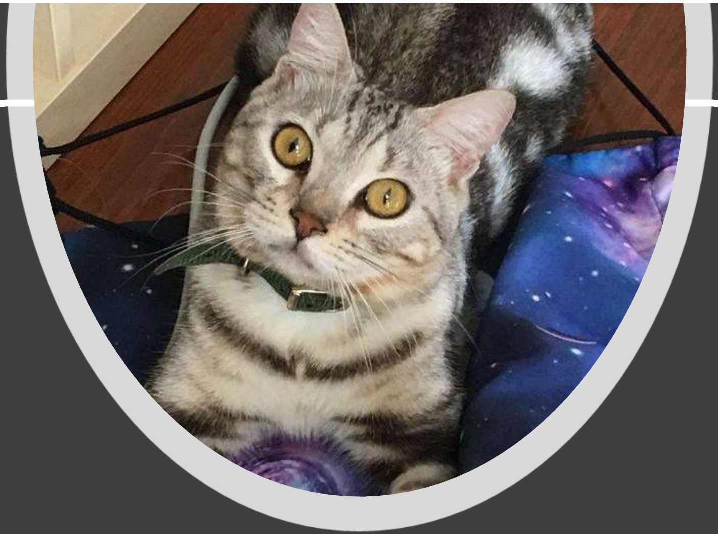
Predicting Pet Adoption Speed

Jordan Weiss – DSI – 11 – Capstone



Executive Summary

- Around the world millions of dogs and cats sit in shelters, waiting for their forever-homes, which many will never find. Over the course of history, our pets become reliant on humans for food, shelter, and companionship. As a result, dogs and cats need humans to adopt them from shelters and bring them into their homes. The demand for shelter animals varies widely based on the characteristics of each animal, as well as the website post for their adoption.

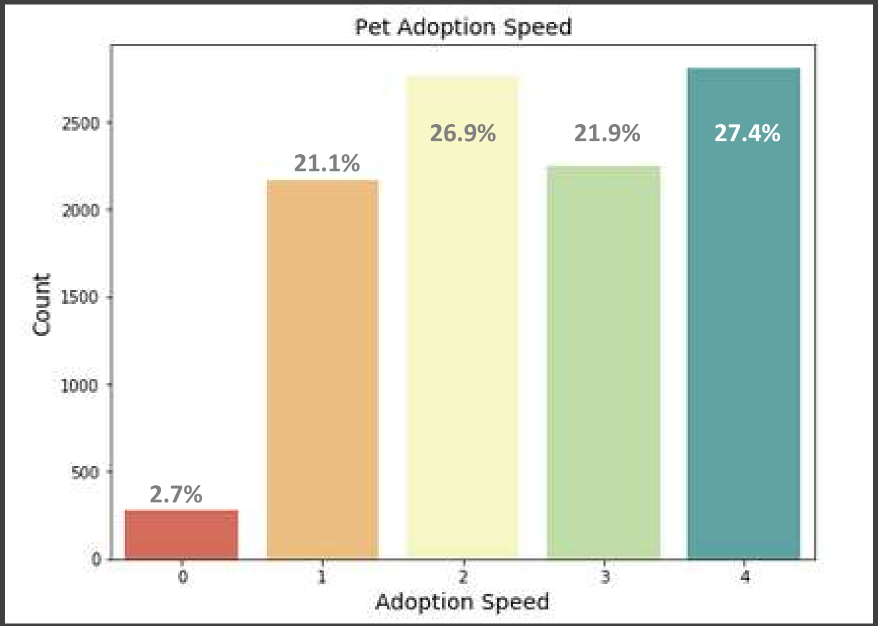
- We will seek to use data from PetFinder (Malaysia) describing each pet, as well as the metadata from their Petfinder posts to predict the speed at which pets get adopted. Specifically, we will make a classifier model that will predict the speed at which the pet is adopted. We will judge our success on the accuracy of our predictions. Additionally, we will set aside a portion of our data as a validation set to confirm our model's ability to estimate new entries.

- By predicting the speed at which pets get adopted, we hope to gain insight into actionable measures we can take to increase the rate and overall number of adoptions.

Initial Data Exploration

- The data include both dogs and cats, representing 54.6% and 45.4% of the entries respectively

- The data include many key features, including:
    - Breed
    - Age
    - Hair Length
    - Size at a Maturity
    - Color
    - Health Information
    - Presence of Pictures and/or Video



- The target for classification is the speed of adoption for these pets on a scale of 0 – 4:

  0 – Adopted the day the pet was listed

  1 – Adopted after the first day, but within the first week of listing

  2 – Adopted after the first week, but within the first month of listing

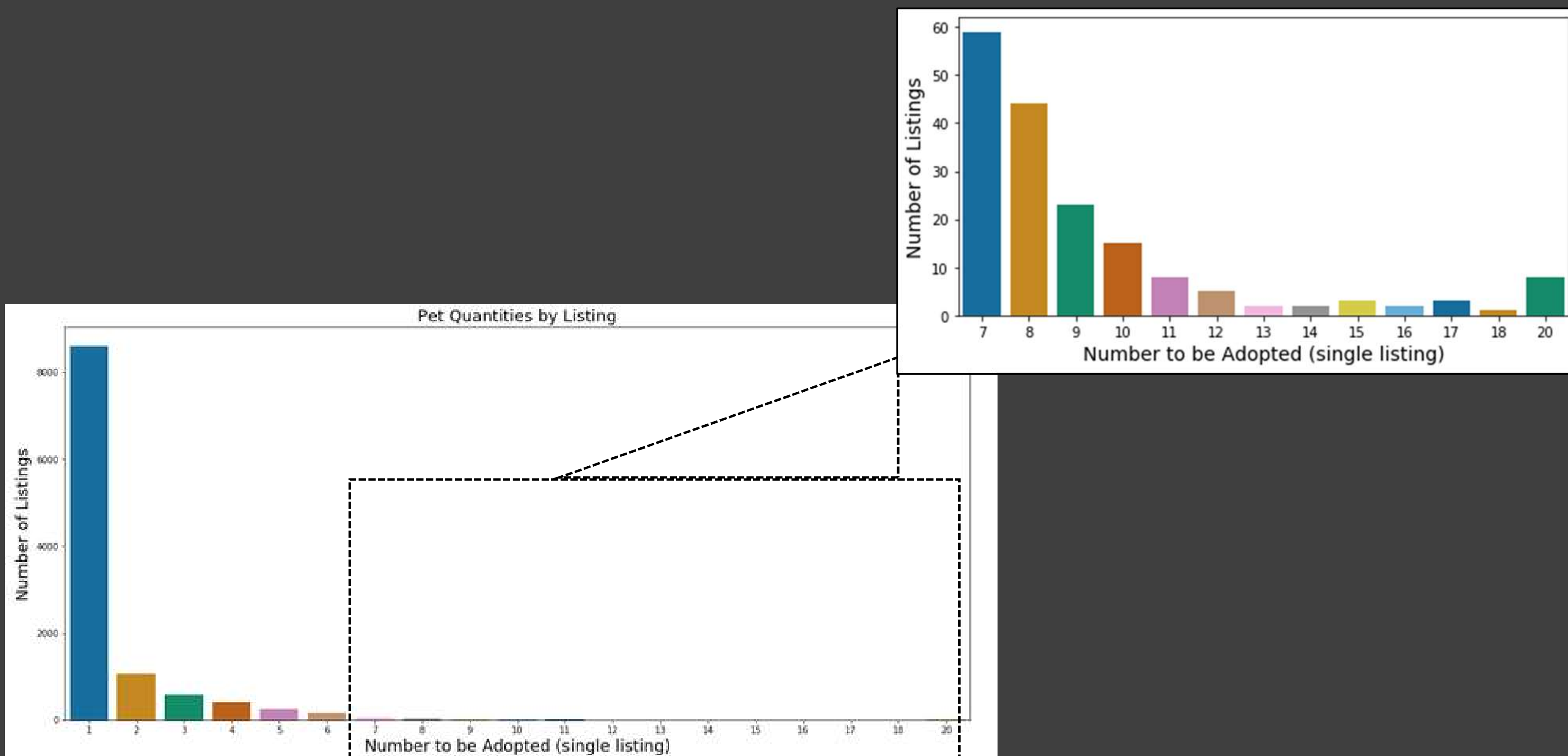
  3 – Adopted between 31-90 days after listing

  4 – Adopted after 100 days, or never adopted

Initial Data Exploration and Analysis

Observations

- Tragically, many of the items included were for entire litters of pets. These were excluded from the analysis.
- Litters and groups of pets ranging in size from single animals to groups of 20 have been listed for adoption.
- For our predictions, only groups of three or fewer were included.



Modeling

- As our goal in this project is to predict the speed of adoption, using the time frames provided in the data, we used classification models to attempt to predict speed of adoptions.
- To assist in the classification, we used a K-means cluster model to add a o the data set with the predicted cluster for each pet.
- We tried a wide range of models to determine the best estimator:
  - K Nearest Neighbors
  - Decision Tree
  - Bagged Decision Tree
  - Random Forest
  - Extra Trees
  - Neural Network
- For fun and completeness we also attempted some regression modeling:
  - Linear Regression
  - K Nearest Neighbor
  - Neural Network Regression
- As predicted, the results from regression modeling were not on par with the classification models.

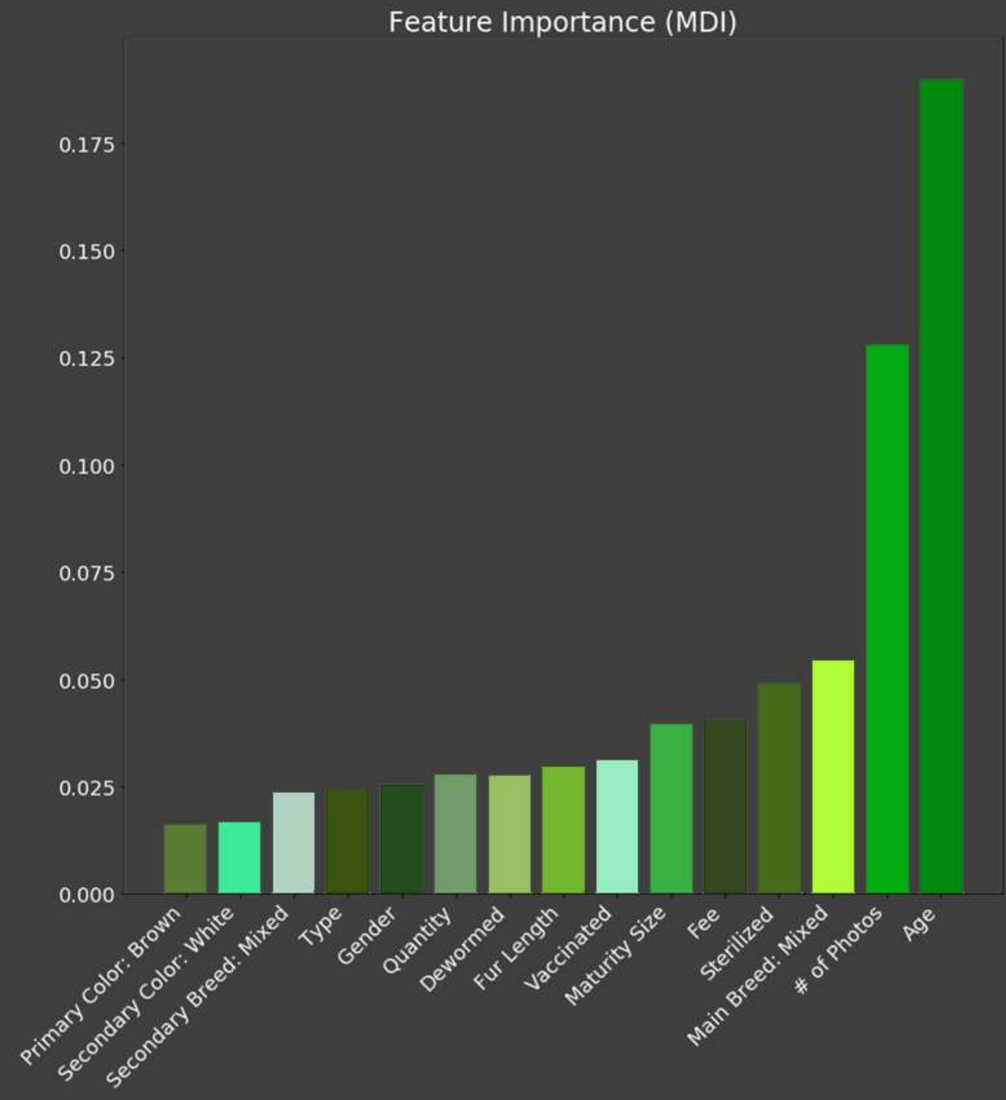
Model Selection

- Most models performed in the high 37% - 39% range
- There is a fair amount of overfitting occurring
- We select the Random Forest model for the rest of our analysis

	Train Score	Test Score	Cross-Validation
Baseline	27.37%	27.37%	27.37%
K Nearest Neighbors	25.59%	25.94%	32.63%
Decision Tree	41.47%	37.65%	38.67%
Bagged Decision Tree	43.36%	37.69%	39.48%
Random Forest	58.76%	38.90%	39.62%
Extra Trees	55.06%	37.85%	38.43%
Neural Network	36.00%	33.24%	N/A

Evaluation – Feature Importance

- We looked at both the feature importance and the permutation importance of features to determine what ‘moves the needle on adoption speed.
- The importance of each feature is determined by the Mean Decrease in Impurity (MDI)



Evaluation – What we got right, what we got wrong

- Out of 2,563 test observations, we were exactly right 1,013 times. This is an accuracy of 39.5%
- Though the exact accuracy isn't great, we have a number of 'near misses' which is still helpful.

Adoption Speed Confusion Matrix

Actual Adoption Speed \ Predicted Adoption Speed	Adopted first day of listing (0)	Adopted first week of listing (1)	Adopted first month of listing (2)	Adopted first 90 days (3)	Adopted 100+ days/ Never (4)
Adopted first day of listing (0)	1	28	16	2	11
Adopted first week of listing (1)	2	185	216	32	109
Adopted first month of listing (2)	0	142	291	65	193
Adopted first 90 days (3)	0	103	197	92	177
Adopted 100+ days/ Never (4)	0	85	153	35	428

Evaluation – What does it mean

- Precision tells us what percentage of the ones we predicted in each category were correct
- Recall is what portion of the pets in each category were accurately identified.
- We created a special metric called Within 1; this metric is precision, but it considers a true positive to include one time-frame ahead as correct (E.G. if the model predicts a 3, but it is a 2, it counts as a true positive)

Evaluation Statistics

Category	0	1	2	3	4
Precision	33.33%	34.07%	33.33%	40.71%	46.62%
Recall	17.24%	34.01%	40.96%	15.11%	61.06%
F1 - Score	3.39%	34.26%	36.28%	21.58%	29.61%
Within 1*	33.33%	39.23%	58.08%	69.47%	65.90%

* Precision with allowance for actual being 1 time category faster than prediction

Conclusions and Recommendations

- With the available data we achieve 39.5% accuracy, however having many ‘near-misses’ indicate, we do have a good sense of what pets may be adopted quickly, compared to those that take longer or are never adopted

- The ability to predict adoption speed will allow us to better allocate resources to find a forever home for as many pets as possible.

- We identified some key factors in the adoption speed of pets:
  - Age
  - Spay/Neuter
  - Number of photos
  - Breed

Recommendations

- Include multiple pictures of the pet
- Even if the breed is unknown or mixed, it is best to take a guess and include a breed rather than 'mixed'
- Spay/Neutering the pet helps in adoption, and helps keep the population down for the future
- Focusing on pets identified as being more difficult to home may increase their likelihood of finding a forever home.

Resources

- DATA SOURCE - https://www.kaggle.com/c/petfinder-adoption-prediction/data
- Background on Multiclass Classification - https://www.linkedin.com/pulse/kaggle-competition-multi-class-classification-image-alexandra/
- Confusion Matrix for MCC - https://stackoverflow.com/questions/53886370/multi-class-multi-label-confusion-matrix-with-sklearn
- Interpretation of Random Forest - https://towardsdatascience.com/interpreting-random-forest-and-other-black-box-models-like-xgboost-80f9cc4a3c38
- Interpretation of Random Forest with Sklearn - https://scikit-learn.org/stable/auto_examples/inspection/plot_permutation_importance.html
- Model Saving/Loading , Pickling - https://machinelearningmastery.com/save-load-machine-learning-models-python-scikit-learn/
- Feature Alignment for Testing - https://stackoverflow.com/questions/44266677/machine-learning-test-set-with-fewer-features-than-the-train-set

ADOPT DON'T SHOP!!