Nora Keenan
June 30, 2019

Petfinder.my
Capstone 2 Final Report

Problem Statement

Animals all over the world are waiting in shelters for adoptive homes. And the length of time before an adopter is found can literally mean the difference between life and death. One organization is looking to data science to help find a solution. Petfinder.my is Malaysia's leading pet welfare platform. Site visitors can search for new pets by location, type, breed, color, gender and age, among other criteria.

Petfinder.my partnered with Kaggle, the data science competition website, to launch a new challenge using pet adoption data pulled from the site. Petfinder.my wanted to know: Can a machine learning model accurately predict how soon a pet will be adopted? With this information they hope to improve pet profiles so animals find forever homes faster.

Description of Dataset

Petfinder.my provided four subsets of data on the same group of 14,993 cat and dog adoptions. The datasets are all available for download from the Kaggle website.
https://www.kaggle.com/c/petfinder-adoption-prediction/data

1. Train.csv consists of feature information regarding each adoption. The label AdoptionSpeed is a numeric value between 0 and 4 represents the length of time the pet was available before it was either adopted or had been available for adoption at least 100 days. 2,337 dogs and 1,759 cats had not been adopted after 100 days.

   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).

2. JPEG file of primary image from each adoption profile on the website.

3. Results from Google's Vision API analysis of each image specifying the likely label annotations.

4. Results from Google's Natural Language API which provides analysis on sentiment and key entities of the description of each pet.

Dataset Cleaning and Wrangling

The following manipulation was done on the train.csv dataset:

1. One hot encoding of categorical features except those with ordinality (fur length, maturity size, health)
2. Reviewed dataset for missing or outlier values. Found one misclassified adoption speed (value = 6, which is not an option). Updated to 4. Otherwise data was clean.
3. Ensured numeric features are in dataset as numeric and not text.

The following transformation was done on the Google VIsion API image metadata:

1. Filtered labels on those with a score of 0.94 or higher to focus on labels which most likely to be correct.
2. Truncated multi-word labels to first word only. For example, "dog like animal" was truncated to "dog."
3. Duplicate PetID/label values deleted
4. One hot encoded labels to pivot unique labels to columns and assign values of 1 or 0 to each PetID/Label interaction.
5. Saved as .csv file

The two CSV files were then merged into one dataset on PetID.

Exploratory Data Analysis and Inferential Statistics

Initial Findings

After cleaning the dataset I completed a correlation analysis of features for dogs and cats. These are the strongest positive and negative correlations to AdoptionSpeed found:

Dogs

-0.105275	Fur Length
-0.070336	Sterilized (Is the animal sterilized? Yes/No/Mixed or Unknown)
-0.038337	Vaccinated (Is the animal vaccinated? Yes/No/Mixed or Unknown)
0.065419	Quantity (Number of animals in the posting)
0.083946	Gender
0.166792	Breed 1

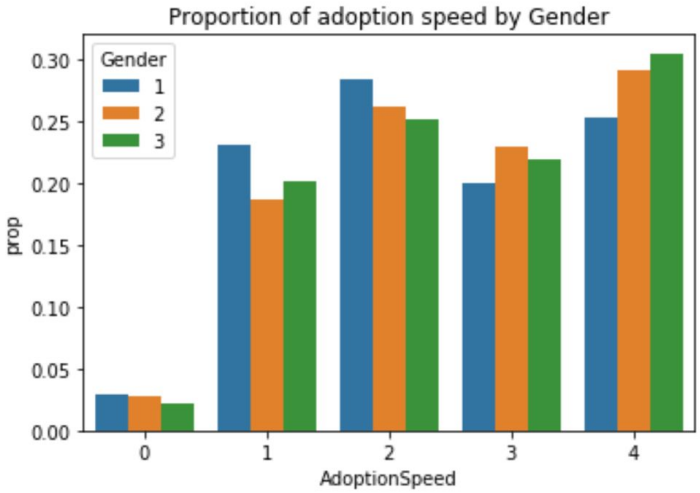
Cats

-0.100239	Sterilized
-0.076639	Fur Length
-0.067206	Vaccinated
0.043446	Gender
0.068437	Quantity
0.143387	Age

Data Visualizations

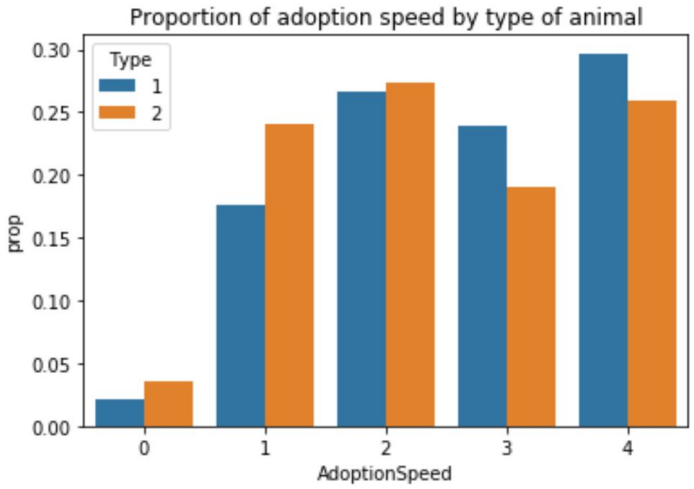
Through a review of data visualization, the following trends and characteristics were noted:

Male animals appear to be adopted faster than female animals.
1=Male, 2=Female, 3=Mixed litter or Unknown



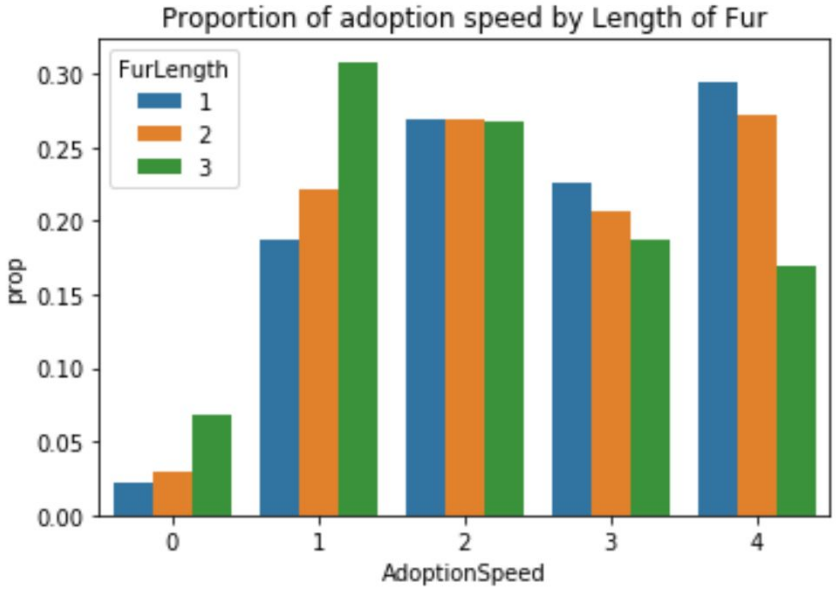
Cats appear to be adopted faster than dogs

1=Dog, 2=Cat



Animals with long fur appear to be adopted faster than those with short fur

1=Short length, 2=Medium length, 3 = Long length



Measures of Central Tendency: Adoption Speed

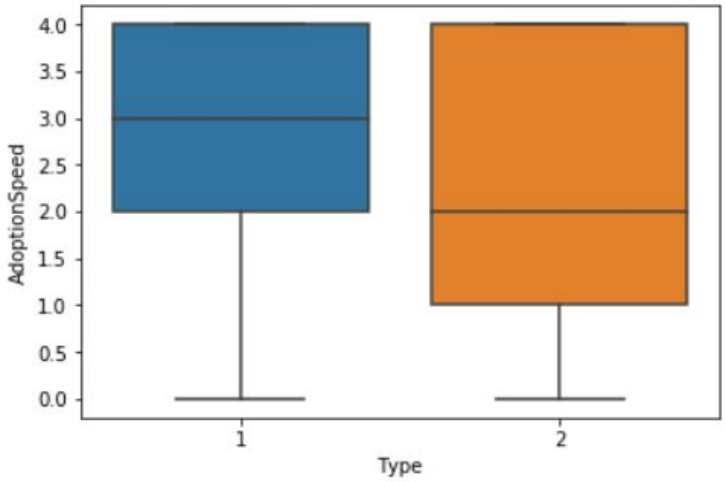
<u>Dogs (Type 1)</u>

Mean: 2.62

Median: 3

Mode: 4

Cats (Type 2)
Mean: 2.34
Median: 2
Mode: 2



Inferential Statistics

The visual representations of the data suggest the following:

1. Cats are adopted faster than dogs
2. Animals with long fur are adopted faster than animals with short fur
3. Male animals are adopted faster than female animals

For this report I will be testing the suggestion that male animals are adopted faster than female animals. Said another way, I will test whether or not the proportion of male animals adopted in the first month is statistically higher than the proportion of female animals adopted in the first month.

Values:
Total number of animals categorized as male or female: 12,510. Remainder are categorized as mixed (for postings with multiple animals) or unknown.

Population 1: Male animals adopted in the first month
Proportion ($p1$) = 0.544523 (2923 of 5368 males)
N ($n1$) = 5368
Population 2: Female animals adopted in the first month
Proportion ($p2$) = 0.478577 (3418 of 7142 females)
N ($n2$) = 7142

Null Hypothesis: The proportion of male animals that are adopted in the first month is equal to or less than the proportion of females that are adopted in the first month. $H_O = P_1 <= P_2$

Alternative Hypothesis: The proportion of male animals that are adopted in the first month is greater than the proportion of female animals that are adopted in the first month. $H_{alt} = P_1 > P_2$

$$p = \frac{(p1*n1+p2*n2)}{(n1+n2)}$$

$$p = \frac{(0.544523*5368+0.478577*7142)}{5368+7142}$$

p=0.507

Standard Error

$$SE = \sqrt{p(1-p) * (\frac{1}{n1} + \frac{1}{n2})}$$

$$SE = \sqrt{(0.507(1-0.507) * (\frac{1}{5368} + \frac{1}{7142}))}$$

SE = 0.0000804

z score

$$z = \frac{p1-p2}{SE}$$

$$z = \frac{.544523-.478577}{.0000804}$$

z = 820.224

Probability of a z value of 820.224 < 0.00001
0.00001 is less than 0.01. The null hypothesis is **rejected**.
Male animals are adopted faster than female animals.

Machine Learning

The primary determining factor for model selection was the nature of the label to be predicted. The label for this dataset, AdoptionSpeed, is ordinal. A low value (0,1, or 2) is a more desirable result than a higher value (4 or 5). I chose to run two types of models against the dataset: an

SKLearn Linear Regression model as a baseline and a Keras deep learning neural network model.

I followed these steps in my machine learning process:

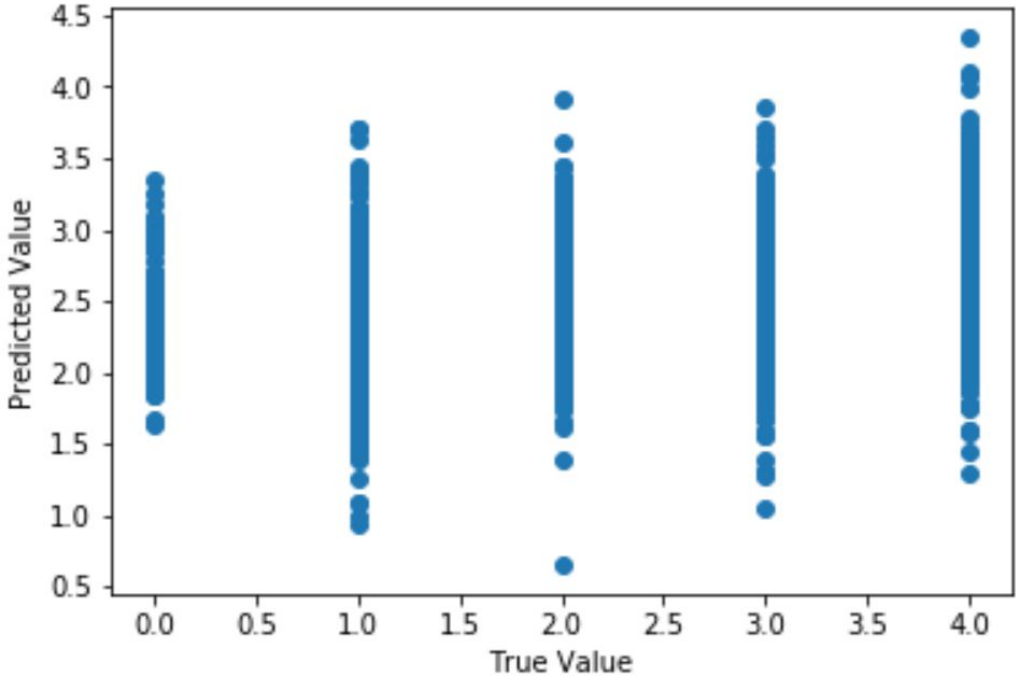
1. Pulled image label data from Google API JSON files
    a. Used only labels with a score of 0.94 or above to focus analysis on labels with a strong likelihood of being correct.
    b. Truncated the multi-word labels to first word only. E.g., “dog like animal” was truncated to “dog”.
    c. Removed duplicate image/label pairs.
    d. Saved as a csv file
2. Merged feature data with image metadata.
    a. Assigned get_dummies transformation to categorical features
    b. Dropped text features (PetID, RescuerID, Name, Description)
3. Ran 20 training scenarios on Keras neural network model with varying numbers of nodes, layers and types of layers, to identify best performing permutation over 20 epochs. Determined that input plus two hidden layers and an output later was best performing.
4. Ran Keras model for 150 epochs before accuracy finally settled around 90%.
5. Built Linear Regression Model
    a. Divided the df_shifted and y variables into training and test sets.
    b. 70% train /30% test ratio

The linear regression model struggled with the dataset, resulting in a mean squared error of 1.30233. The neural network did a better job of estimating the adoption speed with a mean squared error of 0.12008 and an accuracy of 90.59%.

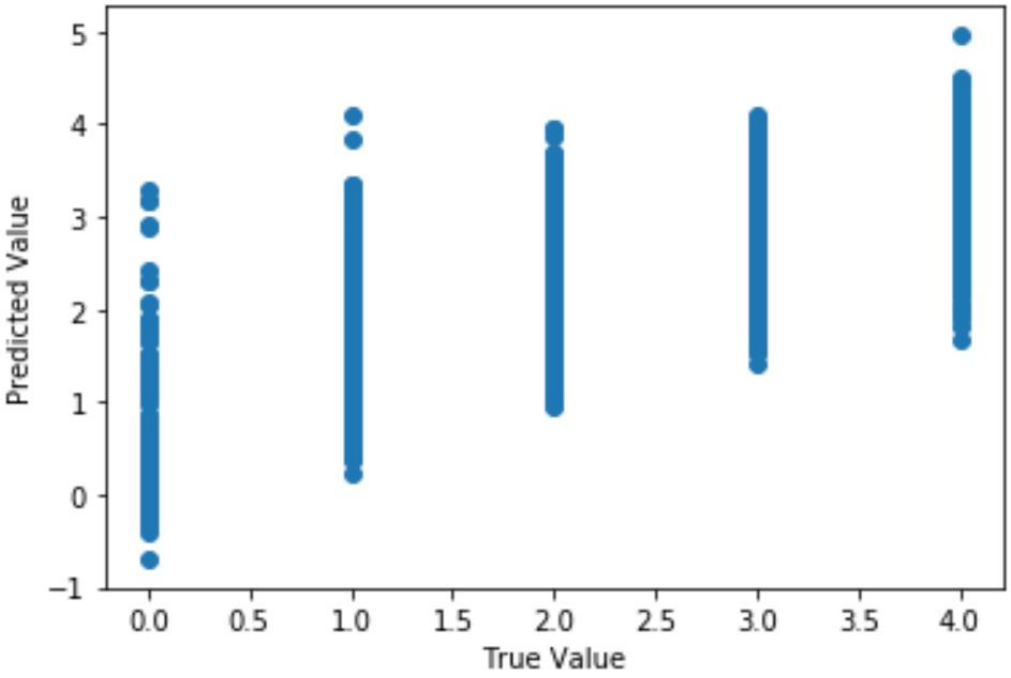
Model details:
The SKLearn Linear Regression model was built with a 70/30 train test split on the data set.

The Keras neural network model had an input layer (475 nodes), two hidden layers (125 nodes each) and an output layer with one node.

Linear Regression Predictions vs Actual:



Keras Neural Network Predictions vs. Actual:



Conclusion

I found this project interesting in that there were definite action items that could be implemented to decrease the amount of time animals were in the shelter. It was also a good project to learn Keras deep learning. The deep learning model performed much better than the linear regression model, which validates the value of the approach.

The project could continue to be expanded on by incorporating the natural language metadata on the pet descriptions. Also, a separate analysis could be done on the images themselves to determine the characteristics that resonate with adopters.