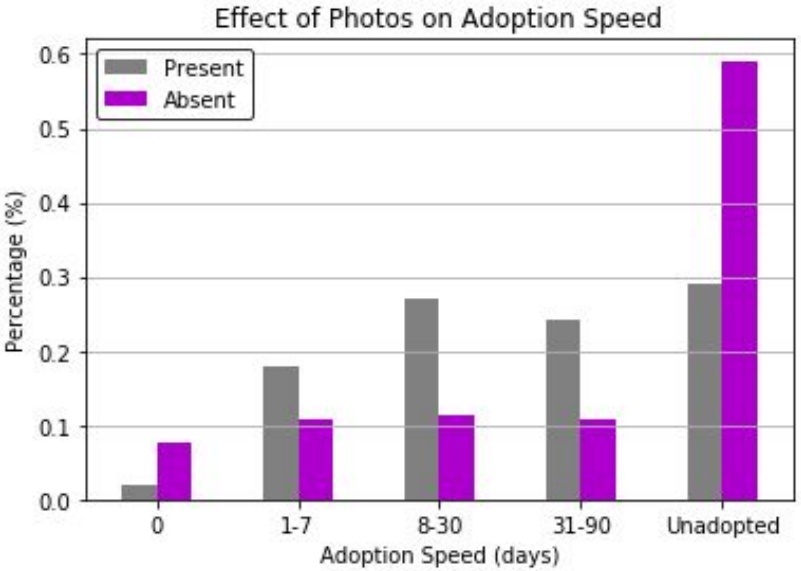


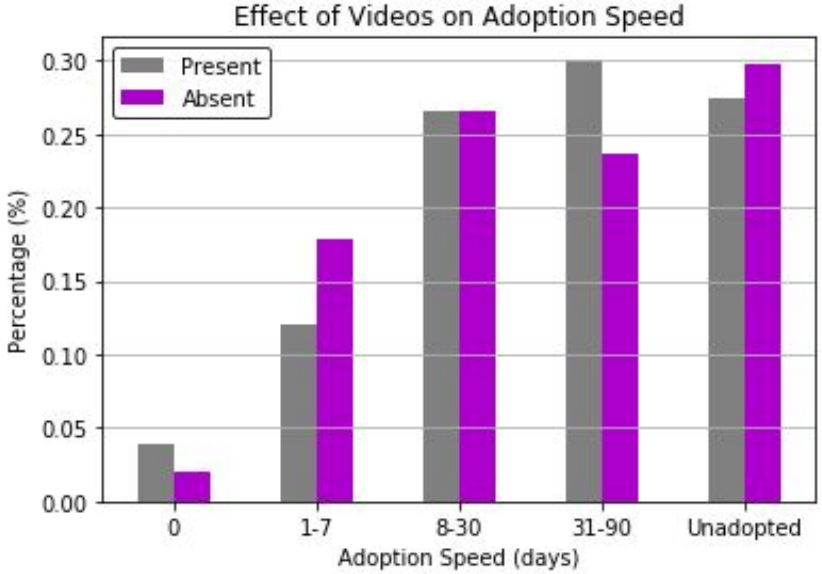
Exploratory Data Analysis - Color



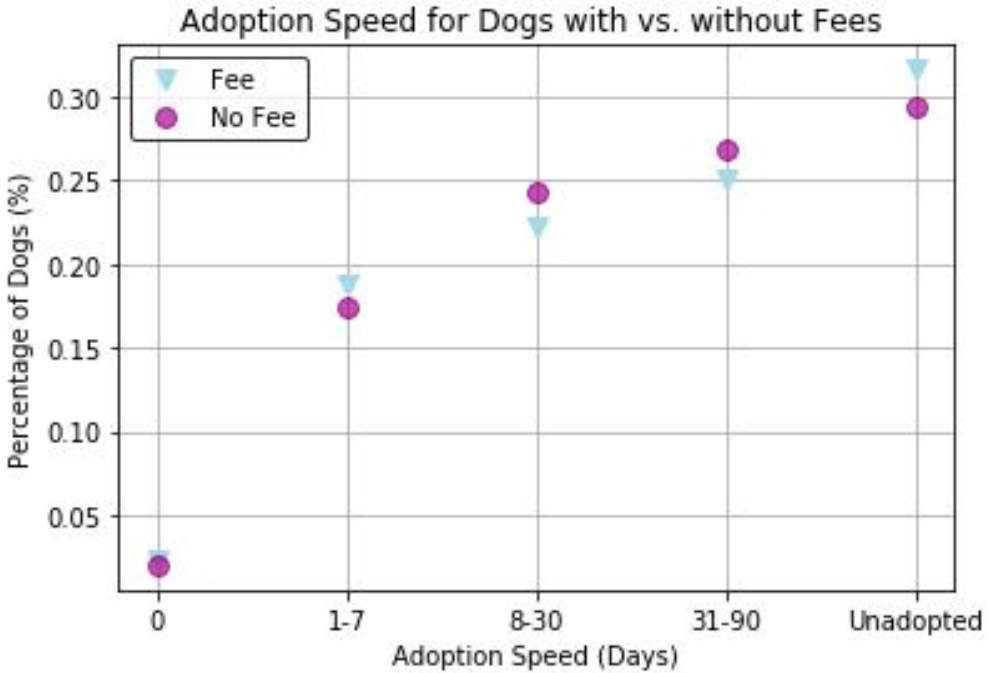


Exploratory Data Analysis - The Listing



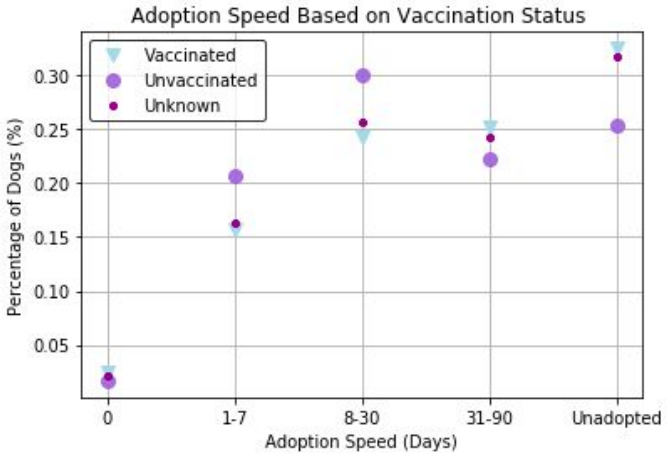


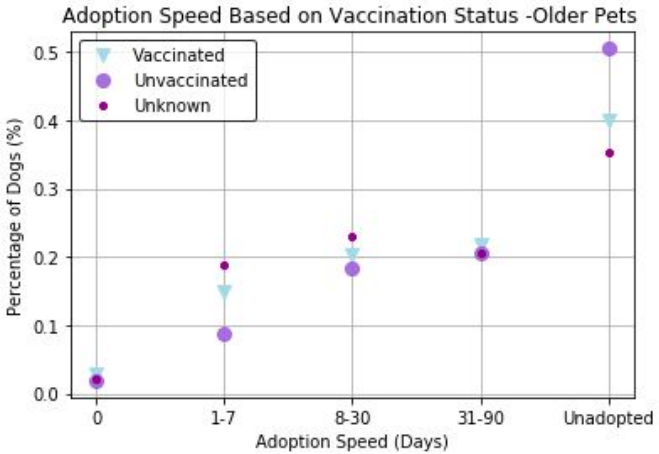
Exploratory Data Analysis - Fee



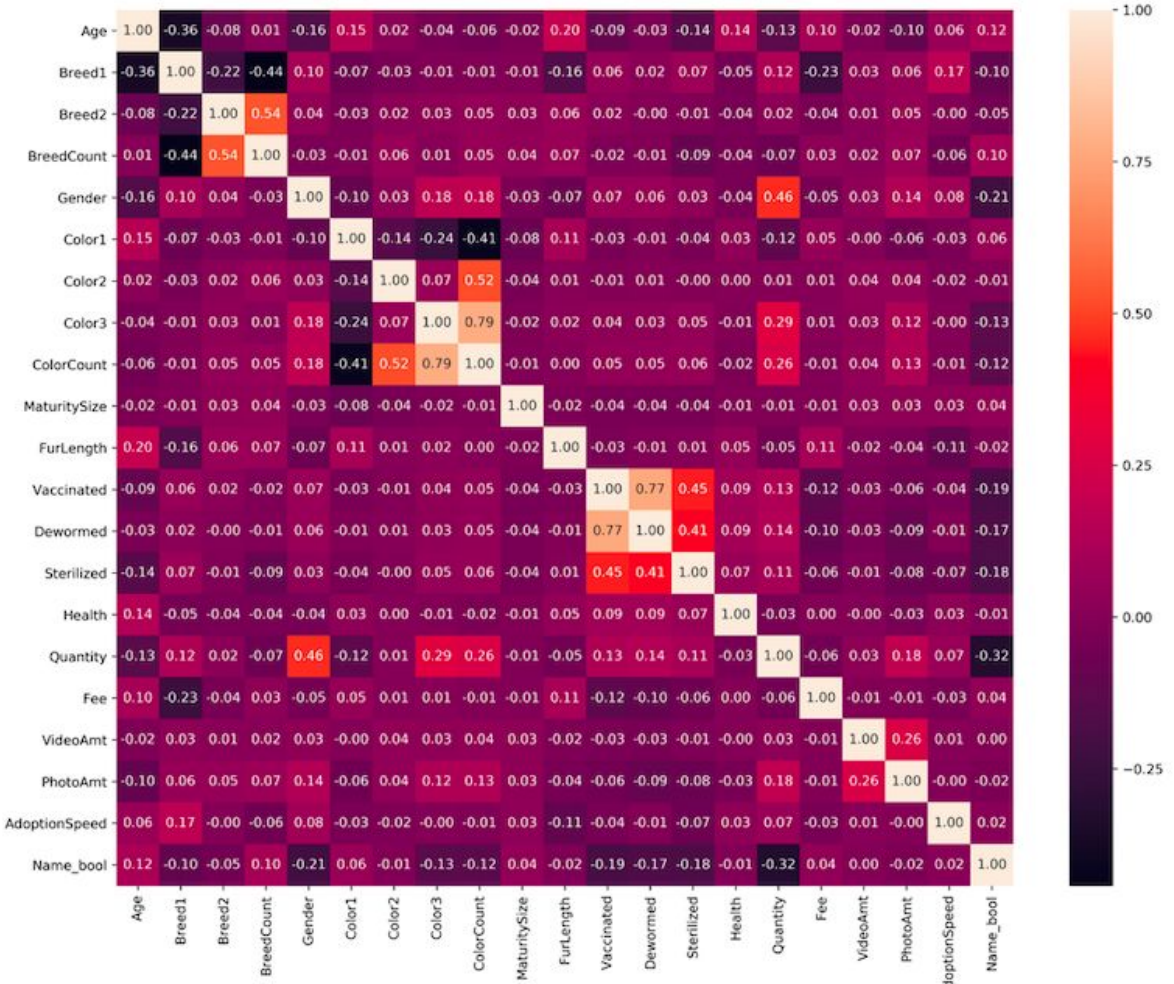
Exploratory Data Analysis - Health

- Health trends have less effect on adoption speed
  - Puppies medical history has a large effect on how health relates to adoption speed
- When removing puppies from the dataset, health only matters for pre-existing medical conditions, which is unsurprising as every other health feature studied is changeable by a veterinarian





Statistical Analysis



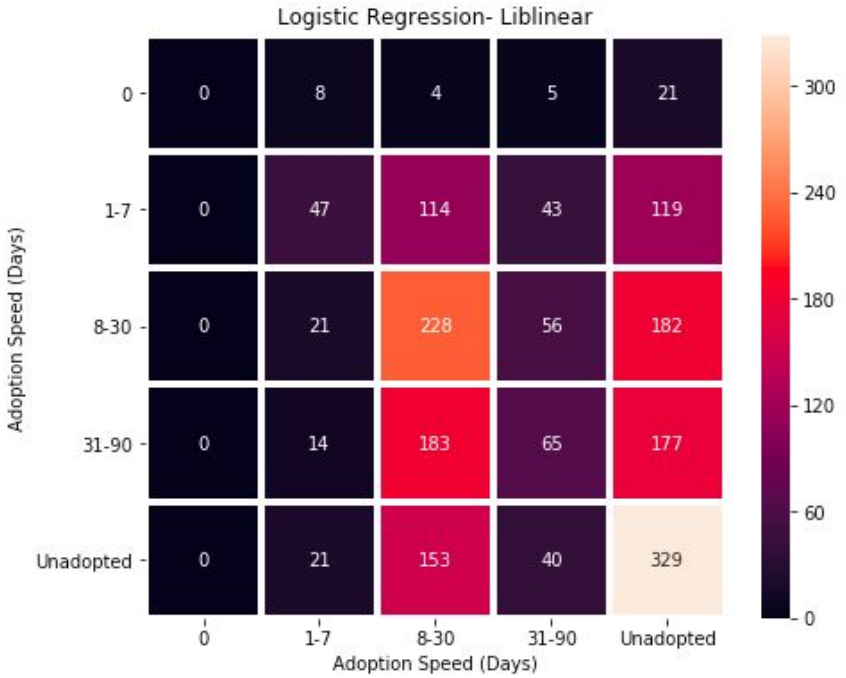
Model Development

Model	Accuracy Score			
	Same Day		No Same Day	
	Training	Validation	Training	Validation
Logistic Regression- lbfgs, restricted features *	0.39208	0.37146	0.40754	0.38645
Logstic Regression - lbfgs, all features	0.38579	0.39237	0.41541	0.37516
Logistic Regression - liblinear, all features	0.38579	0.39483	0.40369	0.37641
Naïve Bayes	0.23880	0.26568	0.33626	0.34003
Linear SVC	0.24918	0.36162	0.27247	0.24467
SVC	0.39071	0.40590	0.42099	0.44668
kNN	0.36885	0.39606	0.39698	0.41656
Random Forest	0.46940	0.47109	0.47571	0.48557
Light GBM	0.45027	0.46740	0.48130	0.47679

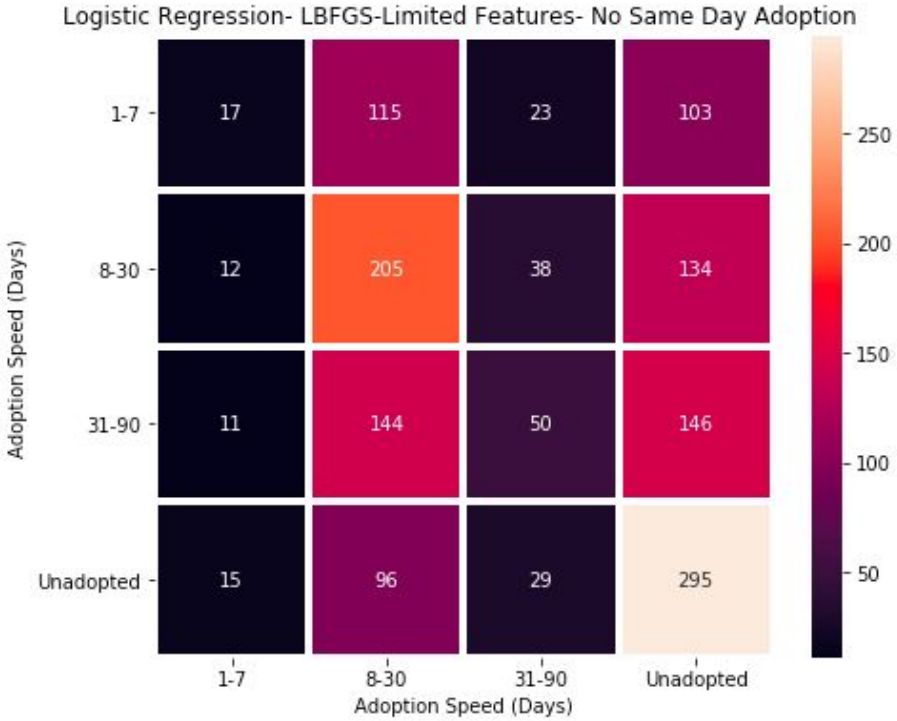
* Features Included: Rescuer ID, pet age, first breed classification, amount of photos

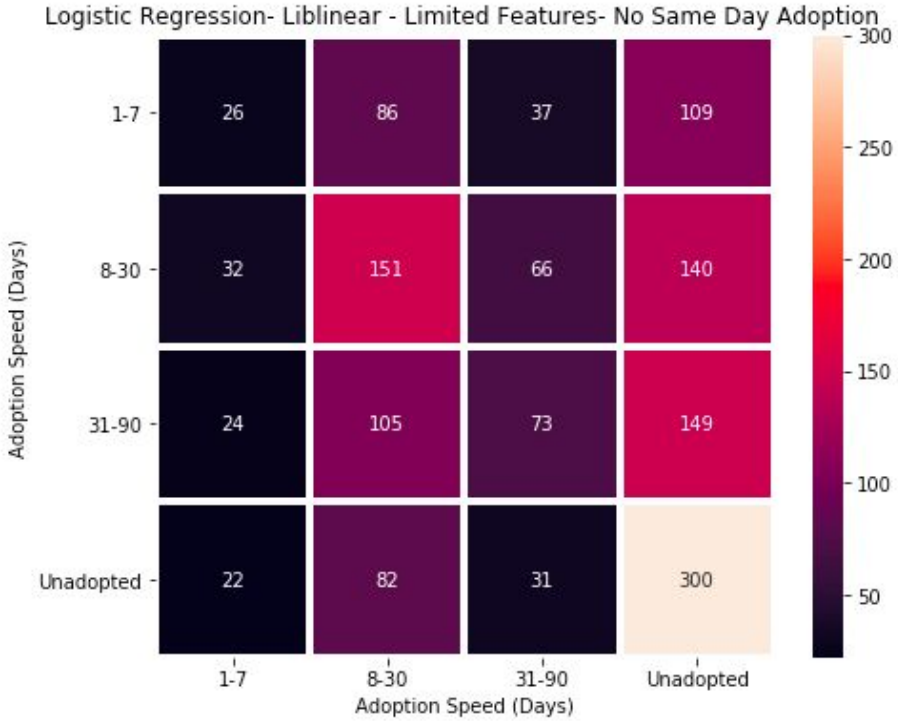
** Kaggle PetFinder Leadboard Winning Accuracy 0.45338

Logistic Regression

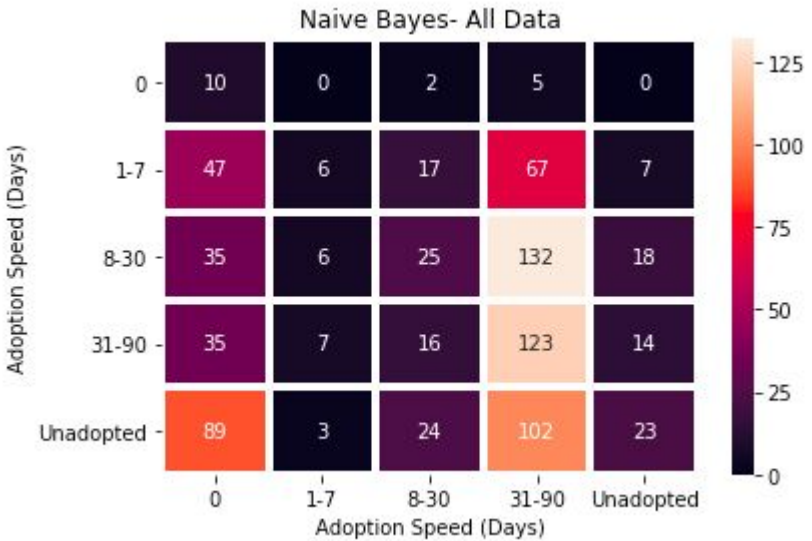


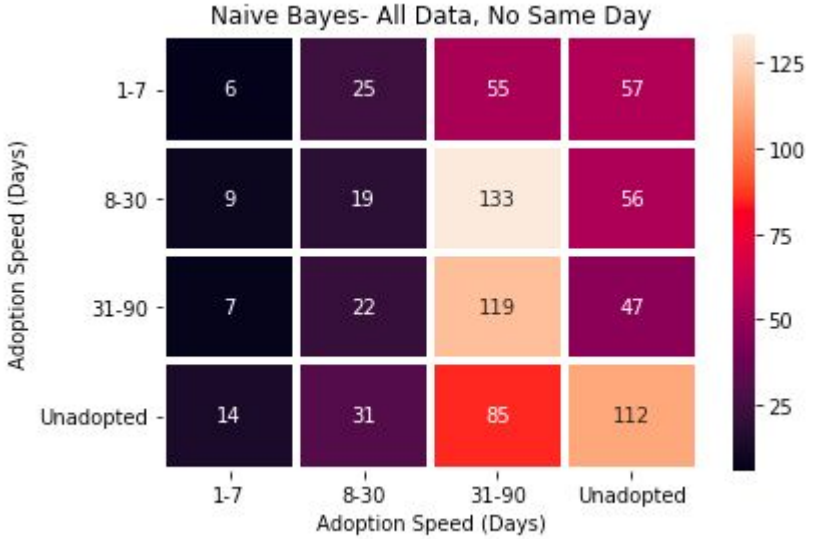
Logistic Regression - No Same Day Adoptions



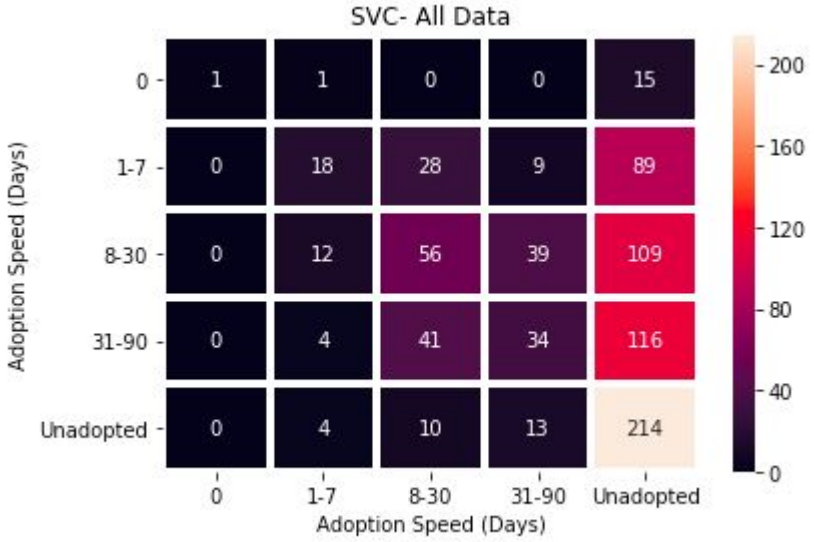


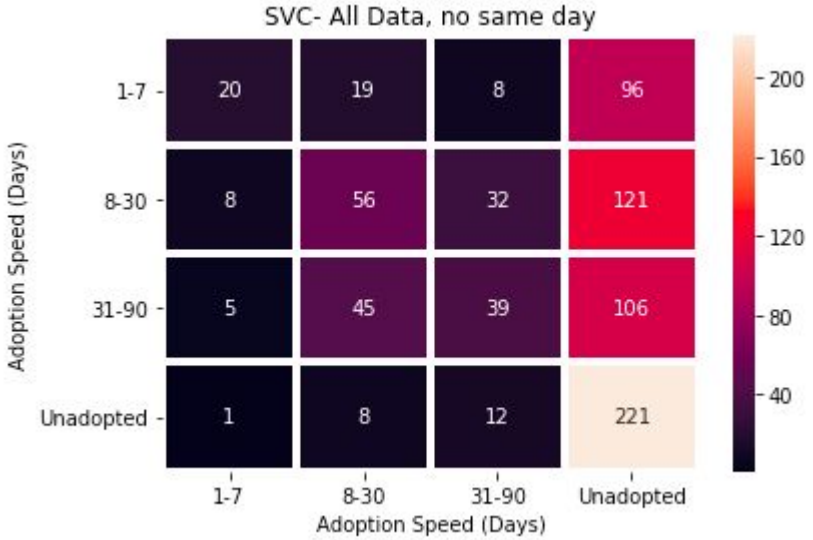
Naive Bayes


Naive Bayes- All Data
Adoption Speed (Days)
0
1-7
8-30
31-90
Unadopted
10 0 2 5 0
47 6 17 67 7
35 6 25 132 18
35 7 16 123 14
89 3 24 102 23
0
1-7
8-30
31-90
Unadopted
Adoption Speed (Days)
125
100
75
50
25
0

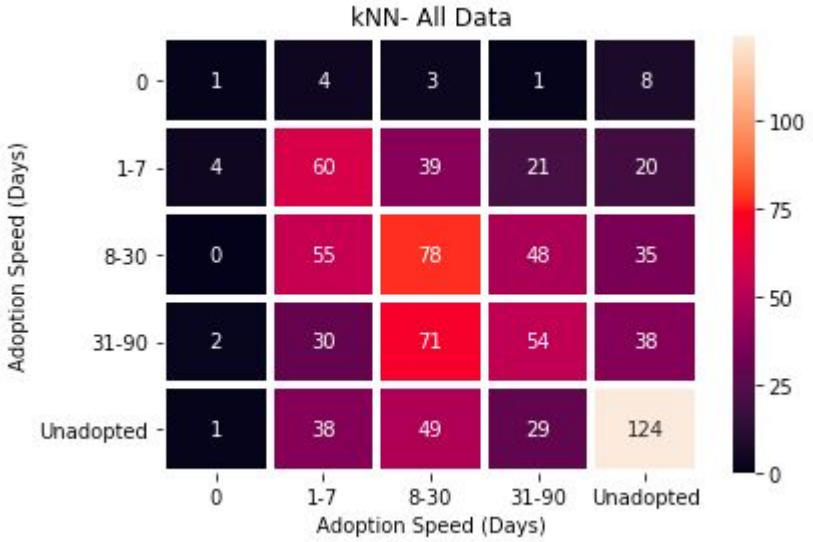

Naive Bayes- All Data, No Same Day
Adoption Speed (Days)
1-7
8-30
31-90
Unadopted
6 25 55 57
9 19 133 56
7 22 119 47
14 31 85 112
1-7
8-30
31-90
Unadopted
Adoption Speed (Days)
125
100
75
50
25

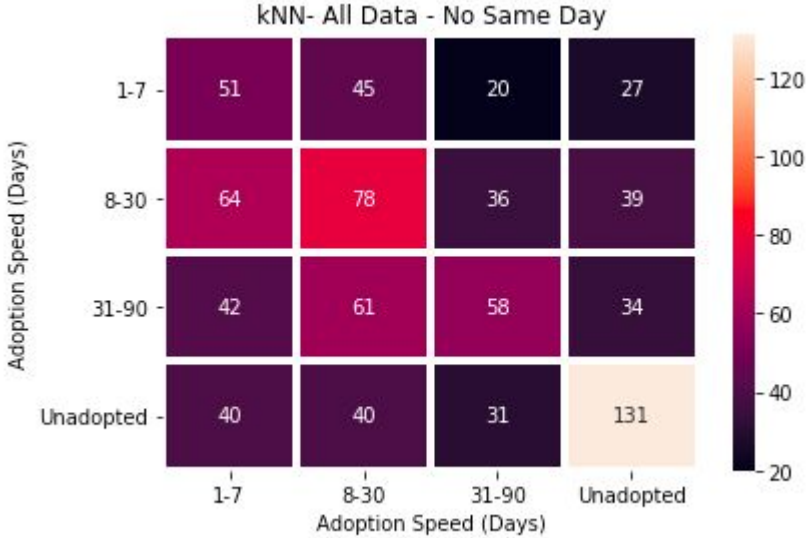
Support Vector Classification



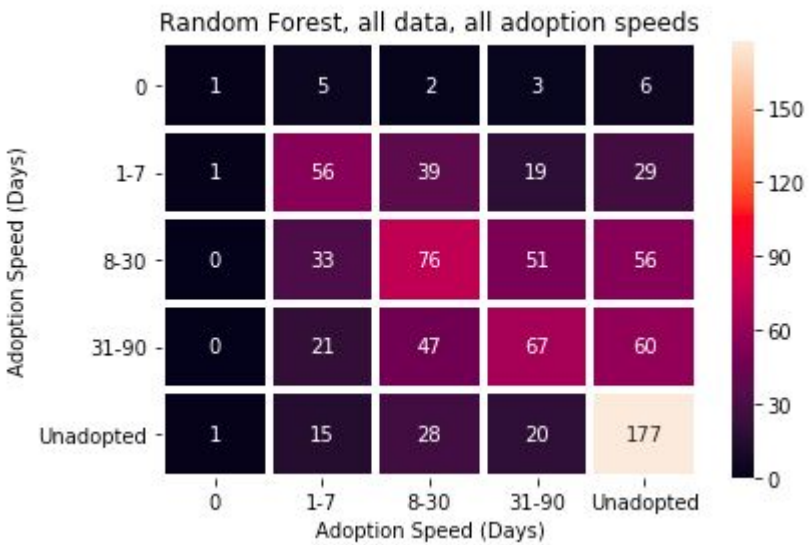


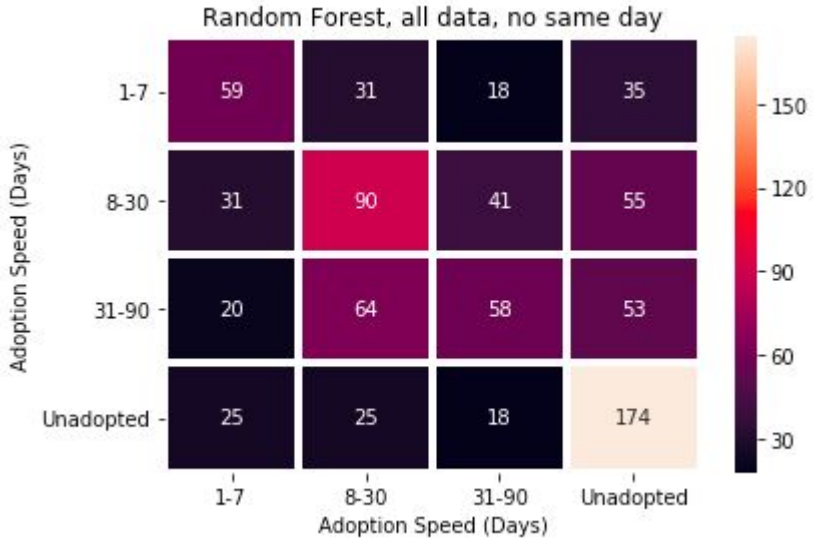
K Nearest Neighbors



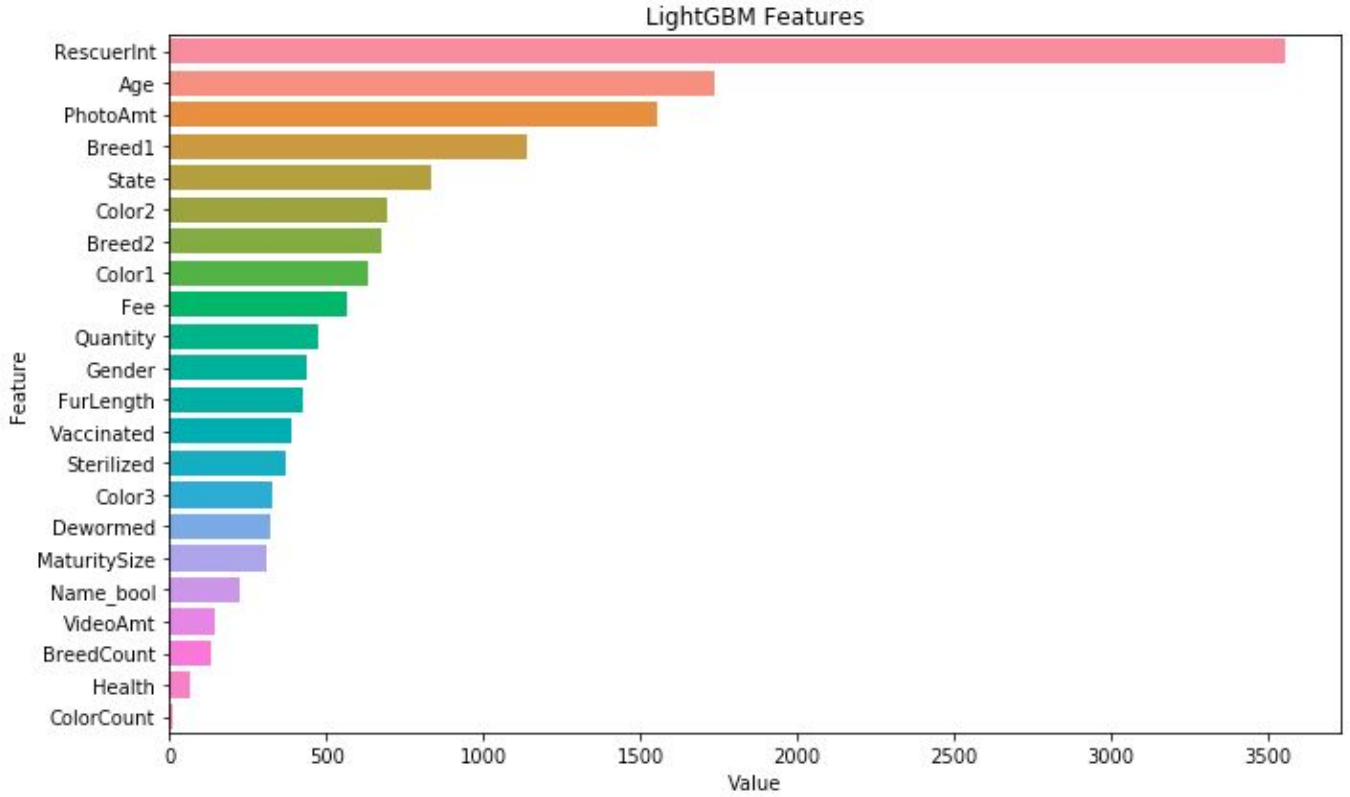


Random Forest

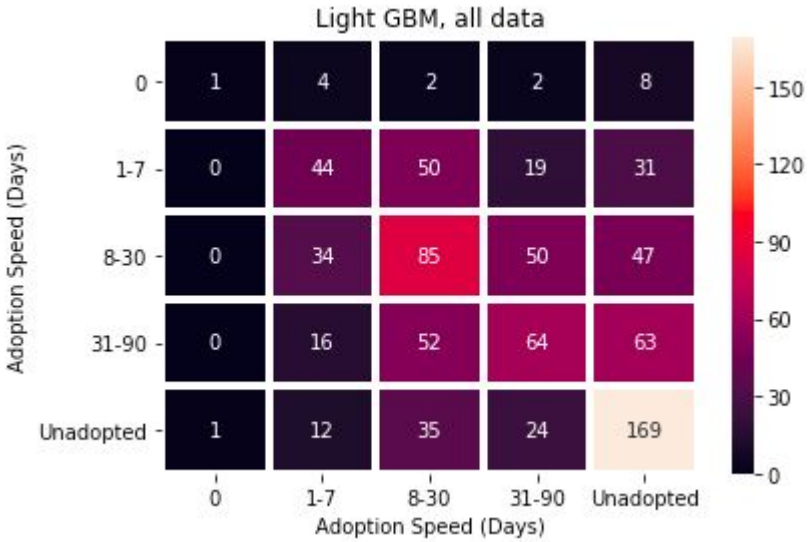


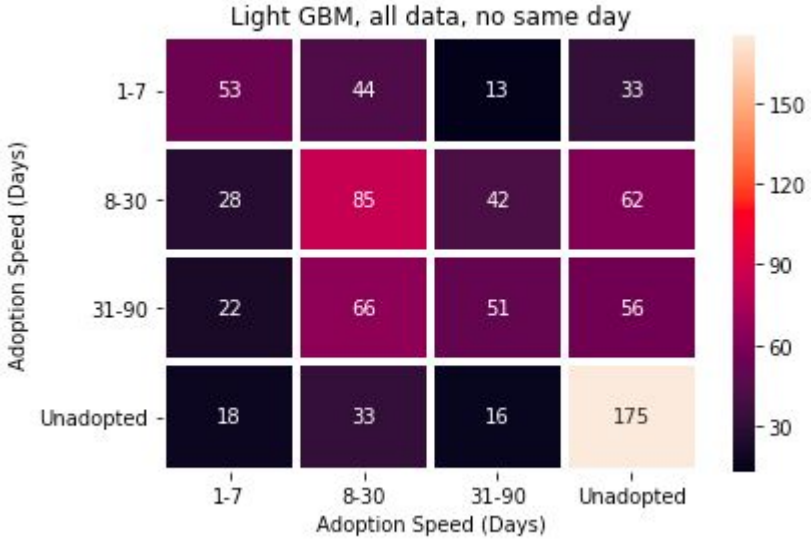


Light GBM - Features Analysis

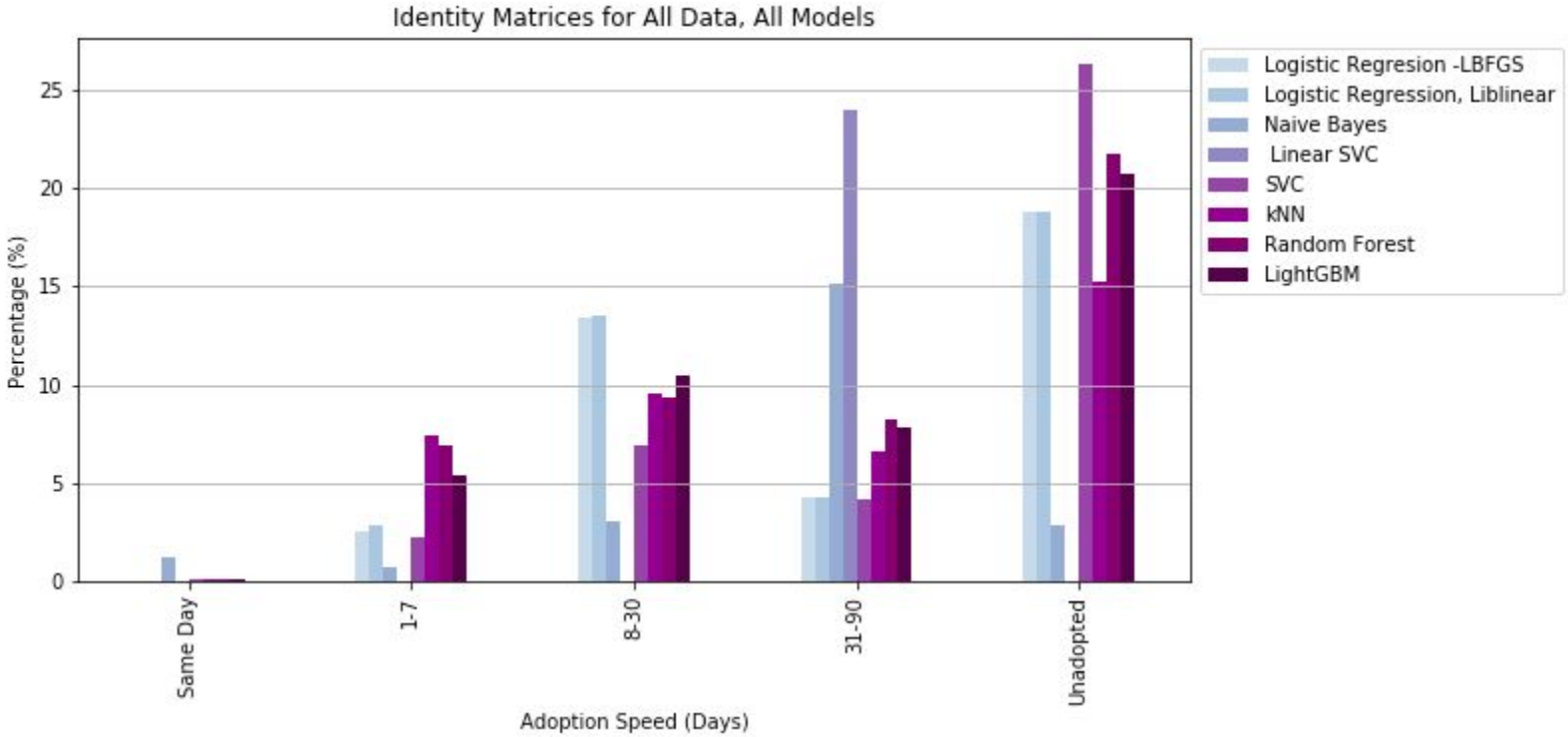


Light GBM

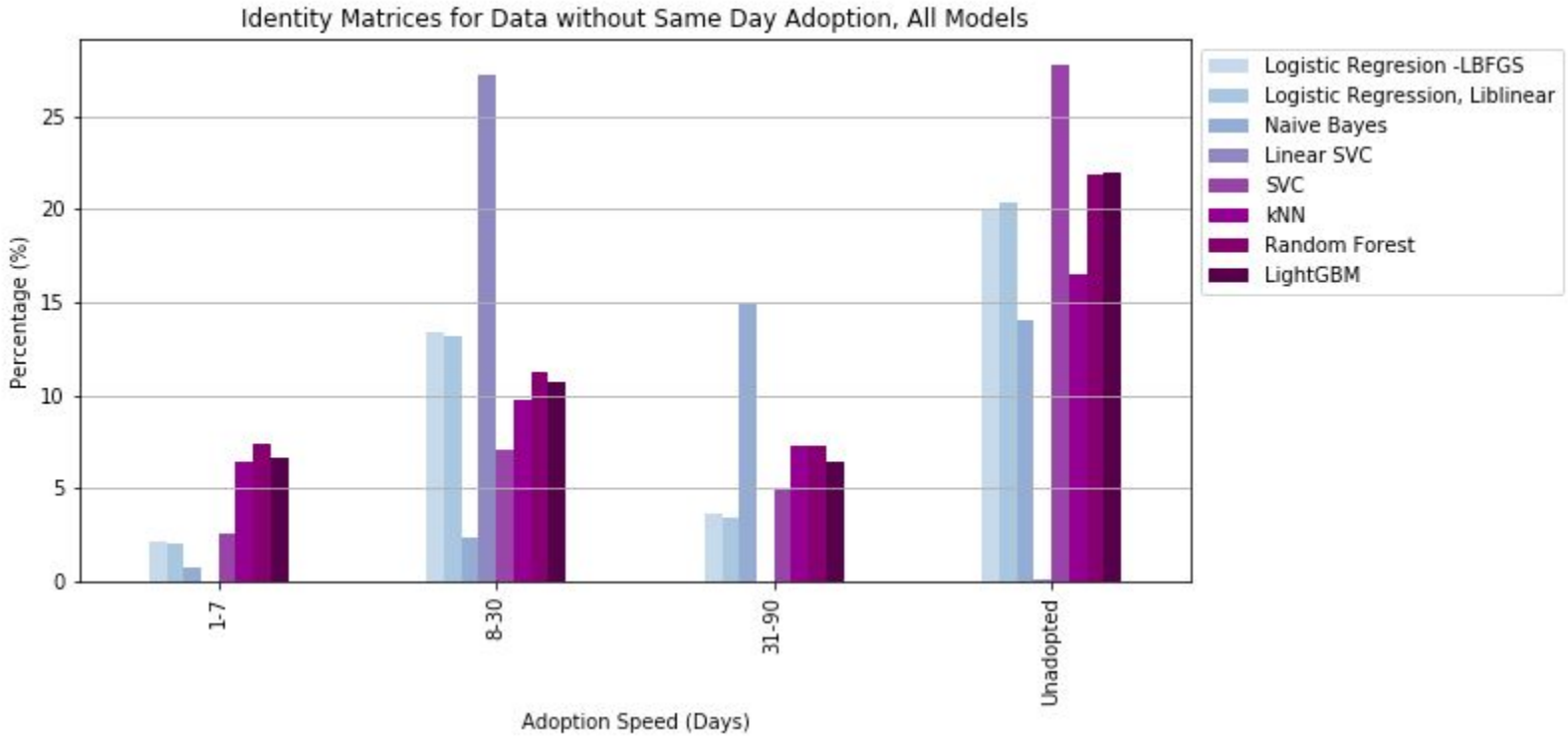

Light GBM, all data
Adoption Speed (Days)
0
1-7
8-30
31-90
Unadopted
1 4 2 2 8
0 44 50 19 31
0 34 85 50 47
0 16 52 64 63
1 12 35 24 169
0
1-7
8-30
31-90
Unadopted
Adoption Speed (Days)
150
120
90
60
30
0


Light GBM, all data, no same day
Adoption Speed (Days)
1-7
8-30
31-90
Unadopted
53 44 13 33
28 85 42 62
22 66 51 56
18 33 16 175
1-7
8-30
31-90
Unadopted
Adoption Speed (Days)
150
120
90
60
30

Model Results- All Pets



Model Results - No Same Day Adoptions



Recommendations

1. Focus more on which rescue has success with adopting out a specific type/breed of animal
2. The best chance of predicting a same day adoption is to use the logistic regression model using a liblinear solver.
3. The second step in analyzing a new pet during intake should be to run LinearSVC, and keep those pets who will be adopted in the first month if capacity is available.
4. To determine those pets who will require a more medium duration stays (<90 days), utilize the Naive Bayes model
5. To predict those pets who will require long term care, and possibly rehoming to an more targeted rescue, specialized training, etc. utilize the following models: SVC, kNN, Random Forest, or Light GBM.
6. An averaged prediction for all models can also be utilized.

Dataset Limitations

1. Understanding interactions between listings opening and closure- how is “adoption speed” defined. Are animals that go unadopted simply not updated listings?
2. Understanding how unique identifiers are assigned - is it possible for a pet to change rescues and have multiple listings for the same animal with different outcomes?
3. Understanding how breed is determined for listings - is it appearance based or DNA based?

Opportunities for Future Work

1. Repeat with pull from US PetFinder API and compare relationships
2. Explore rescuer specific models for each agency
3. Explore different relationships within the data - breed, description, etc.
4. Explore area statistics and correlate to data: state legal policies with regards to animals, economic and housing statistics
5. Develop a model to reliably clean the “name” feature for application acros the entire dataset
6. Expand to study effects and reactions with adoption for cats
7. Develop a predictor model based on location, rescue history, etc. to use fore ach pet’s intake

Thank You

Caitlin Jansson
caitlinjejansson@gmail.com
[Project GitHub Repository]