Capstone Project 2
Petfinder Milestone Report 1

Table of Contents

Capstone Project 2
Petfinder Milestone Report 1

Objectives:

The objectives for this report are to...

1. explore and wrangle the data: identify anomalies in the data set, identify the best approach to address each anomaly, and apply the changes.
2. decide which features to bring forward into machine learning
3. identify the machine learning algorithms to apply the data set to

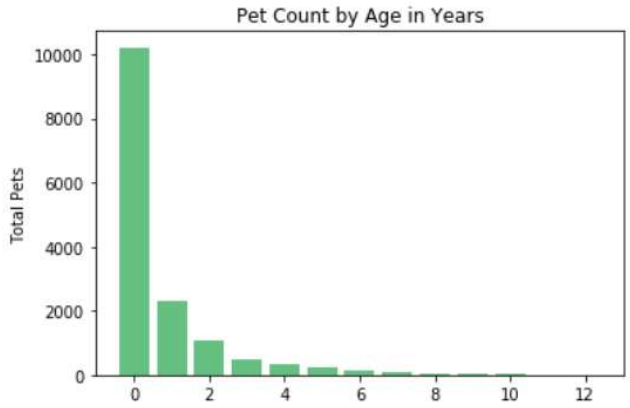
About the Data:

Petfinder has provided the data at this link, which consists of 14,993 pets (54% dogs and 46% cats) that were up for adoption. Along with various features, this data also contains a column called "Adoption Speed", which is the predictor variable for this project.

Explore and Wrangle the Data:

Feature: Age

1. No missing data: all pets have an entry for their age (in months)
2. The scatterplot of age in months shows 10 outliers beyond 12 years old. This seems unlikely, so these entries will be dropped
3. Correlation between age and adoption speed is only 10%
4. Regardless of age, the percent of pets adopted in one day ranges from 2 to 4%
5. On mean, 28% of pets are not adopted.
6. Adoption rate of pets <1 year old rounds up to 80%, while pets over 9 years old have around a 50% adoption rate.
7. The age of pets up for adoption is skewed towards <1 year old (over 10,000 pets)



Capstone Project 2
Petfinder Milestone Report 1

Feature: Type (dog or cat):

1. No missing data: all pets have an entry for their type.
2. The ratio of dogs to cats is reasonably close at 54% dogs and 46% cats
3. Correlation between Type and Adoption Speed is -9%.
4. Mean adoption speed is similar for dogs and cats at 2.6 and 2.4 respectively.
5. The ratio of adoption speeds for dogs vs cats for each year of age varies, but is not consistently higher or lower for one versus the other. The Type feature will probably not be used in machine learning.

Feature: Breed

The Breed feature is represented by two columns: *Breed1* - Primary breed of pet and *Breed2* - Secondary breed of pet, if pet is of mixed breed.

1. No missing data: all pets have an entry for their breed in the training data Breed1 and Breed2 columns. And the Breeding_Labels data also has no missing values.
2. A comparison of the mean adoption speed for dogs with the adoption speed for the breeds with the highest number of samples(>30 pets), shows a negligible difference in adoption speed. The conclusion is that of the breeds with sufficient samples, there is not a particular breed that deviates significantly from the mean.
3. The prior comparison for dogs, also holds true for cats.
4. Correlation between Breed and Adoption Speed is 11%.

Feature: Name

1. 8% (1,200) of the 14,993 pets are not named.
2. Mean adoption speed with a name is 2.5 and without a name is 2.6. Both of these values are close to the overall mean adoption speed (2.5).
3. Correlation between Name and Adoption Speed is only 2%.
4. It seems reasonable at this time to not consider the Name feature when machine learning is started.

Feature: Gender (male, female, or mix – group of pets):

1. No missing data: all pets have an entry for their gender.
2. The ratio of males, females, and mixes is spread across 49% males, 37% females, and 15% group of pets.
3. Correlation between Type and Adoption Speed is 6%.
4. Mean adoption speed is similar for males, femailes, and mix 2.4, 2.6, and 2.6.
5. The %not adopted for groups of pets is higher than the %not adopted for males & females across all years of the pets' lives (ages 0 – 12). In some cases by as much as 10%.
6. In all but one year (year 6), the %not adopted of males is less than %not adopted for females. Or, males have a better chance (in some years as high as 10%) of being adopted than females

Feature: Maturity Size:

1. No missing data: all pets have an entry for their gender.
2. Across all ages (0 – 12 years) of pets, the number of medium size pets is always higher than the other types. As a result, the %NotAdopted pets is higher for medium dogs in over half the years.

3. Correlation between Type and Adoption Speed is 5%.
4. Mean adoption speed is similar across sizes, ranging from 2.35 to 2.57; which is also very close to the overall mean adoption speed for pets.

Feature: Adoption Fee:

1. The vast majority of pets (12,663 of 14,993) have an adoption fee of $0. For the remaining 2,000+ pets the adoption fee ranges as high as 3,000 ringgits ($730 US).
2. The correlation is -0.5%, which makes sense given the number of pets with no adoption fee.
3. Adoption fee will not be carried forward into machine learning.

Feature: Media (Number of Photos and Videos):

1. There are not missing values for photos and videos (though a number of pets have 0 photos and/or videos)
2. Pets with no media have a mean adoption rate of 3.1, versus the 2.5 mean adoption rate across all pets.
3. Pets with at least 1 photo and 1 video have a mean adoption rate of 2.4, which is very close to the 2.5 mean adoption rate across all pets.
4. Pets with at least 5 photos and 2 videos have a 2.5 mean adoption rate, which is the same as the overall mean.
5. Correlations for...
    a. Photos feature: -2.0%
    b. Videos feature: -2.0%
    c. Photos + Videos feature: - 2.0%
    d. With photos or videos = 1, without photos or videos = 0: -8.0%

Feature: Health (Vaccinated, Dewormed, Sterilized, Health):

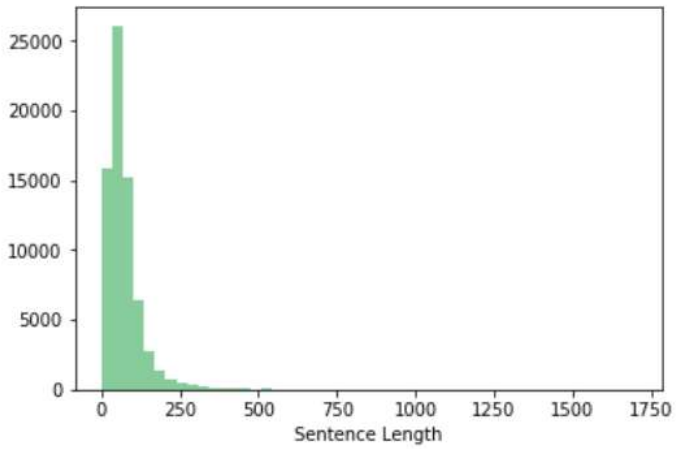
1. There are no missing values for the four health features.
2. The approach for this feature set is to compare the pets with a perfect health score (value of 1 for all features) with pet's that don't have a reasonable health score (decided this is a value > 10 meaning the pet scored less than perfect in several health features.
3. Pets with a perfect health score have an adoption speed mean of 2.91, which is higher than the 2.5 overall mean adoption speed.
4. Pets with a less than perfect health score (> 5) have an adoption speed mean of 2.4 and pets in poorer health (>10) have an adoption rate of 2.9. A general conclusion is that people adopting pets are open to adopting pets with less than perfect health.
5. Using the HealthSum (add scores of all four health features) correlation is -5.0%

Explore and Wrangle: Natural Language Processing (NLP):

Feature: Pet Description:

1. There are no missing Descriptions

2. Combining all the Descriptions to a single text file, there are almost 70,000 sentences with over 900,000 words.



Summary of Key Findings:

1. A number of the features (Age, Type, Breed) have correlation to Adoption Speed in a range of 9 to 11%. Given this relatively low correlation, machine learning should lean towards ensemble and forests with the premise that a number of weak individual predictors when combined can result in a predictor model of high accuracy.
2. "Name" will not be used in machine learning; correlation to Adoption Speed is only 2%
3. "Gender" correlation is only 6%, but given the variance in %not adopted across males, females, and mixed breeds it seems that gender could be a significant feature to include in machine learning.
4. Adoption fee will not be used in machine learning, because all but roughly 2,000 pets have no adoption fee and correlation with adoption rate was only -0.5%.
5. Media (photos and videos) negatively impact the adoption speed when neither is present, but when at least 1 photo and 1 video are present the adoption speed matches the overall mean. The conclusion is that posting many photos/videos does not impact adoption speed, but not posting any media has a negative impact (mean drops from 2.5 adoption speed to 3.1).
6. Health (as measured by the combination of vaccinated, dewormed, sterilized, and health) has a correlation of only -5.0%.
7. Combining all the Descriptions to a single text file, there are almost 70,000 sentences with over 900,000 words.

Exclusions:

1. Did not explore the Fur Length feature because this does not seem like a relevant feature. May reconsider this feature during machine learning.
2. Did not explore Color feature because colors are broken out across three columns and expectation is that color results will be similar to breed results (breed was broken out across multiple columns as well). May reconsider this during machine learning.

Proposed Machine Learning Algorithms:

These are the machine learning algorithms to start with...

1. Classification and Regression Trees (for classification or regression)
2. Naive Bayes
3. k-Nearest Neighbors
4. Learning Vector Quantization
5. Support Vector Machines
6. Bag of Words
7. Word-2-Vec