Capstone Project 1 – Pet Adoption Prediction

Isadora P. Thisted

Springboard

May 4th, 2020

Abstract

Approximately 6.5 million companion animals enter U.S. animal shelters nationwide every year. Limited space forces some animals to be euthanized, so it is important that these organizations place existing animals in permanent homes.

While euthanasia rates in the U.S have significantly declined in the last 10 years due to spaying, neutering and the increased popularity of rescue adoption it is estimated that over 1.5 million animals are still euthanized each year in the U.S.

This project will use data from Petfinder.my (Malaysia's leading online animal welfare platform) which is available on Kaggle. The data consists of metadata with 25 parameters as well as images for approximately 15,000 pets as well as their adoption speeds.

The existing data will be used to predict each animal's adoption speed and explore relationships between adoptability and the various parameters available from the pets' profiles.

The data will be augmented with breed-specific characteristics extracted from the website "dogtime.com" to explore whether canine breed traits impact adoption rates.

Pet shelters and rescue organizations can use the findings of the report and the classification model to enhance their own online pet profiles. This information can also be used as an aid when making decisions on which animals might take priority when being placed in foster care or taken to adoption events.

Keywords: Data Analysis, Pet Adoption, Machine Learning

Data Wrangling

Two datasets were used for this project:

- Petfinder.my data with animal profiles and adoption speeds
- Dog breed traits scraped from "DogTime.com"

The section below explains each of the datasets in detail and the steps that were taken to clean and join them.

Adoption Data

This is the main dataset for this project. The pet adoption data from Kaggle/PetFinder is approximately 2GB in size, with 15,000 pet entries on the training set and 4,000 entries on the testing set. The data is a mixture of text (animal profile descriptions), tabular data (with 25 parameters) and pet images. The pet images have been run through Google's Vision API, and the resulting metadata is available, providing analysis on Face Annotation, Label Annotation, Text Annotation, and Image Properties.

The dataset was clean, and the only missing variables were found in the pet name and pet description columns. The names and description data will not be used for analysis or prediction, so these null values did not need to be filled or removed.

No outliers were found for this data set.

- Data source: https://www.kaggle.com/c/petfinder-adoption-prediction/data

Breed Characteristics Data

Dog breed characteristics were scraped from the website "dogtime.com" for each of the 334 breeds available in the website. The result was a table containing each breed as an entry and thirty unique breed characteristics with corresponding star ratings (on a scale of 1 to 5) for traits

such as friendliness towards strangers, sensitivity level, energy level, and drooling potential. This information will be utilized to explore any associations between various breed traits and adoption rates.

The dog breeds scraped from dogtime.com did not exactly match those existing in the breed labels for the PetFinder adoption data. Several of the names were manually corrected so they would be identical, matching them allowed the breed traits and adoption data frames to be joined using the dog's breed name.

Breeds that did not exist in both databases have missing (null) traits data. For this analysis, the dogs whose breeds were not able to be matched to the corresponding row in the breed traits data will be ignored. The majority of dogs in the adoption database are actually labeled as simply "mixed breed" and had missing data once the breed traits information was joined. While this resulted in a significant amount of missing breed traits data, our sample is still large enough (> 2000 entries) and should be sufficient to derive meaningful insights. As a future exercise, the metadata annotations produced by running the pet images through Google's Vision API might allow us to fill in the breed composition for some of the dogs labeled as "mixed breed" or even those that were incorrectly labeled.

Cat breed traits were not available so only canine data will be used to explore associations between the expected breed traits and adoption rates.

No outliers were found for this data set.

- Data source: https://dogtime.com/dog-breeds/profiles/

Exploratory Data Analysis

Initial Dataset

The PetFinder data includes text, tabular and image data for the pet listings. Exploratory data analysis can be used to derive relationships between the adoption speed and the various parameters available from the pet's profile and suggest improvements to the profiles that would increase the animal's adoptability. Some profiles represent a group of pets. In this case, the speed of adoption is determined by the speed at which all of the pets are adopted. Below is the list of available parameters for each pet entry:

- *PetID* - Unique hash ID of pet profile
- ***AdoptionSpeed*** - Categorical speed of adoption. This is the value we are trying to predict. See below for a description:
    - 0 - Pet was adopted on the same day as it was listed.
    - 1 - Pet was adopted between 1 and 7 days (1st week) after being listed.
    - 2 - Pet was adopted between 8 and 30 days (1st month) after being listed.
    - 3 - Pet was adopted between 31 and 90 days (2nd & 3rd month) after being listed.
    - 4 - No adoption after 100 days of being listed. (There are no pets in this dataset that waited between 90 and 100 days)
- *Type* - Type of animal (1 = Dog, 2 = Cat)
- *Name* - Name of pet (Empty if not named)
- *Age* - Age of pet when listed, in months
- *Breed1* - Primary breed of pet (Refer to BreedLabels dictionary)
- *Breed2* - Secondary breed of pet, if pet is of mixed breed (Refer to BreedLabels dictionary)
- *Gender* - Gender of pet (1 = Male, 2 = Female, 3 = Mixed, if profile represents group of pets)
- *Color1* - Color 1 of pet (Refer to ColorLabels dictionary)
- *Color2* - Color 2 of pet (Refer to ColorLabels dictionary)
- *Color3* - Color 3 of pet (Refer to ColorLabels dictionary)

- *MaturitySize* - Size at maturity (1 = Small, 2 = Medium, 3 = Large, 4 = Extra Large, 0 = Not Specified)
- *FurLength* - Fur length (1 = Short, 2 = Medium, 3 = Long, 0 = Not Specified)
- *Vaccinated* - Pet has been vaccinated (1 = Yes, 2 = No, 3 = Not Sure)
- *Dewormed* - Pet has been dewormed (1 = Yes, 2 = No, 3 = Not Sure)
- *Sterilized* - Pet has been spayed / neutered (1 = Yes, 2 = No, 3 = Not Sure)
- *Health* - Health Condition (1 = Healthy, 2 = Minor Injury, 3 = Serious Injury, 0 = Not Specified)
- *Quantity* - Number of pets represented in profile
- *Fee* - Adoption fee (0 = Free)
- *State* - State location in Malaysia (Refer to StateLabels dictionary)
- *RescuerID* - Unique hash ID of rescuer
- *VideoAmt* - Total uploaded videos for this pet
- *PhotoAmt* - Total uploaded photos for this pet
- *Description* - Profile write-up for this pet. The primary language used is English, with some in Malay or Chinese

Since our goal is to identify the animals at highest risk of no adoption an additional column (labeled *NoAdoption*) was created to easily identify those pets in category 4 of *Adoption Speed*, if the pet was in category 4 a value of 1 was assigned to the *NoAdoption* parameter, otherwise it was assigned a value of zero.

Additionally, to explore numeric relationships in more depth, the number of days until adoption was also added as a column (*DaysToAdoption*), using an average value of the range of number of days until adoption for each category in *AdoptionSpeed*. It is worth noting that since category 4 doesn't provide us with a date range like the other categories 150 days were assumed, which might lead to some inaccuracy when exploring relationships between variables.

The overall distribution of pet adoption speeds on the training set is shown in the figure below:

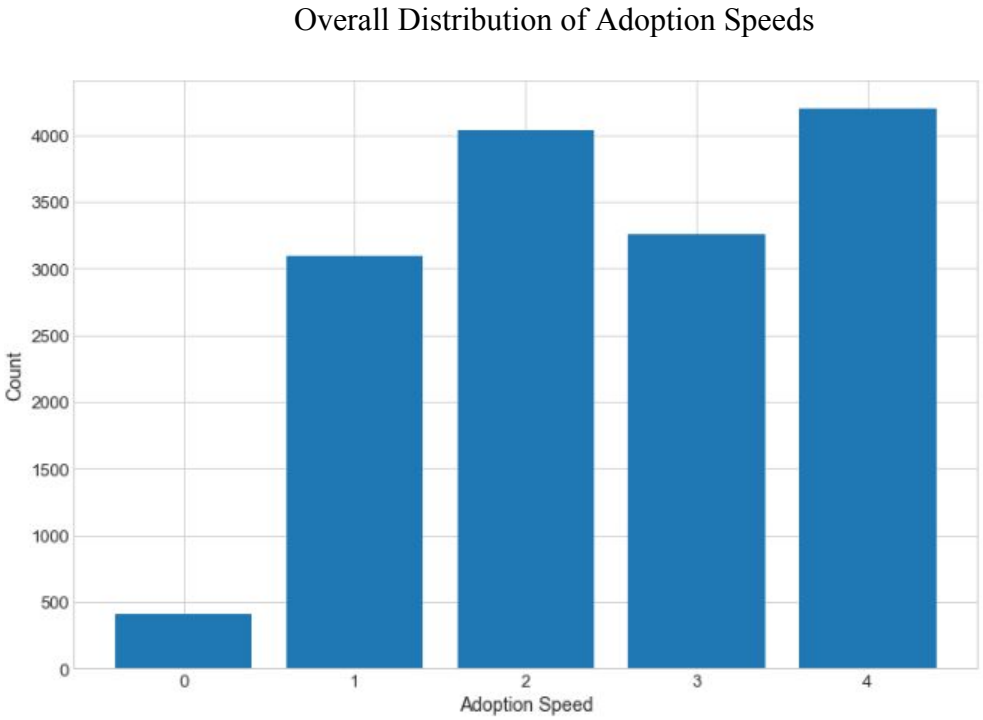

Figure 1. Distribution of adoption speeds on the training set

As seen in the histogram above, only a small proportion of animals gets adopted in the fastest category (same-day adoption) and while the majority of animals were adopted within 90 days (categories 0 through 3) over 4,000 animals took longer than 100 days or were not adopted (category 4). In the sections that follow, we explore relationships between the available parameters and the animal's adoption speed.

Animal Type and Gender

Here we explore the dependency of adoption speed based on the type and gender categories. The animal types in the data were limited to cat or dog and the gender was broken down into three categories (male, female, and mixed). The "Mixed" gender label was used to

account for pet listings with multiple pets and will be ignored for this analysis. The *DaysToAdoption* column was used along with the categorical variable *Type* to generate the boxplots below.

Adoption Speed broken down by Animal Type and Gender

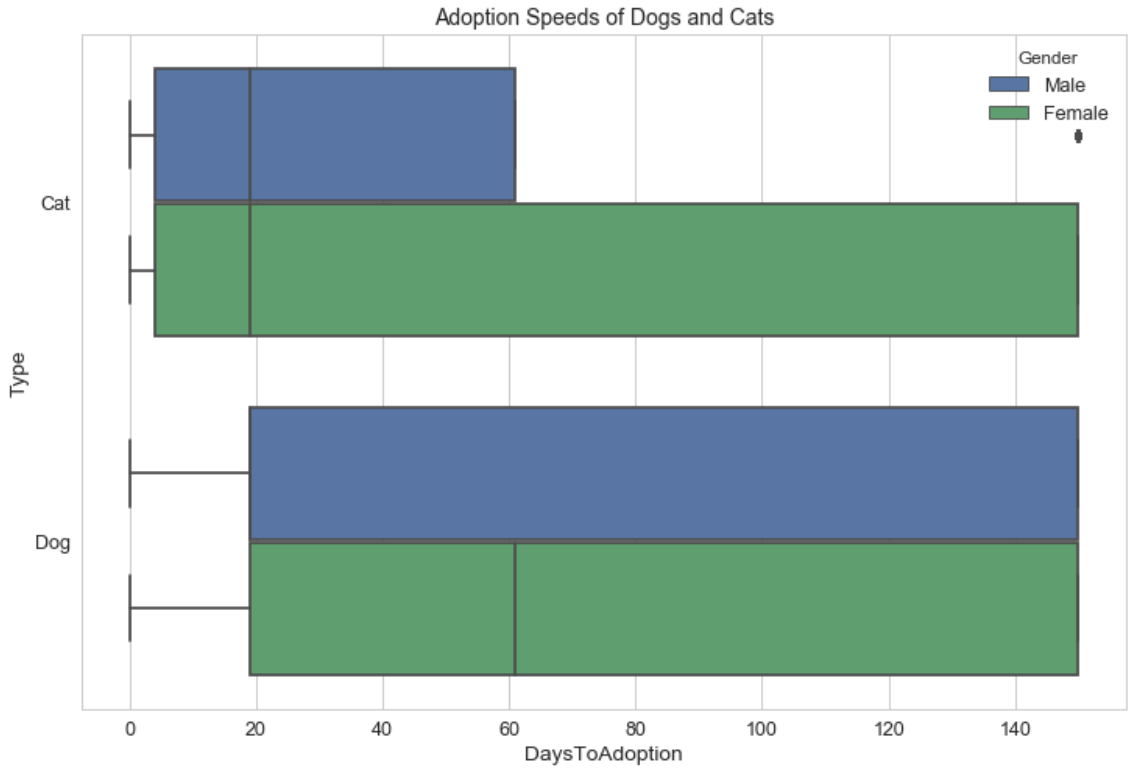

Figure 2. Boxplot showing days until adoption of cats and dogs

Cats seem to have an easier time getting adopted. This difference could be due to the fact that cats are lower maintenance animals and don't require as much tending to. As can be seen above, the median adoption speed of cats is roughly one third that of dogs with 50% of cats adopted within 20 days as opposed to a median of over 60 days for dogs. The 25th percentile for cats is under 10 days and for dogs, it is closer to 20 days. The 75th percentiles for both animals are the same at 150 days (category 4 of adoption speed).

Gender seems to play as much an impact as animal type, with males having a clear advantage of getting adopted quickly. The median adoption speed of females is 3 times as large (60 days) as the median adoption speed for males (20 days).

Surprisingly, gender seems to have a much larger impact on dogs as compared to cats. For cats, the median adoption speed is actually the same for males and females.

A t-test was performed and the resulting p-values (<0.05) confirmed that both gender and type are indeed statistically significant factors impacting adoption speeds.

Animal Age

A quick check for correlation using points biserial between the animals' age data and the NoAdoption data (1 or 0 based on whether the animal had been adopted by the 10th day) proves an association exists (as evidenced by the 0.11 correlation factor and low p-value):

PointbiserialrResult(correlation=0.11043932612065595, pvalue=6.616521713632641e-42)

The age for the animals is given in months in the original dataset, it was converted to years to allow for a more intuitive comparison. Below a scatter plot and linear regression model can be seen.

Animal's Age vs. Adoption Speed

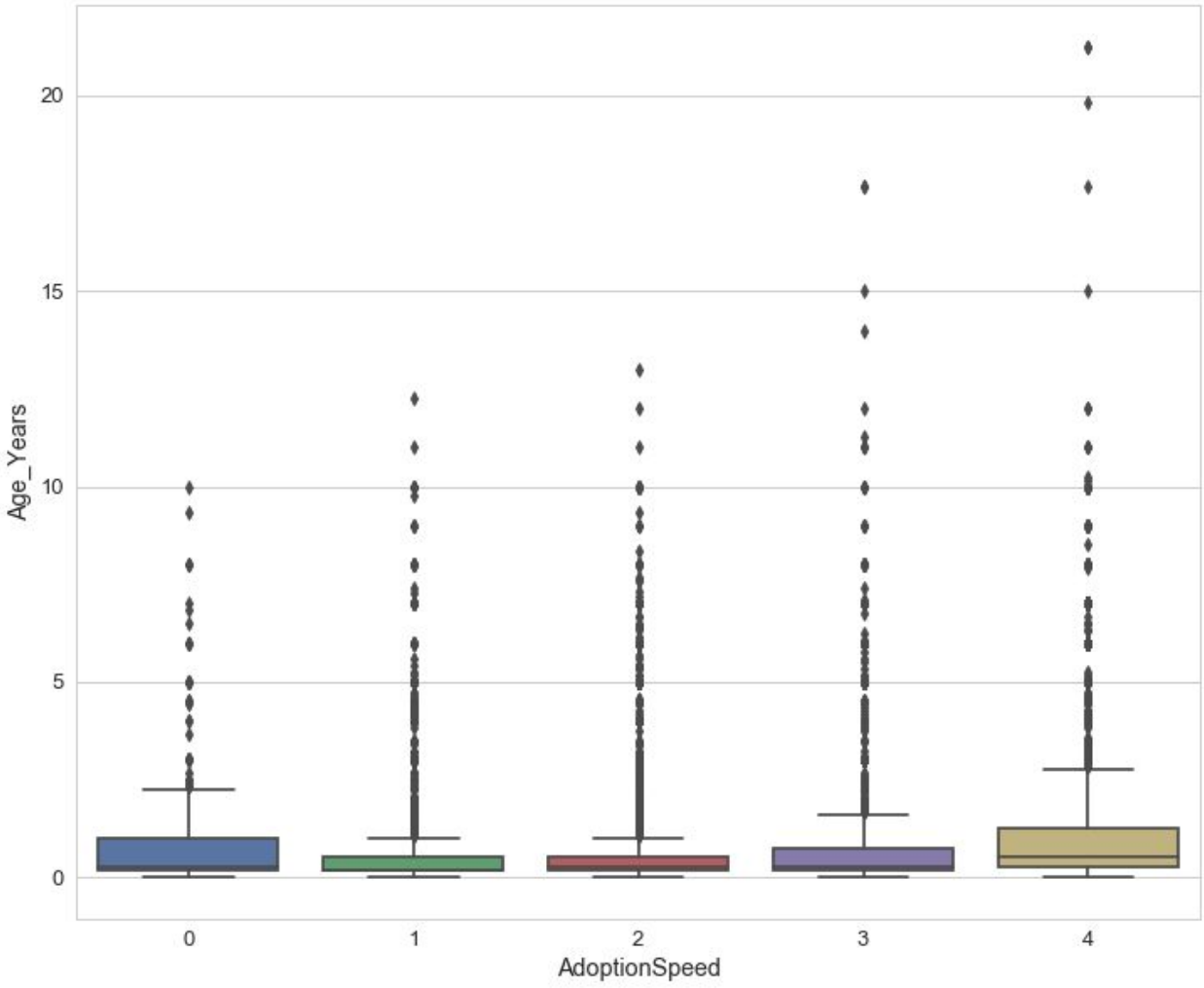

Figure 5. Boxplot for Adoption Speed vs Age

There is an evident association between the animals' age and the number of days they took to get adopted, the younger the animal the more likely it is to get adopted quickly.

Animal Health

It is expected that an animal's health will impact the speed at which it gets adopted, but to what extent? Below we calculate the percentages of adoption speed for the animals in the three health categories (healthy, minor injury, and serious injury).

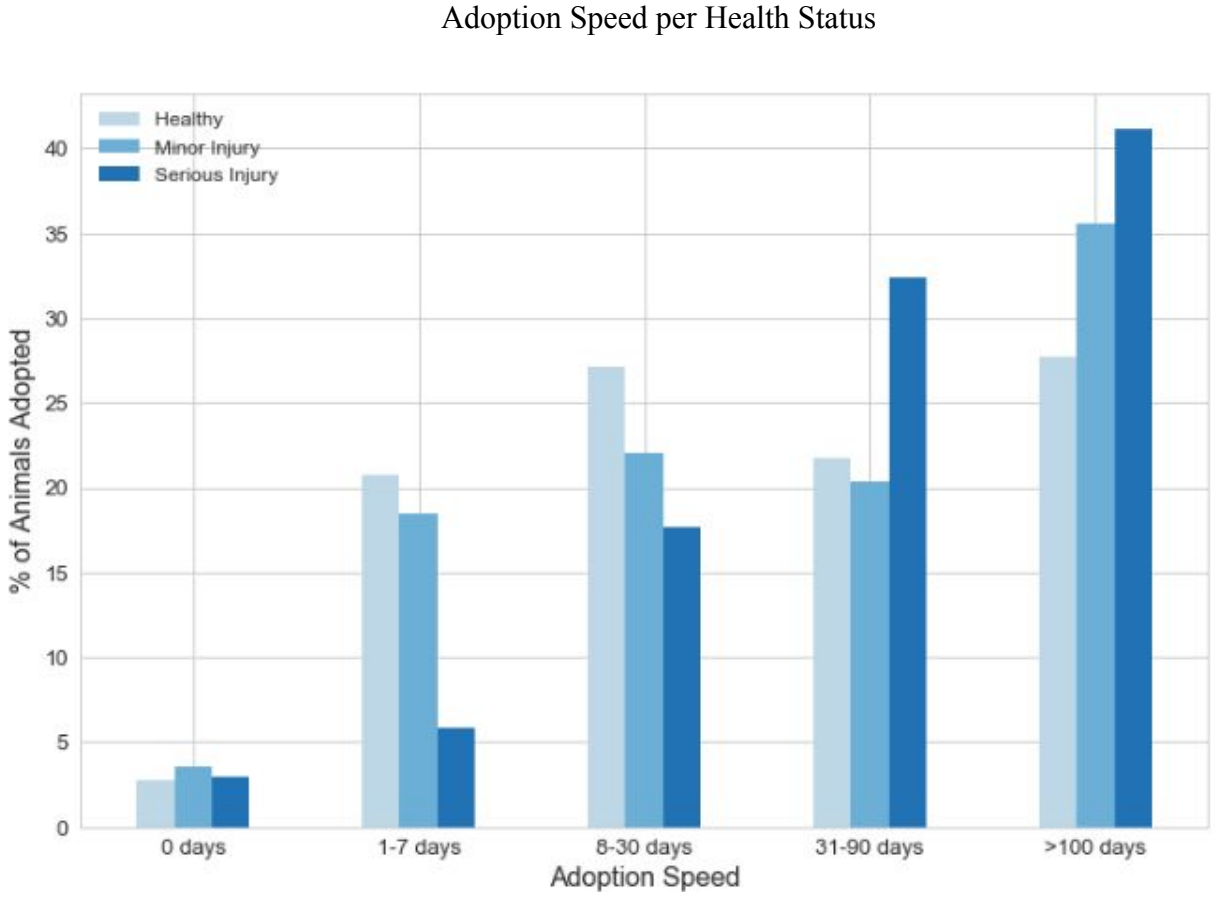

Figure 6. Percentage of Animals Adopted according to Adoption Speed and Health Status

Over 41% of animals with serious injuries are not adopted after 100 days, for healthy animals that number is much lower at 28%. Minor injuries also seem to impact the rate at which animals get adopted but to a lesser degree.

To verify that the difference in adoption speeds due to an animal's health is statistically significant we perform Tukey's HSD (honestly significant difference) test:

Multiple Comparison of Means - Tukey HSD,FWER=0.05

==========

group1	group2	meandiff	lower	upper	reject
Healthy	Minor Injury	9.8641	3.4739	16.2542	TRUE
Healthy	Serious Injury	24.2704	0.5973	47.9434	TRUE
Minor Injury	Serious Injury	14.4063	-10.0604	38.8731	FALSE

There is a statistically significant difference between healthy animals and the remaining groups (animals with minor and serious injuries). However, a statistically significant difference was not found between animals with minor and serious injuries, so we can say that injury does affect adoption speed, but we cannot say confidently that the extent of the injury does as well.

Dog Breeds

Below is a summary table for the 15 dog breeds with the highest total counts from the training dataset, for each of the breeds the total counts and also percentages for each adoption speed category are shown.

Table 1

Breakdown of Dog Adoptions by Adoption Speed and Breed

Breed Label	**Total (Count)**	Adoption Speed 0 (%)	Adoption Speed 1 (%)	Adoption Speed 2 (%)	Adoption Speed 3 (%)	Adoption Speed 4 (%)
Mixed Breed	**5923**	1.37	14.50	26.59	23.32	34.22
Labrador Retriever	**230**	3.04	19.57	29.57	32.17	15.65
Shih Tzu	**189**	4.23	34.92	28.04	18.52	14.29
Poodle	**169**	9.47	30.18	27.81	17.75	14.79
Airedale Terrier	**162**	0.62	17.90	28.40	32.10	20.99
Golden Retriever	**151**	6.62	29.14	28.48	21.85	13.91
German Shepherd Dog	**100**	1.00	24.00	36.00	23.00	16.00
German Spitz	**93**	3.23	12.90	27.96	31.18	24.73
Beagle	**90**	7.78	27.78	16.67	32.22	15.56
Rottweiler	**88**	5.68	30.68	25.00	25.00	13.64
Standard Schnauzer	**69**	8.70	30.43	18.84	23.19	18.84
Jack Russell Terrier	**68**	0.00	19.12	25.00	35.29	20.59
Miniature Pinscher	**67**	1.49	22.39	29.85	29.85	16.42
Doberman Pinscher	**62**	3.23	20.97	20.97	41.94	12.90
Siberian Husky	**62**	3.23	33.87	25.81	19.35	17.74

From the summary table above several insights become clear:

- Dogs simply labelled as 'mixed breed' are the most abundant and constitute over 1/3 of all dogs in the dataset
- Among the top 15 highest count breeds dogs labeled as 'mixed breed' have one of the highest percentages of no adoption after 100 days at 34.2%
- From the top 15 breeds with the highest count, Poodles and Standard Schnauzers have high rates of same-day adoption at 9.5 and 8.7% respectively

For a clearer summary, adoption speeds of 'mixed breed' animals are compared against all other identified breeds in the plot below.

Adoption Speeds Per Breed:

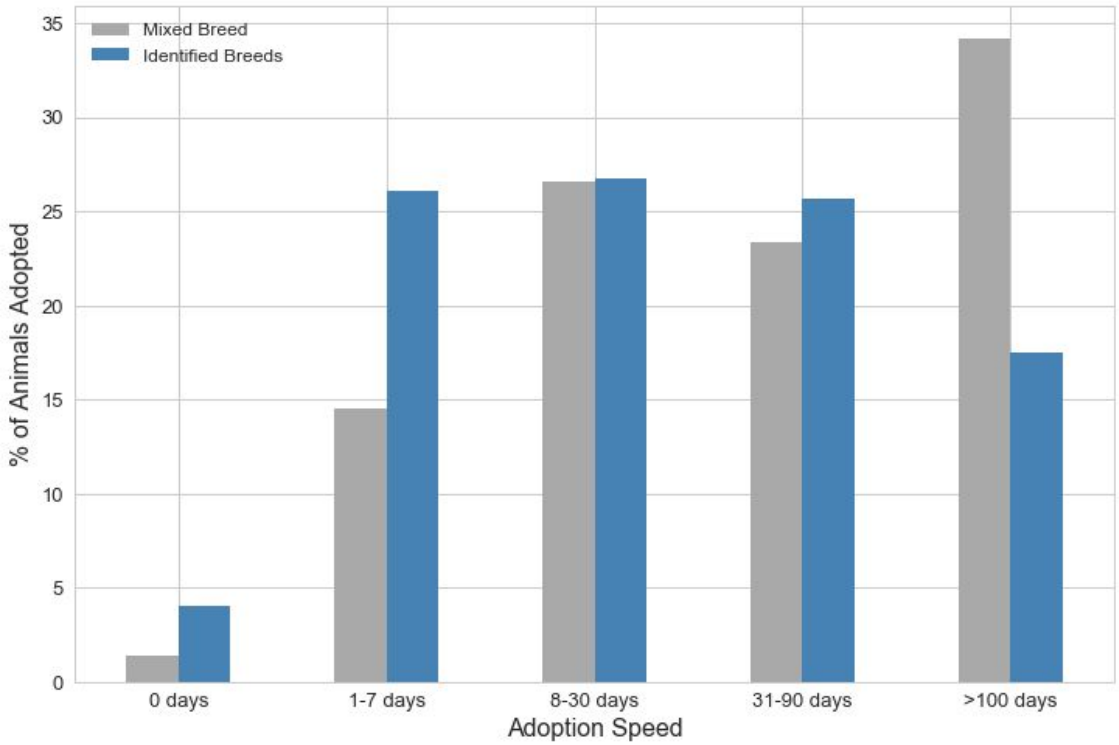

Figure 7. Percentage of Animals Adopted according to Adoption Speed and Breed

The data shown above seems to point to significant differences in the adoption speeds between mixed breed dogs and dogs with identified breeds. Performing Pearson's chi-squared test for independence we can verify that this difference is actually statistically significant (p-value < .001)

Dogs labeled as simply 'mixed breed' have lower rates of adoption in categories 0-3 (adoption within 90 days) and are approximately **twice as likely to end up in category 4** (no adoption after 100 days).

Performing a chi-square test for each breed to identify the ones with statistically significant different rates of no adoption rates we get the plot below, with the Pekingese and Silky Terrier having the lowest rates of no adoption at the top. Dogs labeled as “mixed breed” perform the worst:

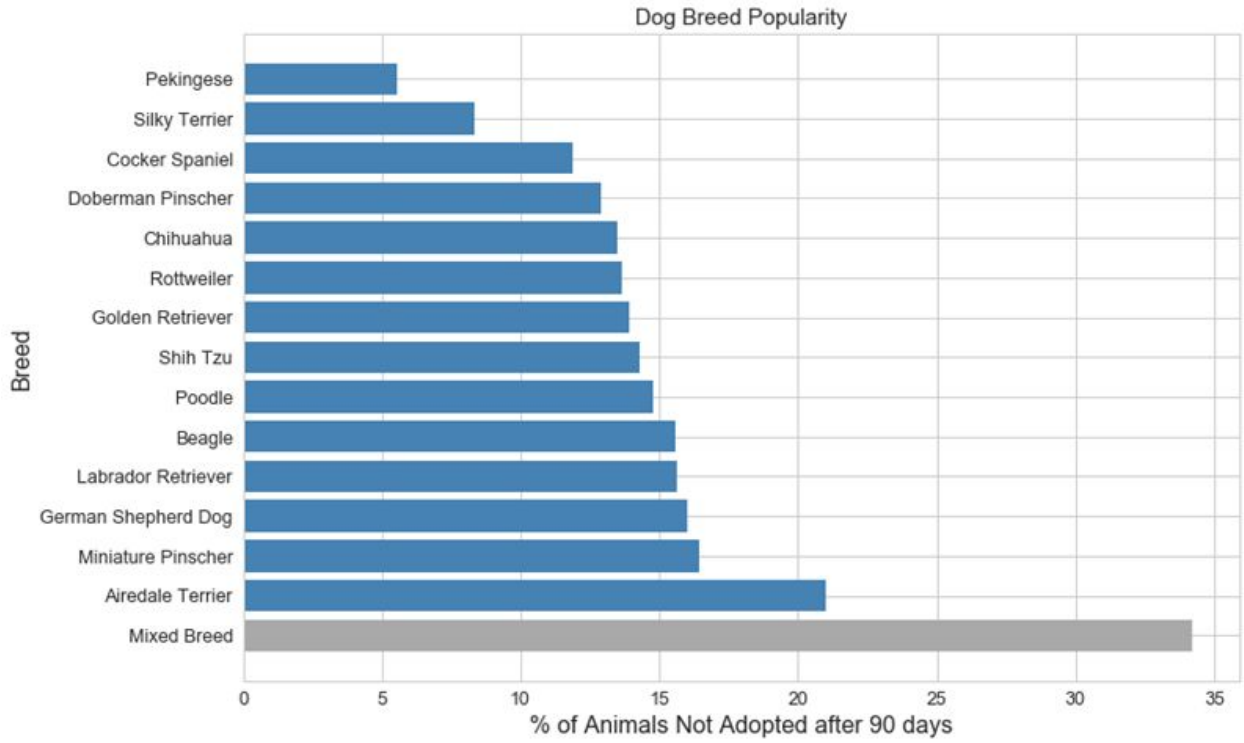

Figure 8. Dog Breeds vs. rates of no adoption

It is worth noting that a large proportion of animals with “identified breeds" in fact had values for both Breed1 and Breed2 variables and so are not pure-bred dogs, but technically mixed breed animals with the only difference being that their mixed breed composition had been identified/labeled. This points to the fact that the public has a negative bias not towards mixed breeds but rather towards dogs that don’t have a specific breed identification. For several of the dogs labeled as “mixed breed” Google’s vision API was able to identify the dog’s breed

composition with a good degree of confidence simply using the images provided in the pet profile. Going forward, this API could be used to assist rescue workers and volunteers in providing specific breed assignments and boost the animal's chances of getting adopted.

Correlation of Variables

Using a combination of Cramer's V (for categorical-categorical relationships), point biserial (for numerical and binary categorical relationships), eta correlation(for numerical and multi-level categorical relationships) and pearson's r for (numerical-numerical relationships), we compute correlation factors for the various variables. A heatmap with the results is shown below.

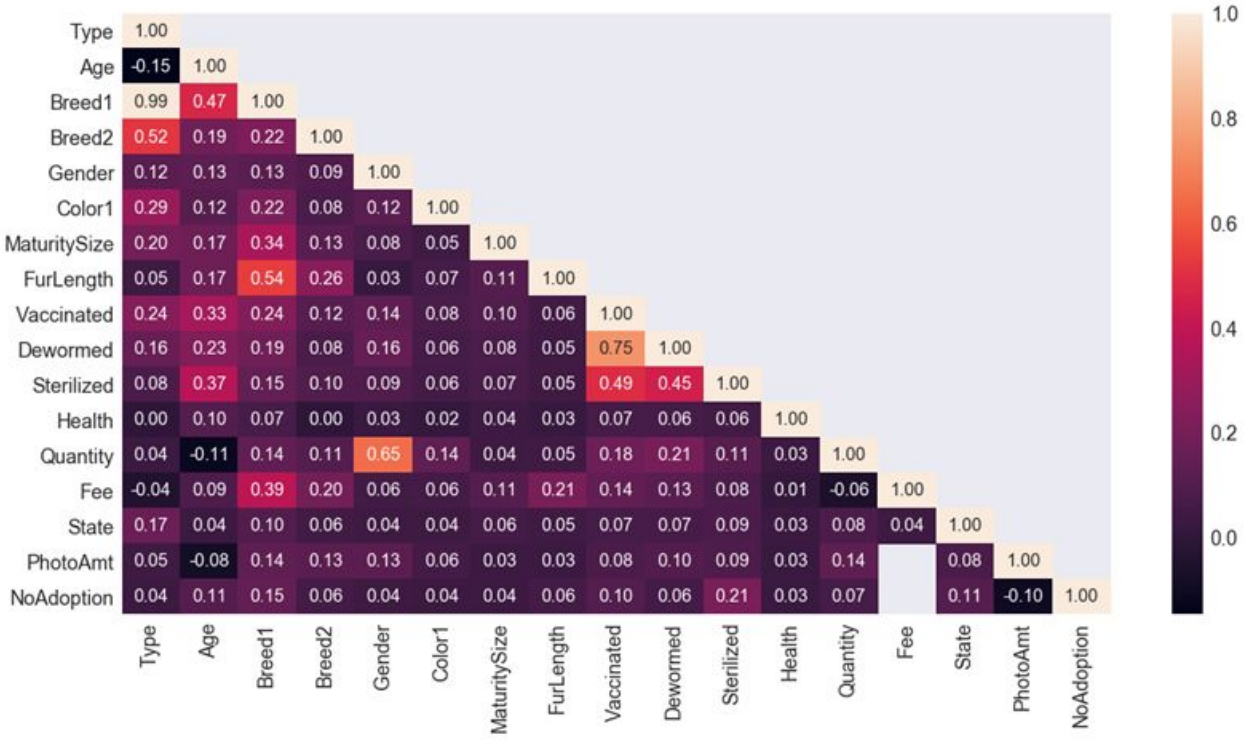

Figure 9. Heatmap of correlation results for categorical variables using Chi Squared

Vaccination, deworming and sterilization status are highly correlated amongst each other. Gender and quantity are also strongly correlated (this is expected as 'Mixed' gender was used

where there was a listing for multiple animals). Breed is strongly correlated to fee, animal type, FurLength, and MaturitySize - all of which were expected.

Surprisingly, breed and age are also strongly correlated. To better understand this finding a Mann-Whitney U Test was performed to identify for which breeds the age distribution was statistically different from other breeds.

Age Mean (Years)	Breed
0.52	Domestic Medium Hair
0.56	Domestic Short Hair
0.63	Mixed Breed
0.75	Siamese
0.82	Domestic Long Hair
1.01	British Shorthair
1.16	Maine Coon
1.22	Persian
1.25	Terrier
1.47	Doberman Pinscher
1.52	Siberian Husky
1.52	Labrador Retriever
1.56	Pit Bull Terrier
1.59	Dalmatian
1.78	German Shepherd Dog
1.86	Spitz
2.00	Dachshund
2.19	Rottweiler
2.21	Husky
2.28	Jack Russell Terrier
2.35	Miniature Pinscher
2.65	Chihuahua
2.72	Poodle
2.88	Pug
2.91	Pomeranian

2.93	Beagle
2.94	Golden Retriever
3.31	Schnauzer
3.66	Silky Terrier
3.66	Shih Tzu
3.76	Cocker Spaniel

Generic dog and cat breeds like ‘Domestic Medium Hair’, ‘Domestic Short Hair’ and ‘Mixed Breed’ and the Siamese tend to be younger than one year. It is harder to label kittens and puppies with specific breeds than it is adult animals, so this likely plays a role in the differences in the distribution. Several breeds of dog tend to be older, this might be due to certain breed temperaments or size changing overtime or people abandoning or losing pets because they are older.

The highest correlation factor for our target variable (NoAdoption) is whether or not the animal is sterilized at a correlation of .21. After that, breed, state and vaccination status come next. While vaccination, and sterilization status all might be valid predictors of how quickly a pet gets adopted they are also closely correlated to the animal’s age with correlation factors of -0.14 and -0.19 respectively. Puppies and kittens need to reach a certain age before these procedures can take place, so these three parameters cannot be considered in isolation, and a valid assessment would only be possible when considered in conjunction with age - as age very well could be a determining confounding factor.

Dog Breed Traits

The section below consists of the exploration of the correlation between NoAdoption and dog breed traits extracted from dogtime.com. While the traits are categorical in nature, they are provided based on a 1-to-5 star rating system, allowing us to treat them as numerical. Only those

dogs that had Breed Label 1 that matched breeds available at “dogtime.com” were used, the sample is still large enough (>2,000 entries) to allow us to derive meaningful insights.

Using points biserial correlation the heatmap below can help us determine which traits (if any) are correlated with adoption speed. A p-value cutoff of 0.05 was used (relationships that weren’t statistically significant per this criteria are not shown).

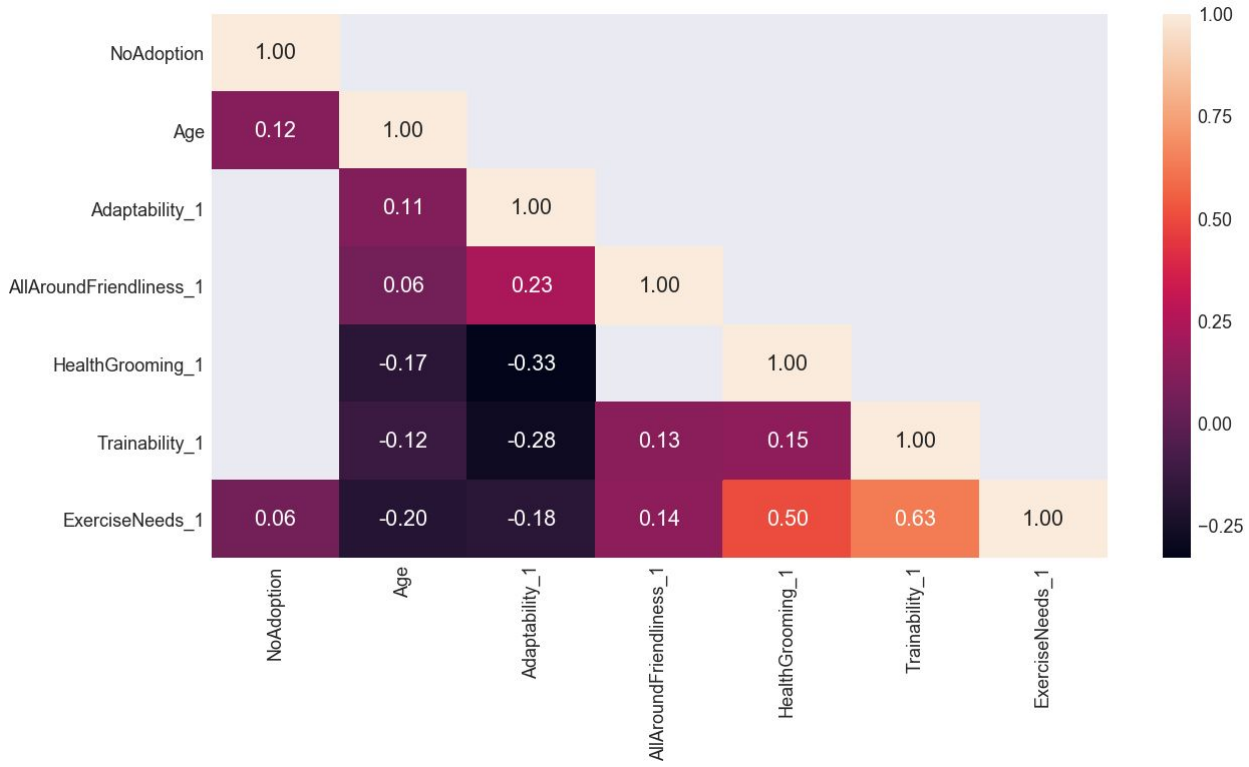

Figure 10. Heatmap of correlation results for adoption speed and breed traits using Points Biserial

The only breed trait found to have a correlation (weak) to adoption rates was ExerciseNeeds, the correlation is weak and conclusive results cannot be drawn from this. Among the breed traits, ExerciseNeeds is shown to be highly correlated to Trainability as well as Health/Grooming.

Machine Learning

The machine learning model aims to take new animal profiles and predict whether or not they are at risk of not getting adopted after 100 days.

Feature Engineering and Selection

A t-test for each breed was performed to see if we could limit the number of breeds to those that were statistically significantly different from others. An additional column was added labeled 'BreedNoAdoptionRate' that contained the average no adoption rates for breeds found to be statistically significantly different from others, for all other breeds the column contained the overall average no adoption rate.

After encoding categorical features, normalizing and fitting the data with the statsmodel logistic regression model, the resulting p-values showed that MaturitySize, and VideoAmount were not significant at the 5% level. A feature set was created based on the 5% threshold.

Checking for multicollinearity between variables using the Variance Inflation Factor (VIF) showed the highest factors for Vaccination and Dewormed status but at 2.0 and 1.7 respectively their colinearity doesn't justify removing either of them from the analysis.

Using the logistic regression model parameters the odds ratios for each variable were also computed, their results are shown below:

Odds Ratios	
Type	0.881
Gender	1.044
MaturitySize	1.020
FurLength	0.957
Vaccinated	1.117
Dewormed	0.921
Sterilized	1.388

Health	1.063
Quantity	1.209
Fee	1.076
VideoAmt	1.007
PhotoAmt	0.811
Breed1NoAdoptionRate	1.526
Breed2NoAdoptionRate	1.073
Age_Years	1.241
Color_Brown	1.012
Color_Cream	0.958
Color_Golden	0.989
Color_Gray	0.972
Color_White	0.963
Color_Yellow	1.021
State_Kedah	1.056
State_Kelantan	1.023
State_Kuala Lumpur	1.149
State_Labuan	1.014
State_Melaka	1.159
State_Negeri Sembilan	1.093
State_Pahang	1.024
State_Perak	1.057
State_Pulau Pinang	1.111
State_Sabah	1.014
State_Sarawak	1.081
State_Selangor	1.046
State_Terengganu	1.027

The highest odds ratio values are for the Breed Adoption rates, reiterating that breed is one of the most significant factors in predicting adoption speed. A second logistic regression model (non normalized) was created to compute interpretable Odds Ratio coefficients. It was noted that each year of age is associated with an average 13.4% higher likelihood of not being adopted.

Model Selection and Optimization

Three models were compared: Logistic Regression, Random Forest and Gradient Boosting, and the resulting ROC-AUC, F measure, and Brier score were used to select the best solution.

The feature selection for the Logistic Regression model was generated by first applying L1 (LASSO) regularization to the model using all features that had been found significant at the 10% level (using p-value results). From the results, features with non-zero coefficients were extracted and using this new set of features, another model was created using L2 regularization and the c value optimized with GridSearchCV. The optimal c value found was 10.

The Random Forest was optimized using GridSearchCV using all features that had been found significant at the 10% level, 5% level and again using the same set of features extracted from the L1 Logistic Regression Model. The higher-performing model (10% level) was used to assess feature importance using the built-in feature importance attribute from scikit-learn:

The Gradient Boosting model was optimized with RandomSearchCV using the same feature set as the optimal random forest model. ROC-AUC values for all three models are shown in the plot below:

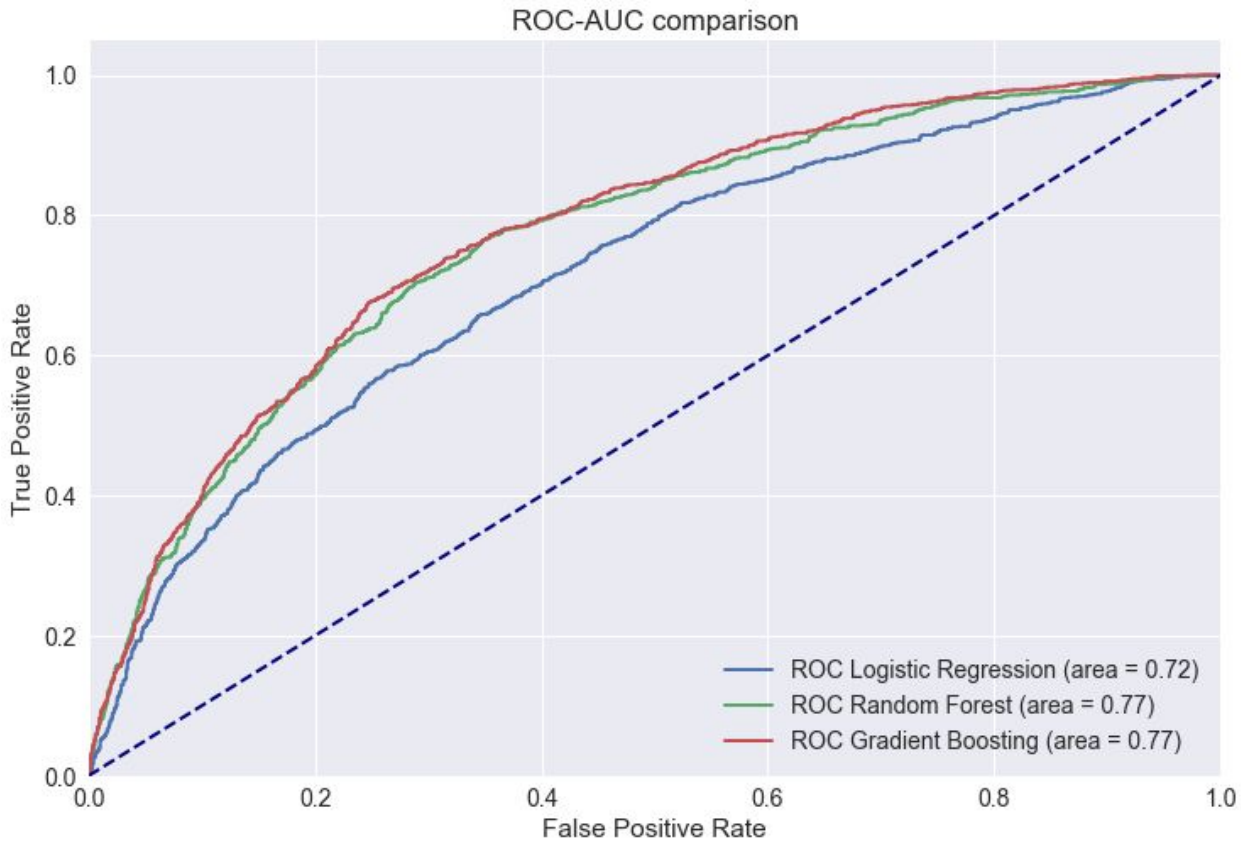

Figure 11. ROC-AUC Comparison for the various models

In addition to the ROC-AUC, the Brier score was also used for the situation in which model probabilities might need to be used. Using probabilities, rescues with a limited amount of space can use the results to get a rank of the animals with the highest risk. The model with the highest ROC-AUC and lowest Brier score (0.17) was the Gradient Boosting Model. If different models had proven to be optimal for the two metrics then different models could be used depending on the desired output.

Choosing A Metric

The purpose of the model is to try to identify animals that have a higher likelihood of not getting adopted. If a Type I error occurs (false positive) and an animal mistakenly gets labeled as "not likely to be adopted" and gets additional exposure to potential adopters as a result, there is no downside to that animal. However, because rescue organizations' resources are limited this additional attention or spot at a foster home might take away from an animal that is truly at risk. If a Type II error occurs (false negative) an animal that is likely to not get adopted might not get additional attention and could potentially end up getting euthanized.

Because we want to put emphasis on minimizing the number of Type II errors the original optimal threshold was selected calculating the F measure with a beta value of 2 but the resulting best threshold identified over 40% of animals as being at risk of not getting adopted, which is not feasible considering rescue organization resources and space for animal intake are limited. Assuming we can prioritize approximately 25% of animals, the F0.5 measure was used instead, and the resulting optimal threshold was 0.38 which identified 24.7% of all animals as being at risk.

Thresholding

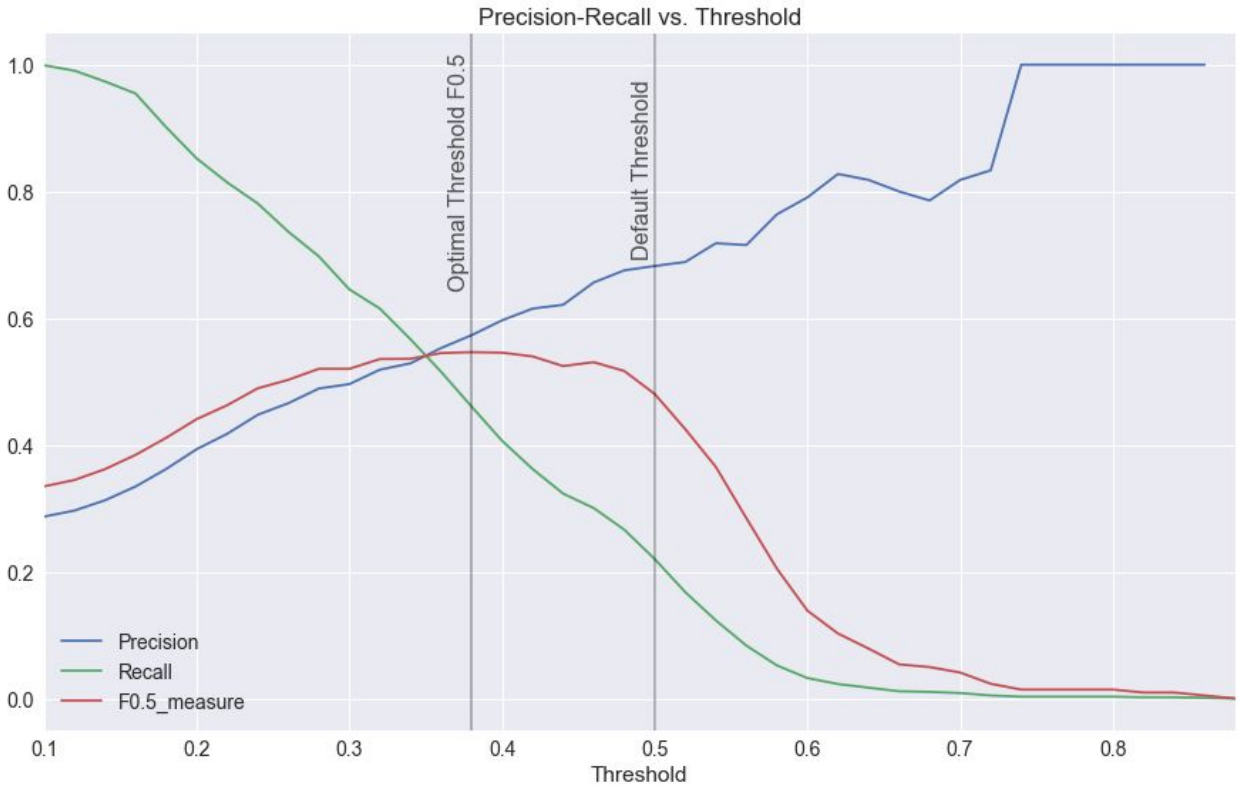

Figure 12. Precision, recall and F0.5 measure values as a function of model threshold

The confusion and classification matrices for the model can be seen below:

	Predicted 0 (Adoption)	Predicted 1 (No Adoption)
Actual 0 (Adoption)	2283	411
Actual 1 (NoAdoption)	537	518

	precision	recall	f1-score	support
FALSE	0.81	0.85	0.83	2694
TRUE	0.56	0.49	0.52	1055
avg/total	0.74	0.75	0.74	3749

While selecting the appropriate threshold is important for binary classification output, the raw ranked probability value for each prediction would be a more useful asset to animal rescue organizations with limited amount of space and resources. Using the raw probability output, the Random Forest model has a good Brier score of 0.17 showing that it can provide helpful guidance when making decisions.

Conclusion

Using the model and the results provided in this report rescue organizations can hopefully target their adoption efforts on animals that are at a higher risk of not getting adopted and increase the number of overall adoptions. One straightforward way to improve a dog's chances of being adopted would be to simply improve the labeling of the animals' breeds on their online profiles by specifying specific breeds rather than simply "mixed breed". Google's Vision API can detect a variety of breeds with a good degree of confidence using the animal's images.

For future work, Google API Vision tags and image analysis can be conducted to explore how other parameters might also impact adoption such as the animal's cuteness or the quality of their profile images. Natural language processing can also be explored on the pet's online profile descriptions.

This project also focused on extracting insights relating to dog breeds, it would be interesting to explore whether cat breeds and their traits enact a significant effect on their adoption rates.

References

1) ASPCA. 2020. Pet Statistics. [online] Available at: <https://www.aspca.org/animal-homelessness/shelter-intake-and-surrender/pet-statistics> [Accessed 1 May 2020].
2) Nytimes.com. 2020. Why Euthanasia Rates At Animal Shelters Have Plummeted. [online] Available at: <https://www.nytimes.com/2019/09/03/upshot/why-euthanasia-rates-at-animal-shelters-have-plummeted.html> [Accessed 1 May 2020].