Doggy Doggy What Now? Executive Summary

Project Team: John P. Harden, Angela Kubena, Jun Bo Lau, Claire Merriman, and Robert Young

Mentor: Evelyn Huszar

GitHub: https://github.com/rty10/doggy-doggy-what-now

Overview: The Humane Society states that over 3 million dogs enter animal shelters around the United States each year, and around 2 million dogs are adopted each year. Shelters are understandably busy, noisy, and fast-moving places where many challenges present themselves. How shelter managers allocate resources for the future could greatly minimize the chaotic nature of the animal shelter process. Our group sought to leverage machine learning and over a decade of observations to predict animal shelter intakes and outcomes, particularly adoptions.

Stakeholders:

- Animal shelter managers: Improved resource allocation.
- Taxpayers and grant providers: Greater certainty regarding effective use of taxes and donations.
- Not-for-profit executives: Potential applicability of these models to their own service areas.

Key Performance Indicators:

- Relative accuracy of time series models predicting intakes and outcomes (1 – NRMSE).
- Accuracy score of random forest model predictions and ROC curves.

Data: We use publicly available data from Sonoma County, CA and Austin, TX animal shelters. We analyze those data sets separately given differing distributions for intakes and outcomes. We focus our analysis on canines only, involving key features such as intake and outcome (sub)types, dog size, breed, color, etc.

Modeling Approach: We use two approaches to leverage machine learning and gather insights.

- First, we use time series models, focusing largely on ARIMA and rolling averages. We use a series of train-test splits and train many models to find the most accurate predictions for various features.
- Second, we develop random forest models via categorical boosting to predict likelihood of dog adoption based on animal intake features as well as season of adoption based on the Sonoma County, CA data.

Results/Model Performance:

- Our best time series models achieve accuracy above 90% for intakes and outcomes in both Sonoma County, CA and Austin, TX.
- We achieve 74% accuracy with our random forest model for our test data.
- Model results indicate using macro-level features like unemployment and inflation can provide insight into the coming days, weeks, and months. Meanwhile, shelter-level features such as the time a dog has been in the shelter, the season, and the size, age, and condition of a dog at intake predict whether a dog will be adopted. Notably, dog breed and dog color do not provide a leg up in predicting adoption.

Opportunities and Future Directions: There are opportunities for improvements across three areas:

- Data-Related: Large scale standardized data (Shelter Animals National Data), different time scales, detailed feature extraction (state-to-state transfer data was not available).
- Models: Neural networks to model state-to-state transfers, extreme value analysis to improve time series forecasts.
- Focus: Analysis could be applied to felines and other species. Additionally, do other non-for-profits that handle intakes fit our modeling approach.