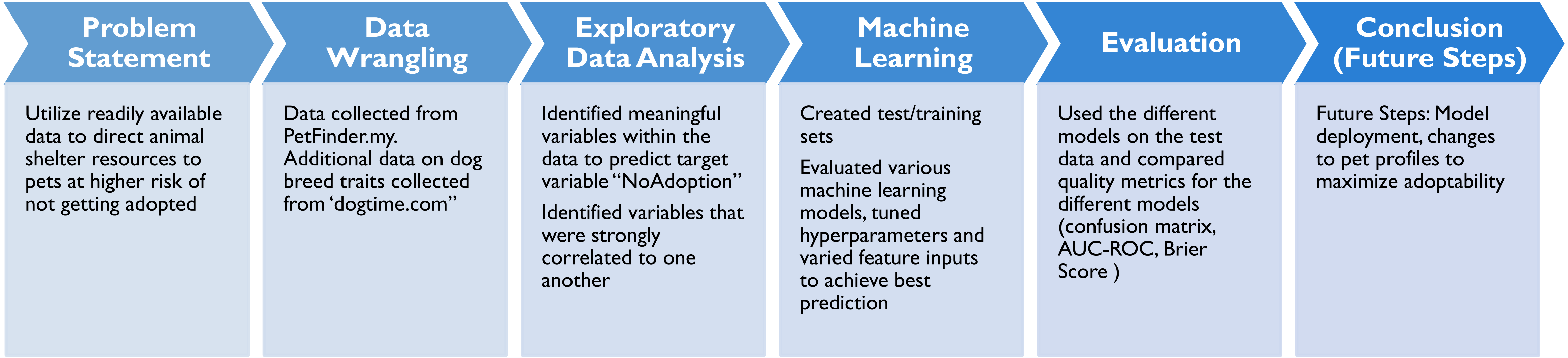
PET ADOPTION SPEED PREDICTION USING PETFINDER DATA

FOR ANIMAL RESCUE ORGANIZATIONS

I. THISTED FOR SPRINGBOARD CAPSTONE PROJECT

OUTLINE



Problem Statement

Utilize readily available data to direct animal shelter resources to pets at higher risk of not getting adopted

Data Wrangling

Data collected from PetFinder.my. Additional data on dog breed traits collected from 'dogtime.com"

Exploratory Data Analysis

Identified meaningful variables within the data to predict target variable "NoAdoption"

Identified variables that were strongly correlated to one another

Machine Learning

Created test/training sets

Evaluated various machine learning models, tuned hyperparameters and varied feature inputs to achieve best prediction

Evaluation

Used the different models on the test data and compared quality metrics for the different models (confusion matrix, AUC-ROC, Brier Score)

Conclusion (Future Steps)

Future Steps: Model deployment, changes to pet profiles to maximize adoptability

PROJECT OVERVIEW

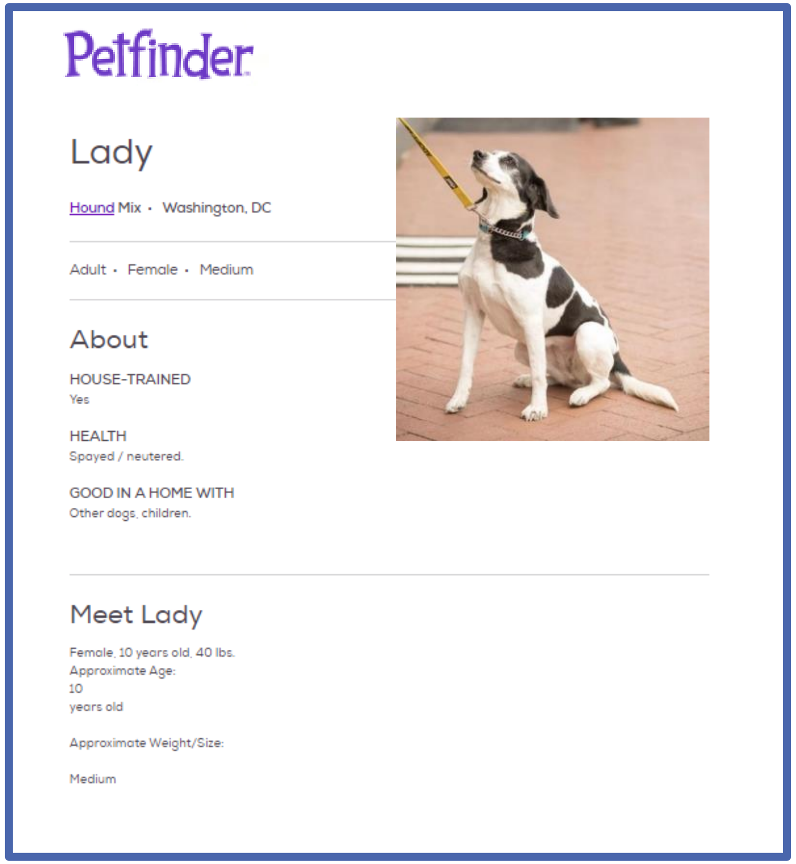
Approximately 6.5 million companion animals enter U.S. animal shelters nationwide every year[1]

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.[1,2]

The goal of this project is to both derive insights and create a machine learning model that organizations can use to help them in identifying animals with the highest risk of not getting adopted.

PROJECT OVERVIEW



Petfinder

Lady

Hound Mix · Washington, DC

Adult · Female · Medium

About

HOUSE-TRAINED
Yes

HEALTH
Spayed / neutered.

GOOD IN A HOME WITH
Other dogs, children.

Meet Lady

Female, 10 years old, 40 lbs.
Approximate Age:
10
years old

Approximate Weight/Size:

Medium

- Petfinder is an online, searchable database of animals who need homes.
- Organizations maintain their own home pages and available-pet databases.
- Typical pet profile contains picture, age, size, sex, health status, house-training status, location and description.
- Data from PetFinder Malaysia is available on Kaggle which also includes pet's adoption speeds.
- The data can be used to explore relationships between adoptability and the various parameters available from the pets' profiles.
- Insights can help organizations to focus resources on animals at higher risk of not being adopted and make informed decisions.

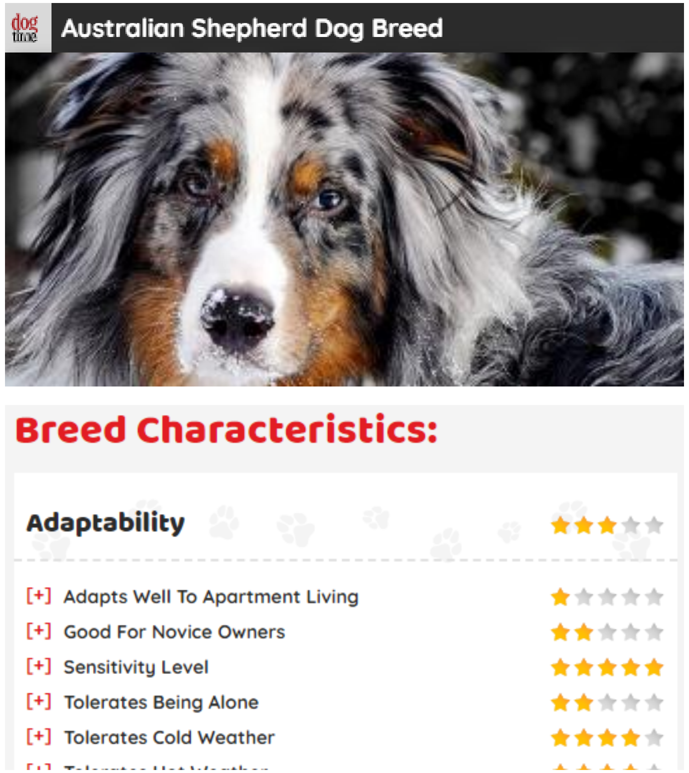
DATA WRANGLING

- Dataset from Petfinder Malaysia:
  - The dataset consists of metadata with 25 parameters as well as images, videos and descriptions for approximately 15,000 pets as well as their adoption speed (our target variable)
  - The data is a mixture of text (animal profile descriptions), tabular data (with 25 parameters) and pet images and videos
  - The pet images have been run through Google's Vision API, and the resulting metadata is available. Each pet profile's description was additionally run through Google's Natural Language API, providing analysis on sentiment and key entities This data was not used for this project
  - The tabular data will be used as the features for our machine learning models
  - The dataset was clean, and the only missing variables were found in the pet name and description columns
  - Checking the range and distribution of the tabular parameters revealed no obvious outliers

DATA WRANGLING

- Adoption speeds are broken up into the following categories:
  - 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)
- While it might be beneficial to predict the various categories of speed, for this project we exclusively focused on predicting those animals in Category 4 since those animals were not adopted
  - Category 4 (no adoption) is the target variable for the machine learning models

DATA WRANGLING



Australian Shepherd Dog Breed

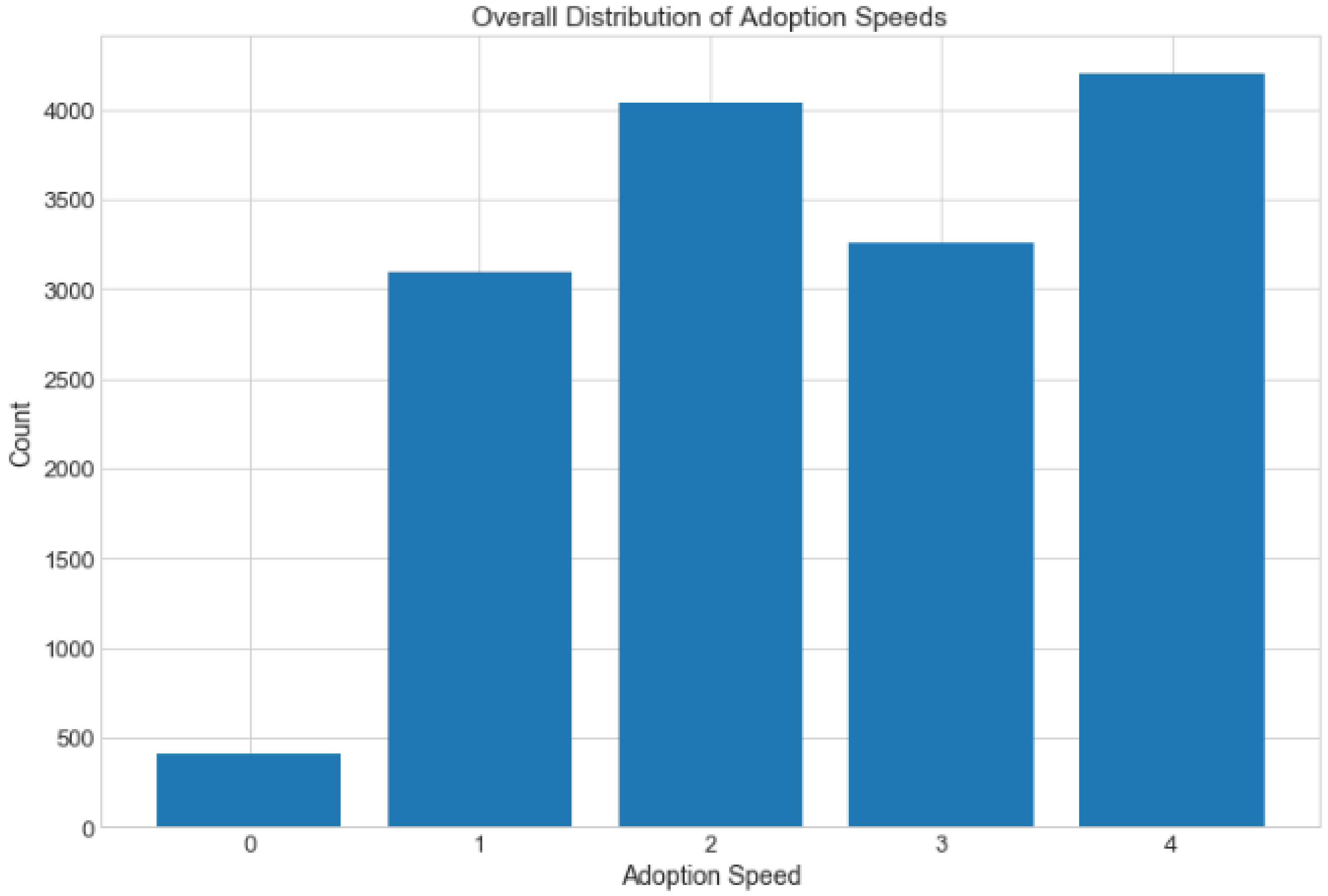
Breed Characteristics:

Adaptability	★★★☆☆
[+] Adapts Well To Apartment Living	★☆☆☆☆
[+] Good For Novice Owners	★★☆☆☆
[+] Sensitivity Level	★★★★★
[+] Tolerates Being Alone	★★☆☆☆
[+] Tolerates Cold Weather	★★★★☆

- Data was augmented with dog breed-specific characteristics extracted from the website "dogtime.com" to explore whether canine breed traits had any observable impact on adoption rates.
- Breed names from the website were manually corrected to match those on Kaggle
- Most dogs in the adoption database are 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 resulting sample size was still large enough (> 2000 entries) and sufficient to derive meaningful insights

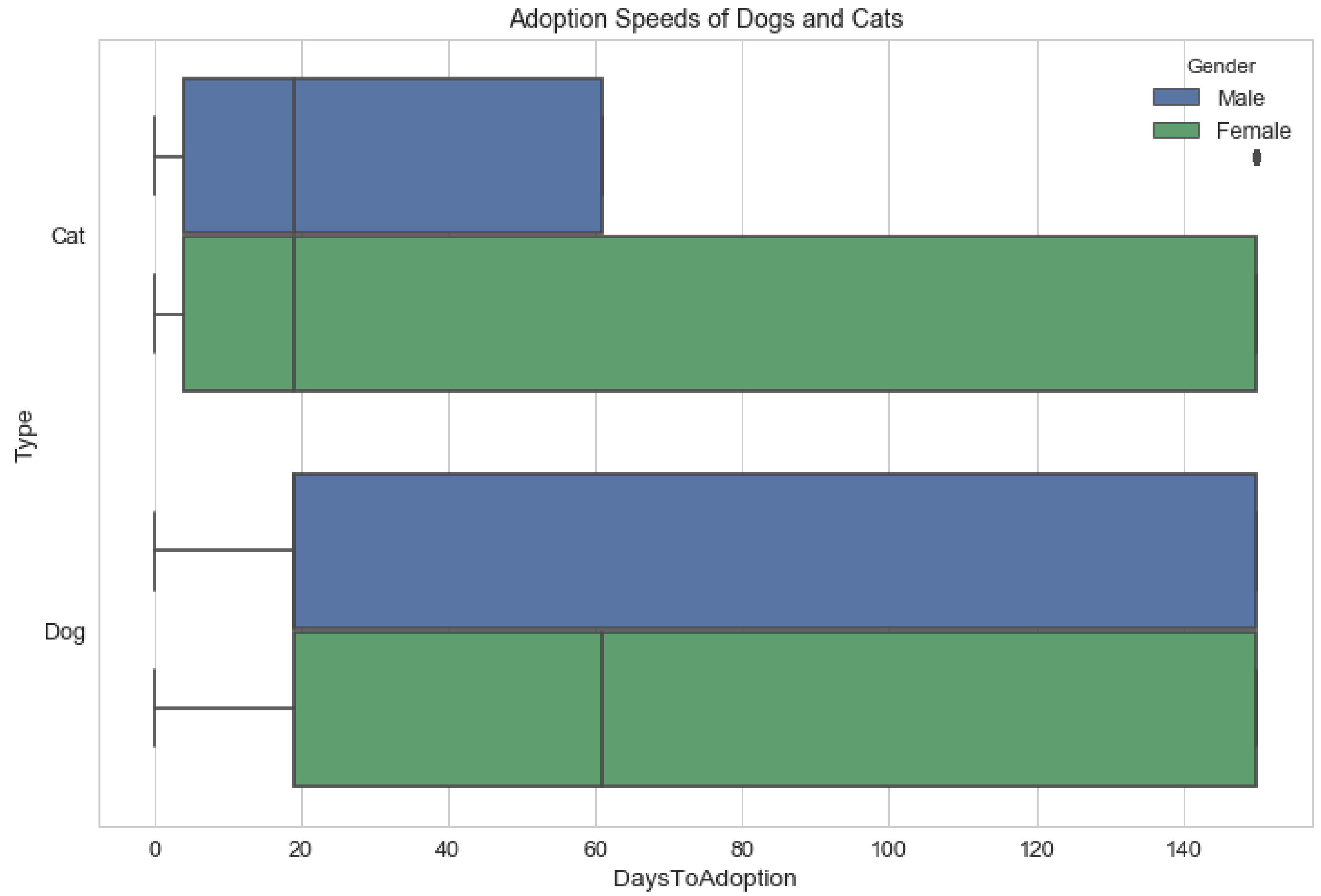
EXPLORATORY DATA ANALYSIS

- Only a small proportion of animals gets adopted in the fastest category (same-day adoption) and while many 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).
- Over the next plots we explore relationships between the adoption speed and the various available features



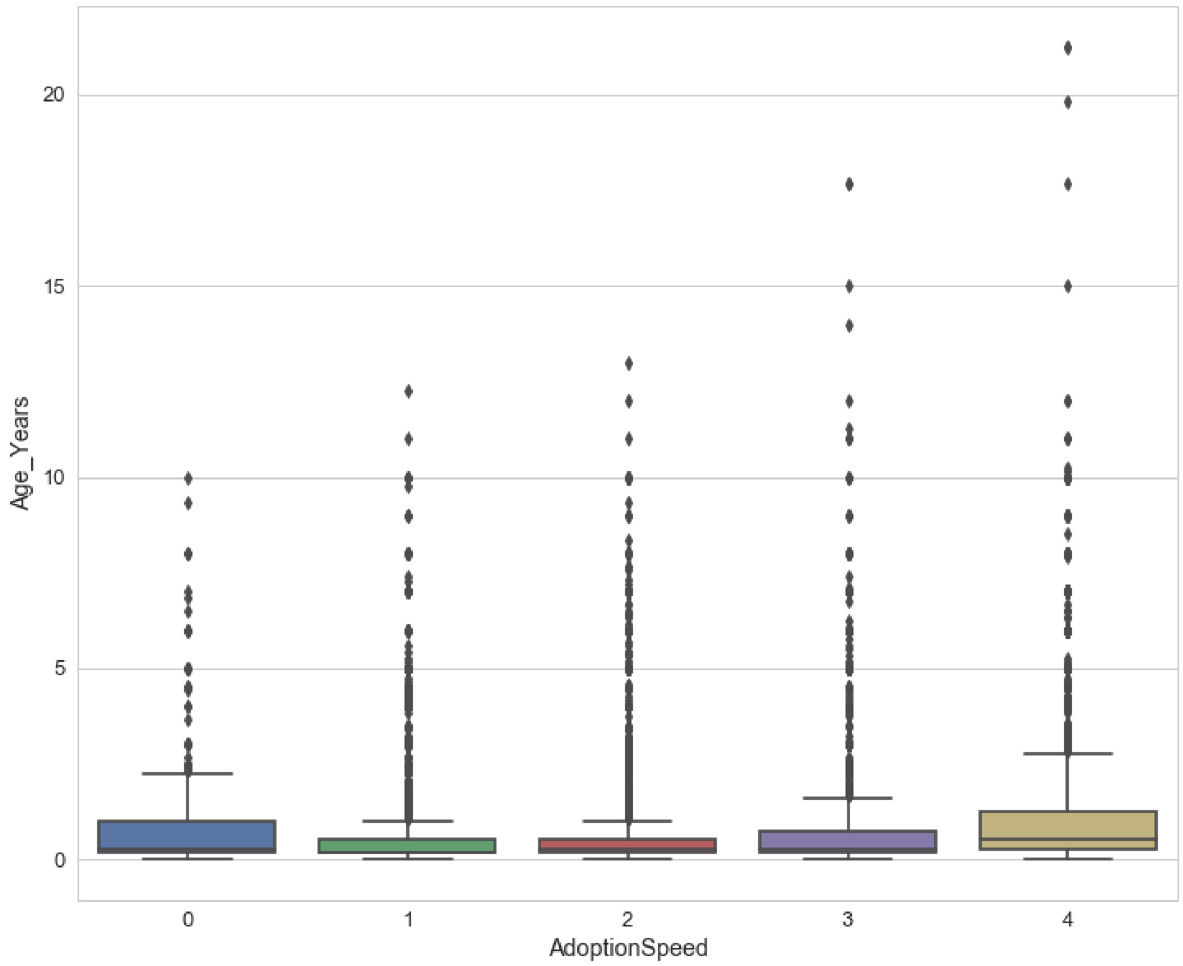
TYPE AND GENDER

- Cats seem to have an easier time getting adopted
- 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
- Gender seems to play as much of an impact as species in the adoption outcomes
- The median adoption speed of females is 3 times as large (60 days) as the median adoption speed for males (20 days).
- Gender seems to have a larger impact on dog adoptions as compared to cats
- T-test revealed both animal type and gender to be statistically significant



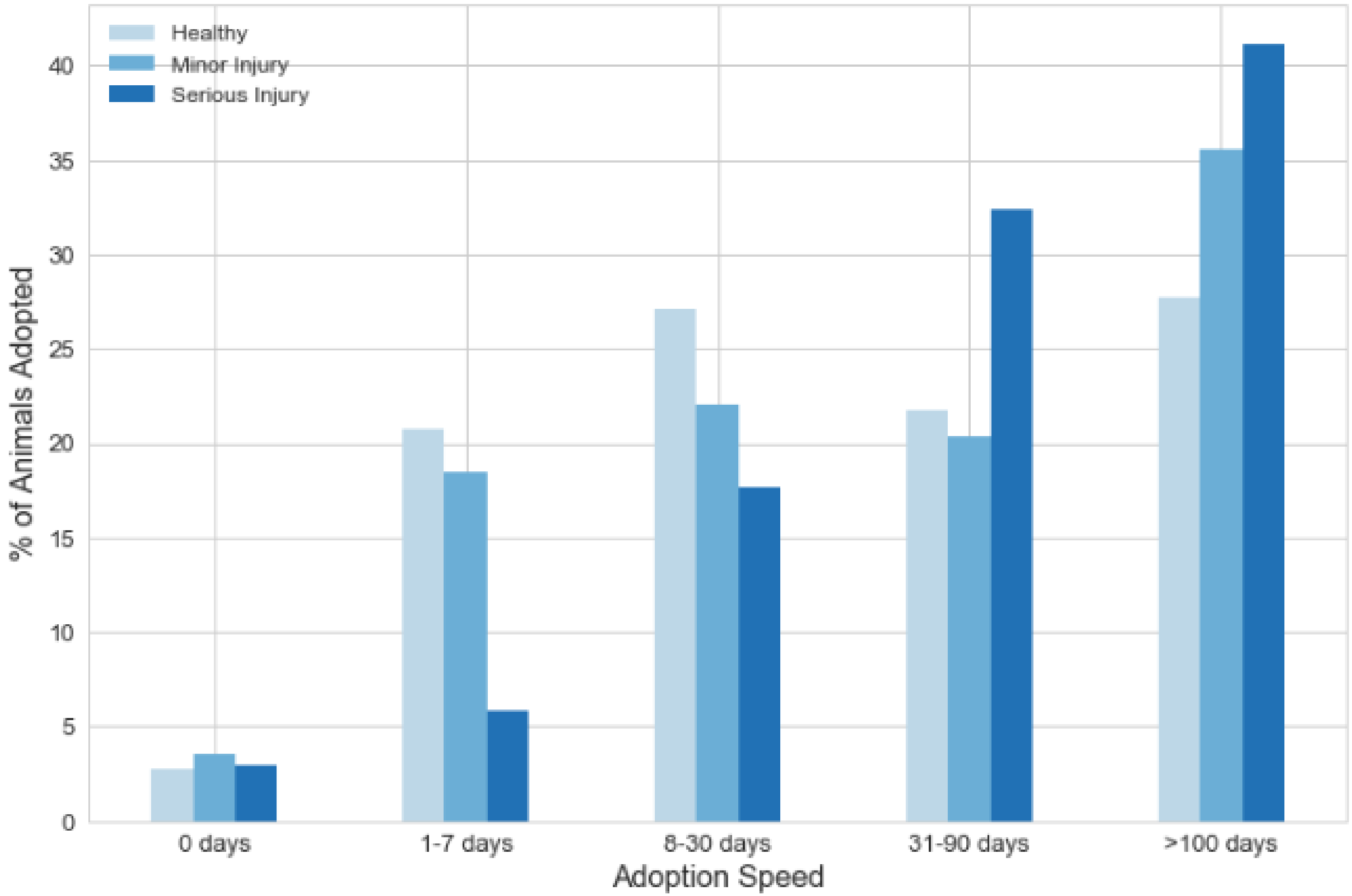
AGE AND ADOPTION

- 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.
- Using a logistic regression model Age has a p-value smaller than 0.0001



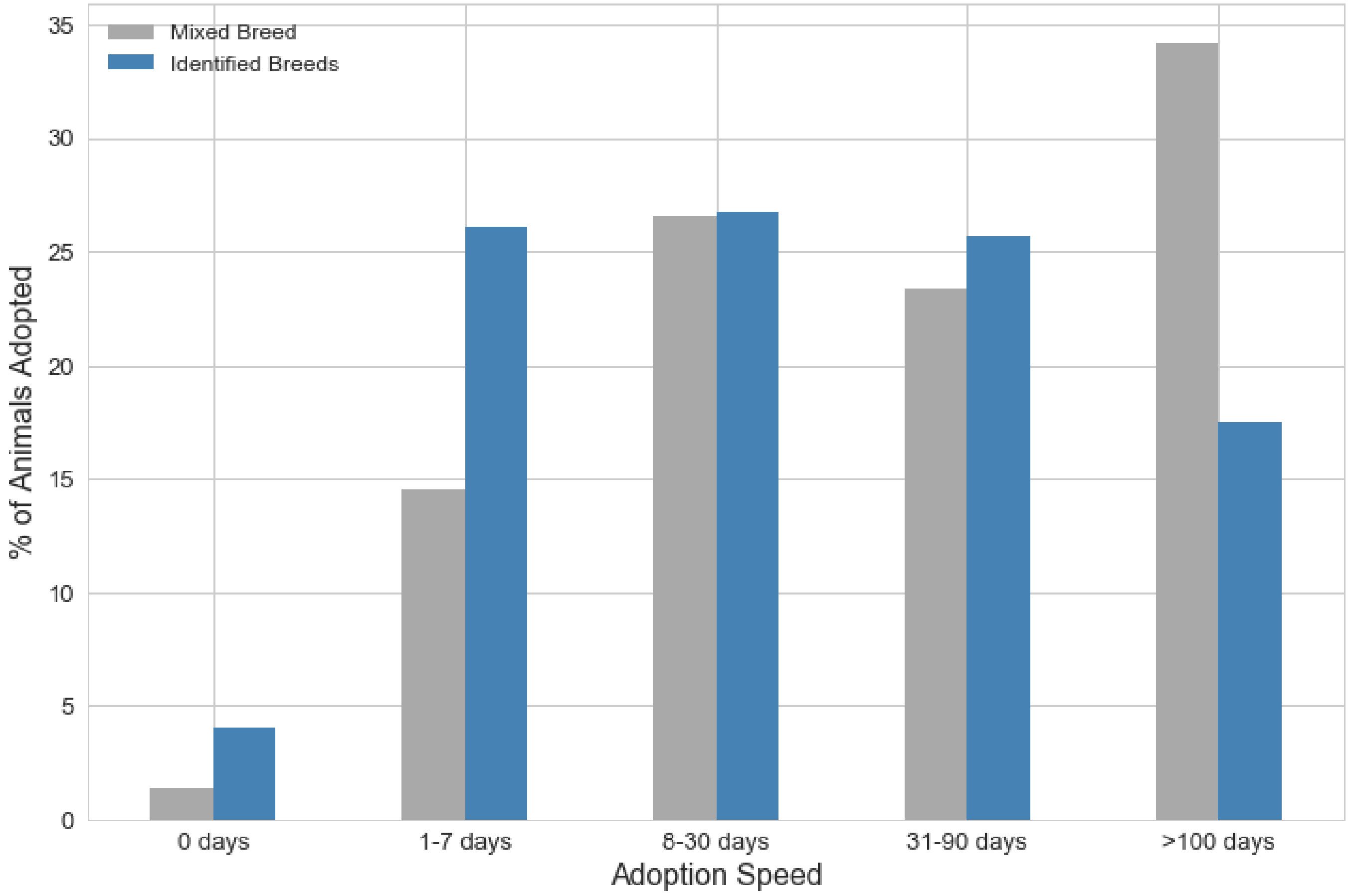
ANIMAL HEALTH

- 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.
- Using Tukey's HSD (honestly significant difference) test, we conclude that while there is a statistically significant difference between healthy animals and the remaining groups, the extent of the injury does not have a clear impact.



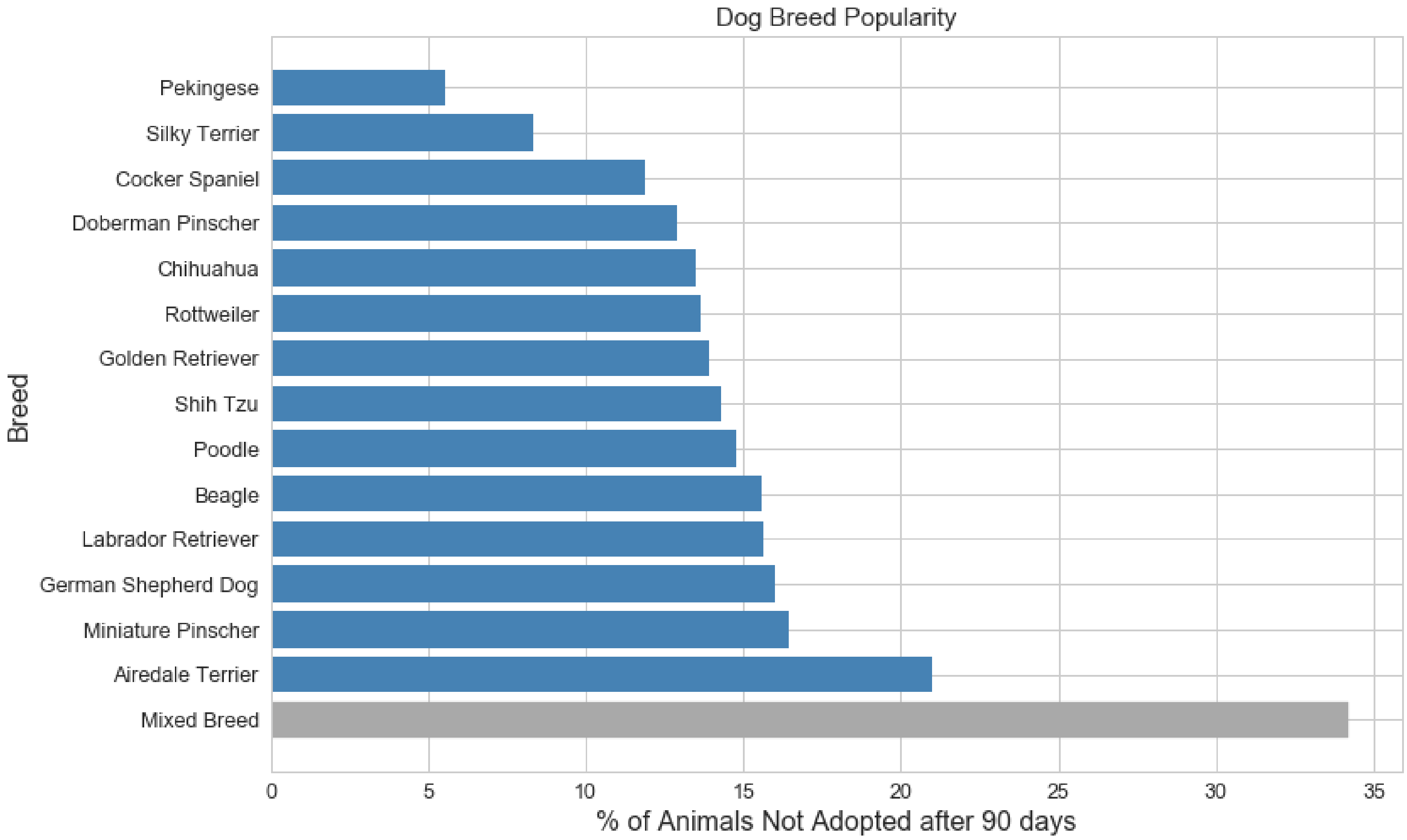
DOG BREEDS

- There are 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 statistically significant
- Dogs labeled as simply 'mixed breed' have lower rates of adoption in categories 0-3 (adoption within 90 days) and have rates **of no adoption nearly double that of other breeds**



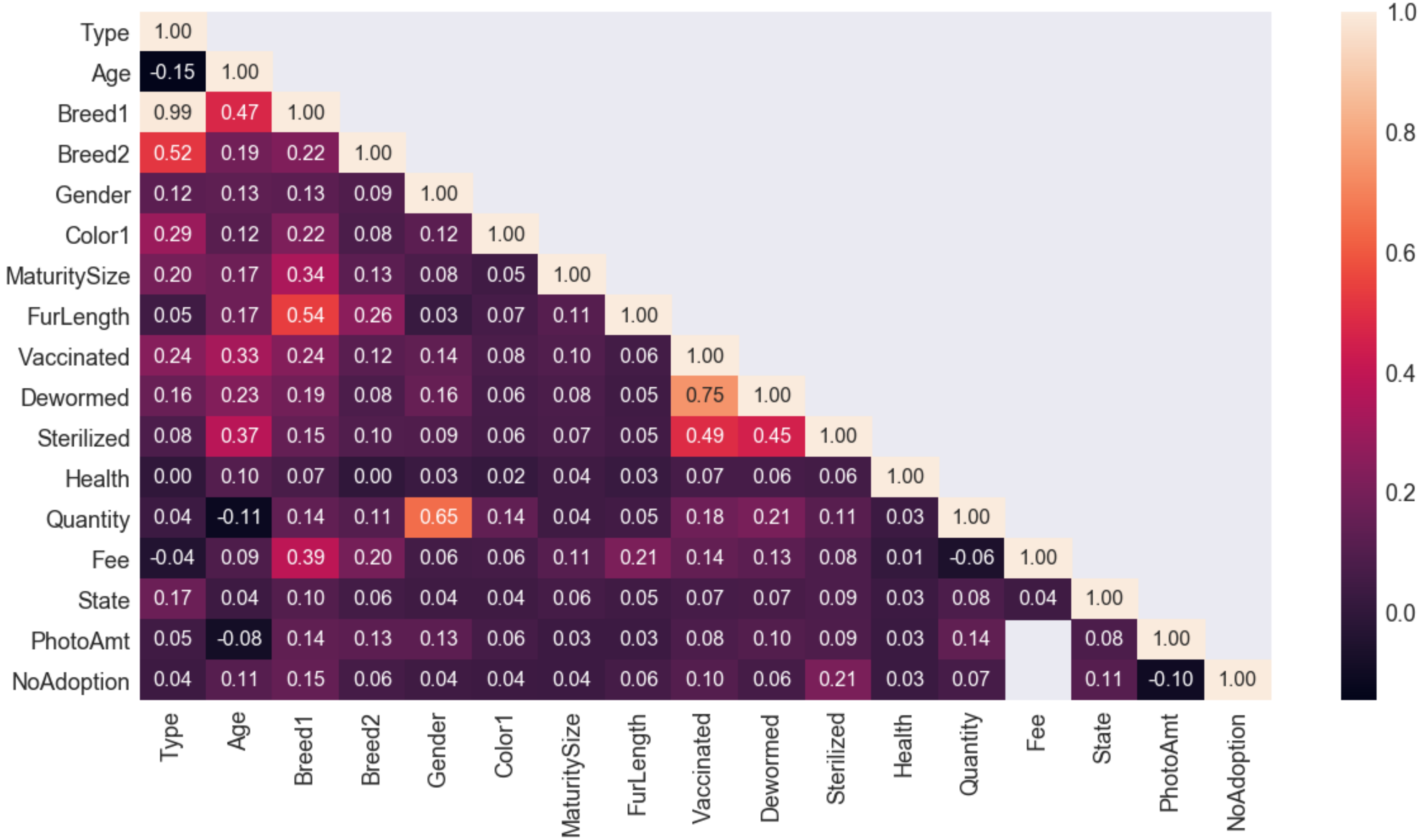
DOG BREEDS

- Among breeds found to be statistically significant different from others the Pekingese and Silky Terrier are the most popular in Malaysia with low rates of no adoption at 5% and 8%
- The worst-performing breed comes from dogs identified as "mixed breed"



CORRELATION OF VARIABLES

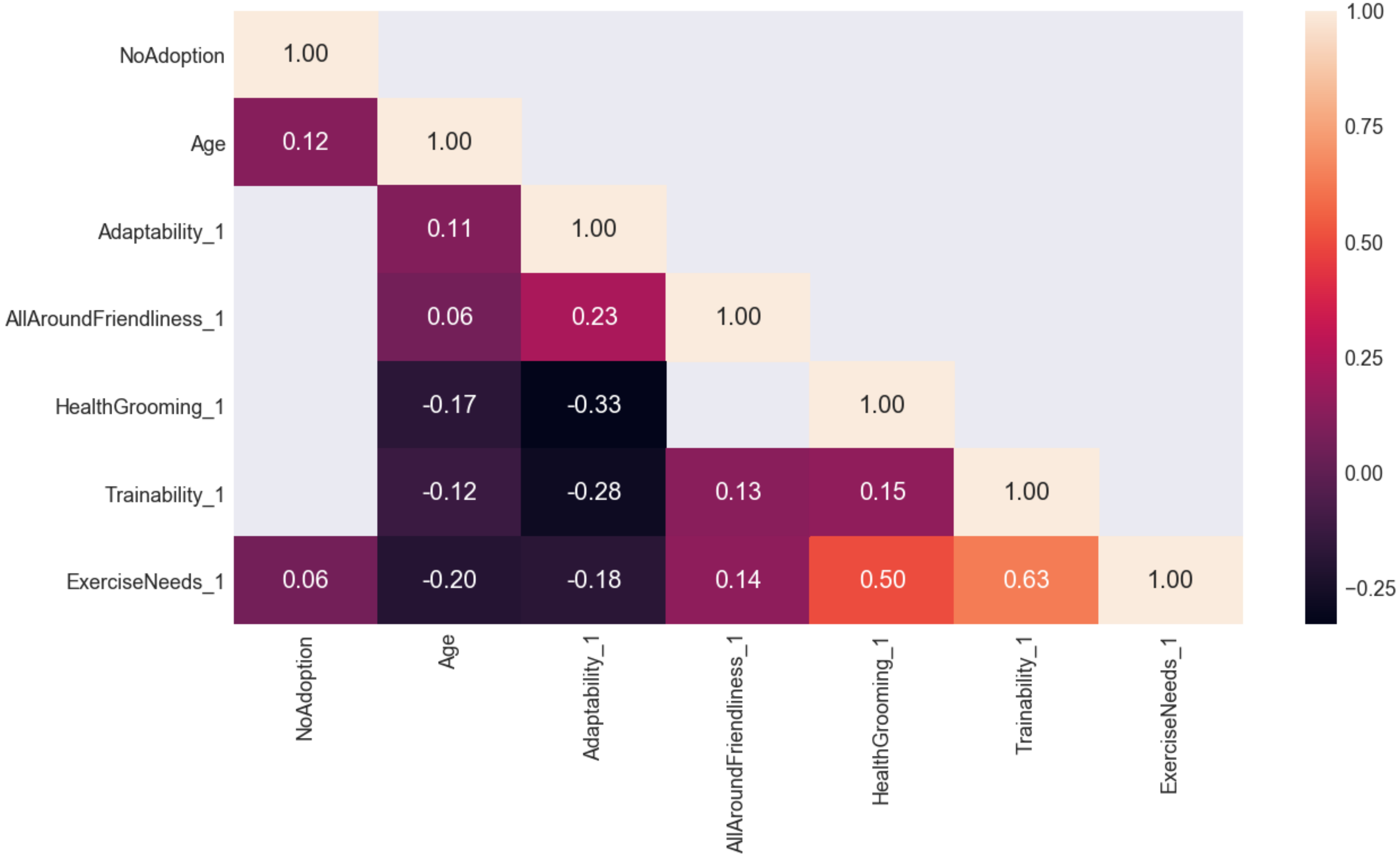
- Using a combination of Cramer's V, point biserial, eta correlation and Pearson's r, correlation factors for the various variables are shown
- Features highly correlated amongst each other are:
  - Breed and Type, Fee, Maturity Size and Fur Length
  - Quantity and gender
  - Vaccination, Sterilization and deworming status
  - Age and breed
- The highest correlation factor for our target variable (NoAdoption) is whether the animal is sterilized at a correlation factor of .21



	Type	Age	Breed1	Breed2	Gender	Color1	MaturitySize	FurLength	Vaccinated	Dewormed	Sterilized	Health	Quantity	Fee	State	PhotoAmt	NoAdoption
Type	1.00																
Age	-0.15	1.00															
Breed1	0.99	0.47	1.00														
Breed2	0.52	0.19	0.22	1.00													
Gender	0.12	0.13	0.13	0.09	1.00												
Color1	0.29	0.12	0.22	0.08	0.12	1.00											
MaturitySize	0.20	0.17	0.34	0.13	0.08	0.05	1.00										
FurLength	0.05	0.17	0.54	0.26	0.03	0.07	0.11	1.00									
Vaccinated	0.24	0.33	0.24	0.12	0.14	0.08	0.10	0.06	1.00								
Dewormed	0.16	0.23	0.19	0.08	0.16	0.06	0.08	0.05	0.75	1.00							
Sterilized	0.08	0.37	0.15	0.10	0.09	0.06	0.07	0.05	0.49	0.45	1.00						
Health	0.00	0.10	0.07	0.00	0.03	0.02	0.04	0.03	0.07	0.06	0.06	1.00					
Quantity	0.04	-0.11	0.14	0.11	0.65	0.14	0.04	0.05	0.18	0.21	0.11	0.03	1.00				
Fee	-0.04	0.09	0.39	0.20	0.06	0.06	0.11	0.21	0.14	0.13	0.08	0.01	-0.06	1.00			
State	0.17	0.04	0.10	0.06	0.04	0.04	0.06	0.05	0.07	0.07	0.09	0.03	0.08	0.04	1.00		
PhotoAmt	0.05	-0.08	0.14	0.13	0.13	0.06	0.03	0.03	0.08	0.10	0.09	0.03	0.14		0.08	1.00	
NoAdoption	0.04	0.11	0.15	0.06	0.04	0.04	0.04	0.06	0.10	0.06	0.21	0.03	0.07		0.11	-0.10	1.00

CORRELATION OF VARIABLES – DOG BREED TRAITS

- 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.
- ExerciseNeeds is shown to be highly correlated to Trainability as well as Health/Grooming.



MACHINE LEARNING

- Machine Learning Models built:
  - Logistic Regression
  - Random Forest
  - Gradient Boosting
- Generated new features from existing ones to improve model performance
- Varied the input feature sets and tuned each model’s hyperparameters
- Quality Metrics: confusion matrix, ROC-AUC, Brier Score

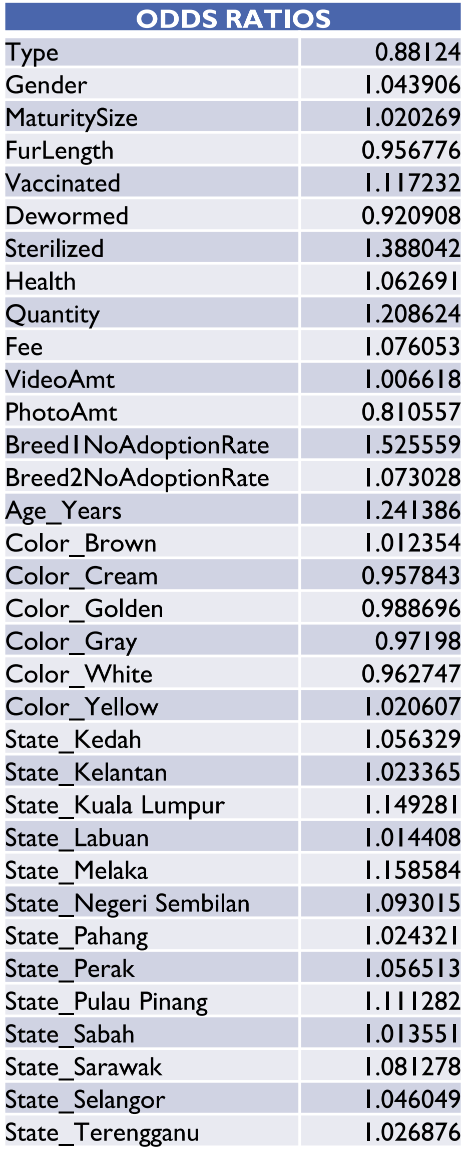
FEATURE ENGINEERING AND SELECTION

- Removing features that exhibit multicollinearity
  - Features were normalized using StandardScaler then their Variance Inflation Factor computed
  - No features exhibited strong collinearity (largest VIF were 2.0 and 1.7)

- BreedNoAdoptionRate
  - Chi-square test (stats model proportions_chisquare) 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 of no adoption rate.
- All models were tested with 3 feature sets:
  - All features
  - Features with non-zero coefficients extracted from logistic regression model with L1 penalization and optimized C value with GridSearchCV
  - Feature sets extracted from statsmodel logisitic regression results (p-value at 5% and 10% significance thresholds)

ODDS RATIO

- Features were normalized and fitted to the data with a statsmodel logistic regression model
- Categorical features were transformed (with one-hot-encoding or changed to numerical)
- Resulting p-values showed that MaturitySize, and VideoAmount were not significant at the 10% level
- Odds Ratio were calculated for the normalized model to assess feature importance.
- 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.

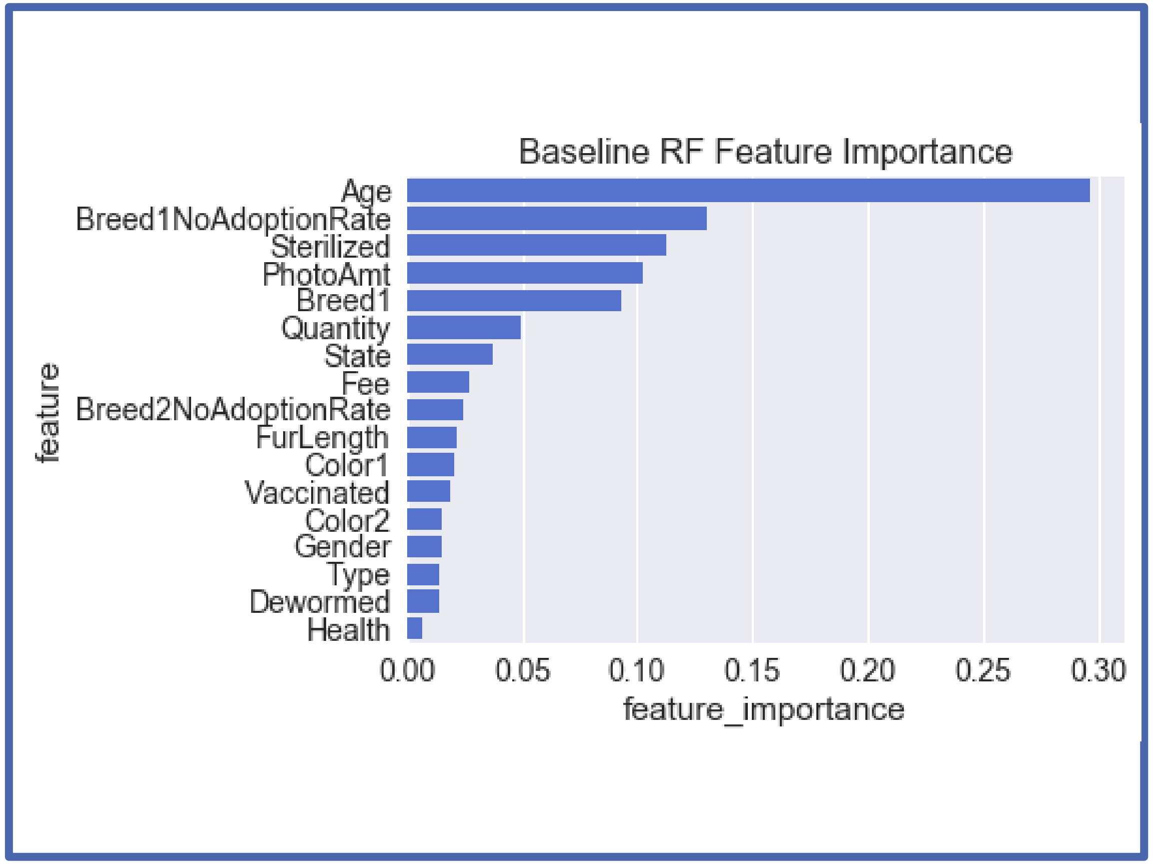


ODDS RATIOS	
Type	0.88124
Gender	1.043906
MaturitySize	1.020269
FurLength	0.956776
Vaccinated	1.117232
Dewormed	0.920908
Sterilized	1.388042
Health	1.062691
Quantity	1.208624
Fee	1.076053
VideoAmt	1.006618
PhotoAmt	0.810557
Breed1NoAdoptionRate	1.525559
Breed2NoAdoptionRate	1.073028
Age_Years	1.241386
Color_Brown	1.012354
Color_Cream	0.957843
Color_Golden	0.988696
Color_Gray	0.97198
Color_White	0.962747
Color_Yellow	1.020607
State_Kedah	1.056329
State_Kelantan	1.023365
State_Kuala Lumpur	1.149281
State_Labuan	1.014408
State_Melaka	1.158584
State_Negeri Sembilan	1.093015
State_Pahang	1.024321
State_Perak	1.056513
State_Pulau Pinang	1.111282
State_Sabah	1.013551
State_Sarawak	1.081278
State_Selangor	1.046049
State_Terengganu	1.026876

LOGISTIC REGRESSION

- Logistic Regression
  - Feature set selected:
    - Features with non-zero coefficients from L1 logistic regression model
  - Tuned Hyperparemeter (GidSearchCV):
    - C: 10
- Model Scores (L2)
  - ROC_AUC: 0.681
  - Brier score: 0.185

RANDOM FORESTS

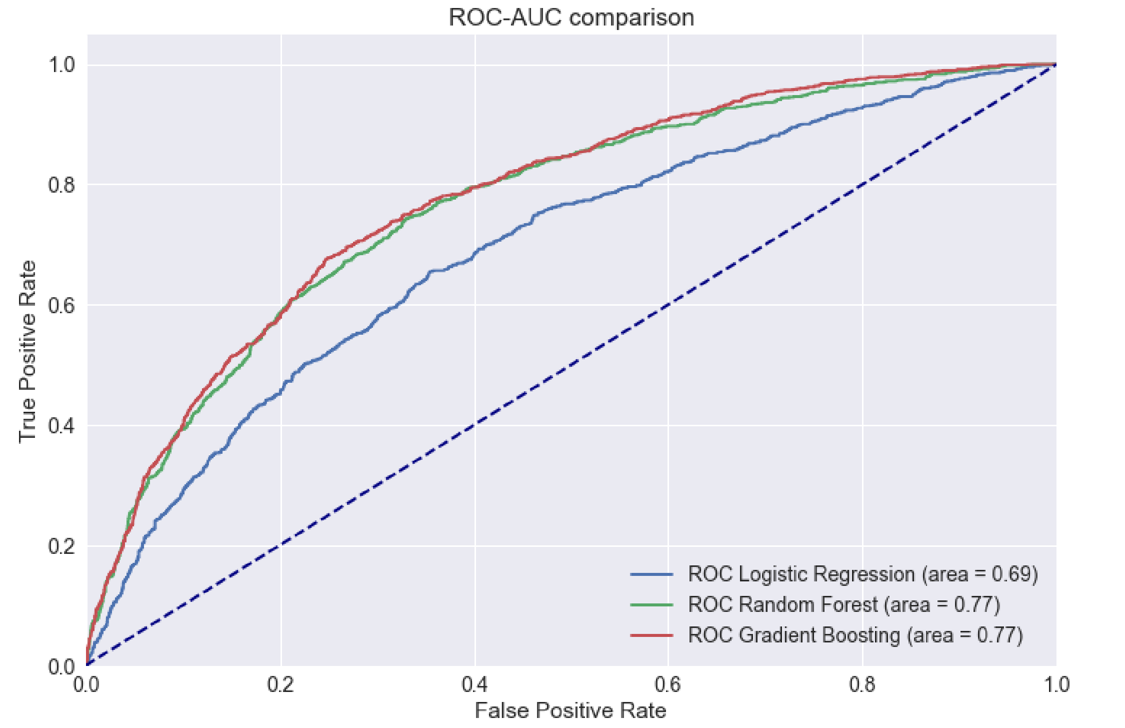


- Random Forest
  - Features set selected:
    - Feature set extracted using 5% significance p-value threshold
    - Feature importance extracted using scikit-learn feature importance attribute
  - Tuned Hyperparemeters (GridSearchCV):
    - N_estimators: 200
    - Max_depth: 7
- Model Scores
  - ROC_AUC: 0.77
  - Brier score: 0.168

GRADIENT BOOSTING

- Gradient Boosting
  - Features set selected:
    - Feature set extracted using 5% significance p-value threshold
  - Tuned Hyperparemeters (RandomSearchCV):
    - N_estimators: 100
    - Max_depth: 4
    - Learning_rate: 0.1
- Model Scores
  - ROC_AUC: 0.77
  - Brier score: .163

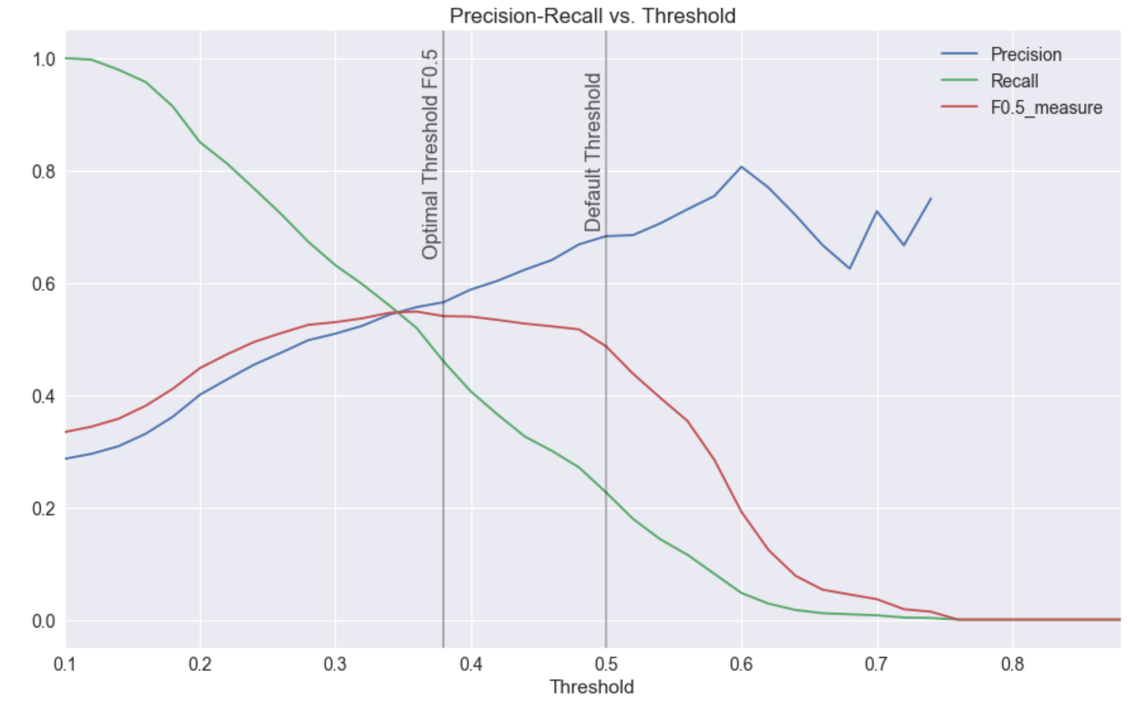
MODEL METRICS



Gradient Boosting Model provides best ROC-AUC

- Gradient Boosting model yields best ROC AUC and conveniently also the best Brier score at 0.163
- Model can be used for the purposes of both classifying animals and obtained ranked probabilities of animals at highest risk of not getting adopted

THRESHOLD



Optimal Threshold for Classification Model

- To minimize 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
- Assuming we can prioritize ~ 25% of animals, the F0.5 measure was used instead, and the resulting best threshold is 0.38 which identified 24.7% of animals of being at risk.
- While selecting the appropriate threshold is important for binary classification output, the raw ranked probability value for each prediction would be more useful asset to animal rescue organizations. Using the raw probability output instead, the Gradient Boosting model has a good Brier score at 0.17.

CONCLUSION

The insights derived from this project can aid workers in both understanding the factors that impact adoptability and in implementing changes to pet's profile

- The Random Forest model can be used to identify and rank animals at highest risk of not being adopted

- Simply adding breed composition for the animal can boost their chances of adoption

Future work:

- Study whether animals' cuteness or quality of their photos have an impact on adoption

- Use Google's Vision API to identify animal's breeds

- Use NLP to derive insights from the animals' description

- U.S. PetFinder profiles have more features available (e.g.: house-trained) so we could use those to improve the predictive power of the model

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].