Predicting Length of Stay for Shelter Dogs

Rachael Latimer

note: see *GitHub* repo for processing files

As people stayed home for most of the early days of the coronavirus pandemic, the demand for goods and services increased. Some of the increased demand was not surprising and likely expected; items such as home workout equipment, trampolines, and lumber. However, some of the demand took the industry by surprise: yeast for baking, and pets. In fact, the interest in pet adoption increased so much that shelters were regularly reporting empty kennels and sifting through dozens of adoption applications for a single puppy.

Unfortunately, as vaccines were rolled out and people began returning to work and school, shelters and foster groups filled up with animals that were no longer compatible with people's lifestyles. The decrease in demand for dogs means that people can be more selective in the kind of dog they adopt. However, the information provided by animal shelters and rescue groups are often based on a short period of time with the animal and the animal's appearance. One shelter's Border collie mix might be another's spaniel or shepherd mix. This best guess breed identification can have significant impacts on a dog's future and could be the difference between adoption and euthanasia. Objectively identifying the impact of a dog's listed breed on the length of stay in an animal shelter could provide shelters with the information needed to shift away from listing a dog's breed as the primary information for potential adopter and toward a more holistic evaluation of a dog's temperament and future needs.

Data

The Austin Animal Center in Texas is the largest no-kill shelter in the US. The shelter maintains data on the intake and outcomes of animals beginning from October 2013 to present. This data set was obtained from **kaggle**. It was originally provided by the Austin Animal Center in Austin, Texas. The data include information about the intake and outcome of the animal, and details on the type and condition of the animal. A brief examination of the data revealed that the animal shelter takes in animals in addition to typical domestic pets (cats and dogs). For the purposes of this study, the following types of animals were excluded: cats, birds, and animals that were classified as other, including rabbits, bats, snakes, raccoons, ferrets, reptiles, and other wild animals that live in close proximity to humans. Additionally, dog breeds with sample sizes less than 20 were excluded as this small sample made it difficult to accurately model the length of stay for the breed. The final data set included variables of the animal (breed, age on intake, sex, condition of the animal),

circumstances of the animal arriving at the shelter (type of intake, month of intake), and specifics of the outcome of the animal (outcome, month of outcome, time spent in the shelter, measured in days).

Initial data visualization was performed to understand the data available. This included visualizing the number of animals taken into the shelter each month (Figure 1) and further exploring the number of each type of animal taken in each month (Table 1).

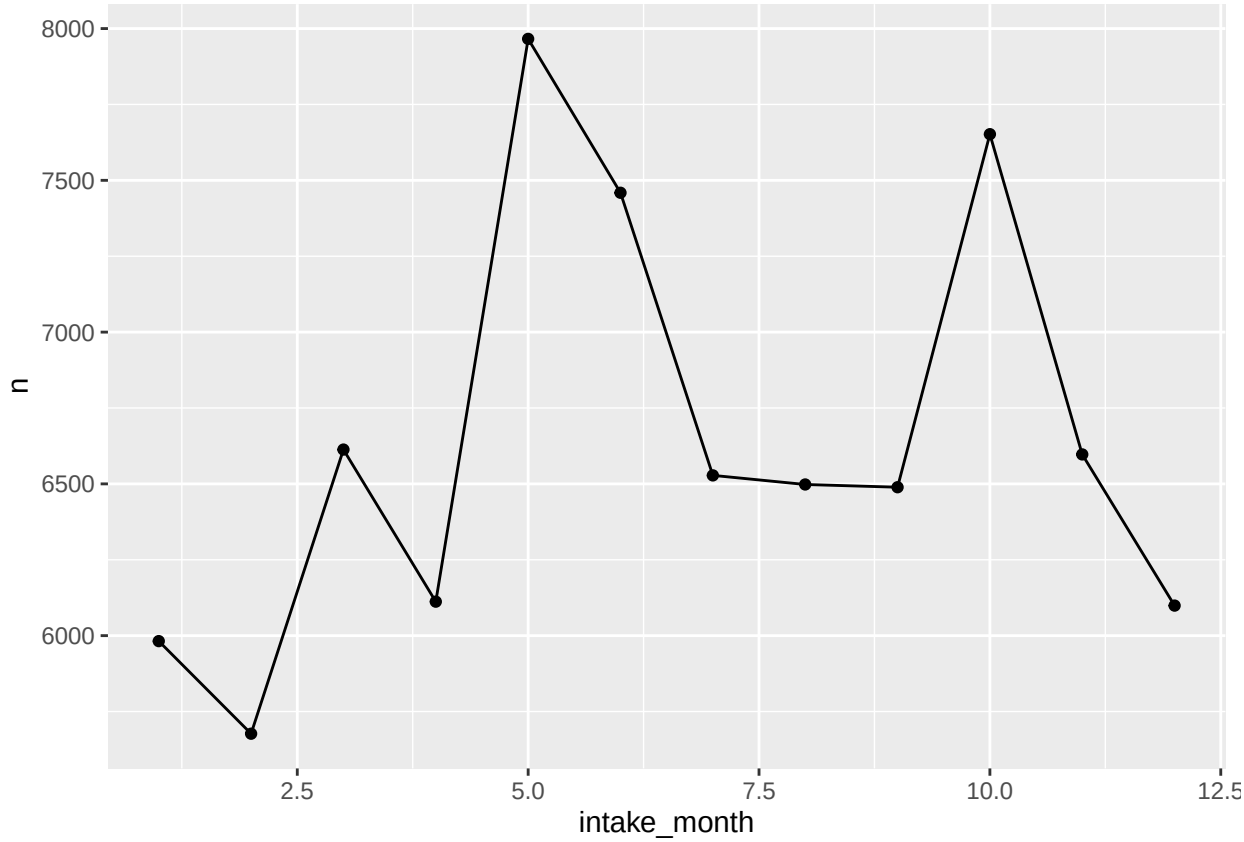

Figure 1. *Shelter Animals Taken in Each Month*

Table 1. *Type of Shelter Animal Taken in Each Month*

	Animal Type			
Intake Month	Bird	Cat	Dog	Other
1	19	1529	4181	253
2	61	1320	3970	326
3	26	1600	4096	891
4	35	2330	3353	394
5	38	3741	3867	320

6	24	3519	3556	360
7	25	2893	3312	298
8	22	2819	3271	386
9	23	2756	3420	290
10	29	3003	4215	405
11	19	2285	4038	255
12	18	1744	4087	250

It is also of interest to explore the relationship between the average length of stay in the shelter of each breed with the frequency of that breed present in the shelter. Table 2 provides the average length of stay in the shelter for the 25 most frequently taken in dog breeds.

Table 2. *Average Length of Stay (in days) for Most Common Dog Breeds in Shelter*

Breed	N	Avg. Length of Stay (days)
Pit Bull	6865	28.028652
Labrador Retriever	6260	16.906381
Chihuahua Shorthair	5726	11.841401
German Shepherd	2612	12.771461
Australian Cattle Dog	1442	19.901094
Dachshund	1239	8.671506
Boxer	948	19.438599
Border Collie	928	12.298963
Miniature Poodle	841	7.600594
Beagle	642	12.259929
Siberian Husky	641	8.506920
Catahoula	638	23.371538
Australian Shepherd	624	10.586646
Jack Russell Terrier	623	10.260468
Yorkshire Terrier	621	7.316022
Rat Terrier	598	10.200137
Miniature Schnauzer	584	5.601945
Great Pyrenees	516	10.899884
Rottweiler	493	18.182440
Shih Tzu	463	4.929401
Pointer	448	22.045905
Chihuahua Longhair	446	7.845548
Cairn Terrier	436	8.984851
Staffordshire	418	28.627440
American Bulldog	381	39.187939

Models

In order to predict the length of stay of shelter dogs, three types of modeling approaches were explored: linear regression, linear regression with ridge penalty, and bagged trees. These models were chosen for their increasing complexity to determine the extent to which the increasing complexity added value to or impacted the predictions and importance of variables

used in the predictions. All models were fit with 10-fold cross-validation for comparison purposes.

The linear regression with ridge penalty will build on the linear regression without regularization by standardizing the variables, and tuning the hyperparameter, lambda. The `caret` package will be used to fit the model.

The bagged trees model will further build on the previous two models by fitting multiple models and aggregating the results. Tuning the hyperparameter, the number of tree models, will be done using a for loop. The `caret` package will be used in conjunction with the `ranger` package to perform cross-validation and train the model.

The performances of the models will be compared using the values of R-squared, MAE, and RMSE.

Results

The linear regression without regularization produced an RMSE of 37.66, and MAE of 16.77 and an r-squared of .11. Even though I used 10-fold cross-validation while training the dataset, the r-squared of .11 on the testing data was much smaller than the r-squared on the training data of .65, suggesting the model is overfitted to the training data. This drop in performance means the linear regression model without regularization is not a realistic method for predicting the length of stay of shelter dogs. Figure 2 illustrates the underperformance of this model.

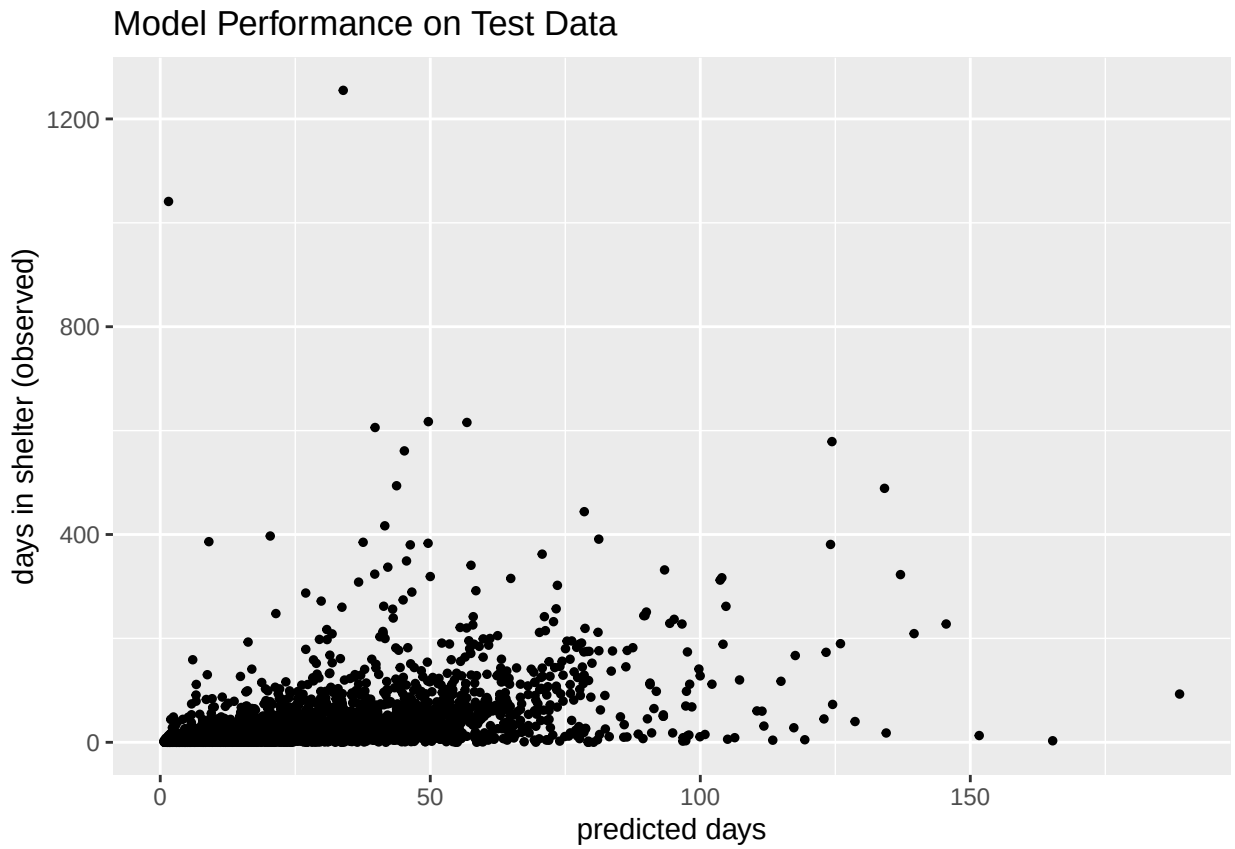

Figure 2. *Linear Regression without Regularization Model Performance*

The linear regression with ridge penalty produced an RMSE of 37.66, MAE of 16.57, and r-squared of .11. These results are very similar to those of the linear regression without regularization. The model also revealed the top ten predictors of length of stay for shelter dogs, which included eight specific breeds (Table 3).

Table 3. *Top Ten Predictors of Shelter Length of Stay from Linear Regression with Ridge Penalty Model*

	Predictor
na_ind_outcome_type	-29.54081
breed_American.Bulldog	24.37111
breed_Bulldog	20.68927
(Intercept)	19.10515
breed_Collie.Smooth	18.35328
breed_Tibetan.Spaniel	-18.22182
breed_English.Coonhound	17.58684
breed_American.Pit.Bull.Terrier	17.40489
breed_Silky.Terrier	-16.77498
breed_Flat.Coat.Retriever	14.29413

The bagged trees model had a similar RMSE as the linear regression without regularization and the linear regression with ridge penalty (34.45), a lower MAE of 12.36, and a higher

r-squared at .26. Due to the similar RMSE and MAE of all three models, I would choose the bagged tree model due to the higher r-squared value. The three models are compared in Table 4.

Table 4. *Model Comparison*

Model	RMSE	MAE	Rsqd
Linear Regression	37.65603	16.76878	0.1077589
Linear Regression with Ridge Penalty	37.66347	16.57493	0.1068479
Bagged Trees Model	34.45266	12.36240	0.2590010

Discussion

As Table 3 illustrated, the breed of the dog is highly predictive of the amount of time a dog spends in an animal shelter. Although there are a number of large breed dogs, there is not an obvious category of dog that is more predictive of the length of stay. For example, there are representatives of a variety of AKC dog groups: the non-sporting group (Bulldog, Tibetan Spaniel), the herding group (Collie), hound (Coonhound); sizes: the Terrier and the Tibetan Spaniel being small dogs, and the Retriever and Bulldog being larger dogs, and temperament. This suggests there may be some prejudice or unfamiliarity of some breeds.

The three models performed similarly with respect to RMSE. This was not surprising based on the examples we have worked with in class that have performed similarly as well. The difference between the r-squared values of the bagged trees model to the linear regression models was surprising, even though all values were low.

Conclusion

This wide variety of dog breeds in the top ten category is surprising. I would have thought that the larger dogs and bully-type dogs would occupy the top spots. I was not, however, surprised that breed was a top predictor of the length of stay in the shelter. The breed of the dog is often the best predictor of temperament of a dog, however if the breed listed is only based on a dog's appearance and not genetics or the dog's history, then breed is less powerful of a predictor. Shelters may find they have better success matching shelter dogs with their forever families if they provide a more thorough behavioral evaluation of the dog and assessment of the dog's future needs to potential adopters.