Connor McCormick
May 23, 2017
Stat Learning

Shelter Outcomes: Approach and Analysis

Looking at shelter outcomes for animals in the Austin Animal Shelter, my final focuses on finding a way to better predict animal outcomes in shelters. 2.7 million animals are euthanized every year, so this Kaggle competition aimed to provide better insight into what leads to certain outcomes for shelter animals. With 7.6 million animals entering US shelters every year, nearly 1/3 of them do not find a forever home. Using information including name, breed, color and age, I examined what determines outcome types (adoption, transfer, return to owner, euthanasia, died) for animals.

The outcome I was looking at was the final outcome for animals that reach the Austin Animal Center, which is categorical. I was looking to predict whether an animal was adopted, transferred, returned to owner, euthanized or died. Since the outcome was categorical, I looked at both CART (Categorical and Regression Trees) and KNN (K-Nearest Neighbors). I created a series of variables from the existing data set to better understand the data and make it more easily analyzed by the models. I wanted to examine the effects of sex, month of outcome, age, mixed breed, fixed, dog and solid coat.

For my exploratory data analysis, I examined a few of the variables more closely and looked for trends. One of the major defining factors is the effect of being fixed on outcome type. 97% of all the animals put up for adoption at the Austin Animal Shelter have been spayed or neutered at the time of adoption. Adoption makes up about 40% of the outcomes at this shelter, so most animals who are fixed end up being adopted. On the other side of the spectrum, a majority the animals who were transferred were not fixed. This led to my best CART model, which has a single split for whether an animal is fixed and predicting adopted or transferred. I ran a second CART model without the variable for fixed animals, but that model performed worse, even with a greater number of splits. The code for these models is in Extras.

Using k-nearest neighbors, I used cross validation to find the ideal number of neighbors for the model. I used the MLmetrics package to find the Log Loss Error Rate, that is used in the Shelter Outcome Kaggle Competition. I more easily cross validate the best number of neighbors using this package, finding the minimum Log Loss rate for a wide range of values. After applying that model to the test set, the result was slightly better than my first CART model, so I decided to use knn. Simple models are better and the CART model is a single split using whether an animal is fixed as a predictor. The number of steps included in the knn model generation and cross validation was significantly higher than that of the CART model. I used a built-in function, the complexity parameter, to find the best fit for the CART model, but I had to write my own loop to cross validate the knn model.