BEASTIE BOYS' DATA MINING PROJECT

April 29, 2019

Cassie Chang
Jace King
Allie Miller
Hanmi Zou

Contents

1 Introduction

Many animal shelters are becoming overcrowded due to the sheer number of stray animals and lost pets coming through their doors. The shelters must also compete with local adoption centers and puppy mills, which tend to appeal more to the public. Due to this variation in popularity and the means by which each group obtains their animals, the general inflow of a shelter tends to be larger than the outflow, and shelters must figure out a way to continuously serve the local human and animal populations in the best way that they can. In this study, we focus on the Austin Animal Center in Texas to analyze the outcomes and the shelter times of the various animals that the shelter served over the course of five years. The Austin Animal Center has more services than a traditional animal shelter has, but still reflects the general animal shelter or center. Our goal was to see which, if any, variables affect an animal's predicted time in the animal center as well as their overall outcome after having been in the animal center.

The Austin Animal Shelter collects and records intake and outcome data for each animal the shelter makes contact with, and this data is made publicly available online. Three total datasets are available for download—one with intake information, one with outcome information, and lastly one with both intake and outcome data. We chose to work with the dataset with complete information about intake and outcome conditions for all animals. One column unique to the separate intake and outcome files was the animal names, but in order to preserve this information we joined this column to our dataset by a unique Animal ID. Other features in the dataset include type of animal (cat, dog, bird, other), breed, condition at intake (e.g. normal, sick, injured), intake condition (stray, euthanasia request, owner surrender), location found, type of outcome (e.g. adoption, return to owner, euthanasia), and length of stay in shelter (days) wth a total of 79,673 observations spanning the years 2013–2018. The majority of the data is thus categorical or simply strings, but some continuous features were also available, like age, date/time found, date/time of outcome, and length of stay in shelter. Thus, our models favor methods which allow for categorical features and predictions.

We performed exploratory analysis, looked for patterns using unsupervised learning, and created supervised learning models. And, as promised, we named all the animals that didn't already have a name.

2 Exploratory Analysis

Before doing any machine learning, we performed some exploratory analysis. This included both a cursory search for trends in the data and geocoding the address data into coordinates to see the geographical density of animal intakes.

By far the most common animal that the center encountered were dogs, followed by cats, though the shelter also dealt with birds and other animals (raccoons, bats, opossums, etc.). Since the shelter is in an urban location, it is no surprise that the main types of animals the shelter deals with are household pets. Most animal intakes were in normal condition, and there were a similar number of sick animals as injured animals. Comparing time spent in the shelter across the different species, we can see that on average cats and dogs tended to spend more time in the shelter (Figure 1). Of course this makes sense considering cats and dogs are more likely to be adopted, and adoption takes time.

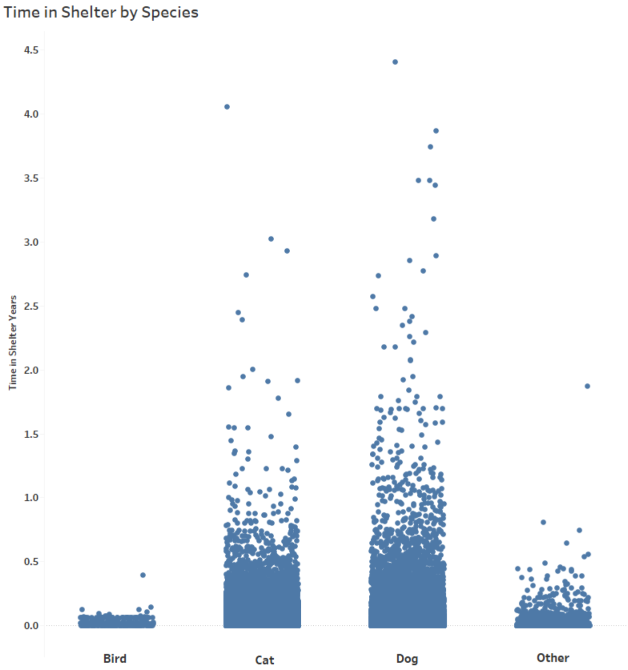
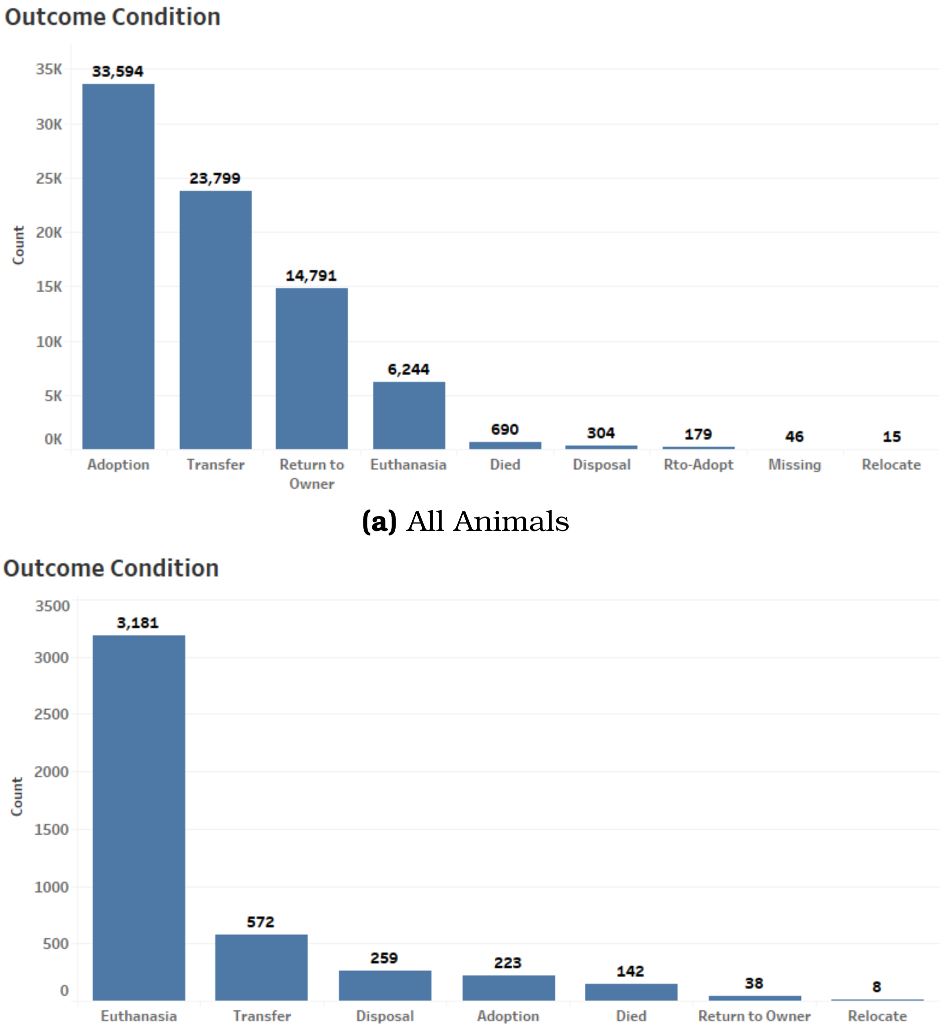


Figure 1: Time in Shelter by Animal Type

We made extensive use of Tableau for our analysis. Figure 2 shows the outcome condition for all animals compared to just others. From this we can see just how different "other" is from animals at large We see that euthanasia is a very common outcome for others animal



(a) All Animals

(b) Filtered on Other

Figure 2: Outcomes

types (which makes sense, considering many are wildlife). Thus, although others contain a wide variety of animals, it is likely that supervised learning models predicting animal type should perform decently well when predicting other.

Through further exploration with the dashboard by filtering we observe that dogs, cats, and birds are all relatively similar, with one key difference being that cats are transferred as often as they are adopted. Thus, we would expect cats, dogs, and birds to be treated similarly in supervised models (although we must note the relative lack of bird observations, with only 0.43% of the observations).

2.1 Correlations

We are also interested in finding possible correlations between features like age (in days or in years) and other numeric features. Finding relationships like this could yield useful information for the shelter. We used `corrplot` to see the correlations between different numeric features (Figure 3).

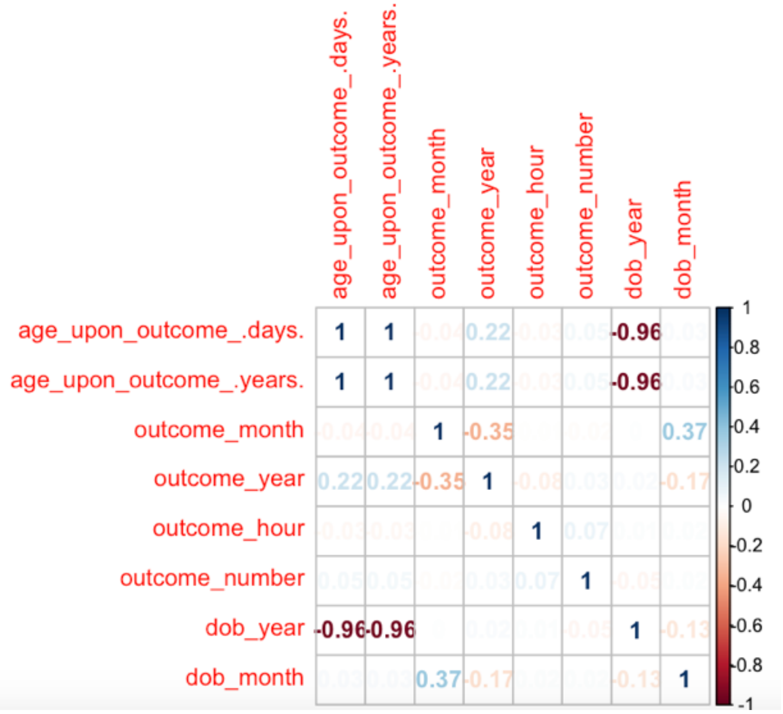
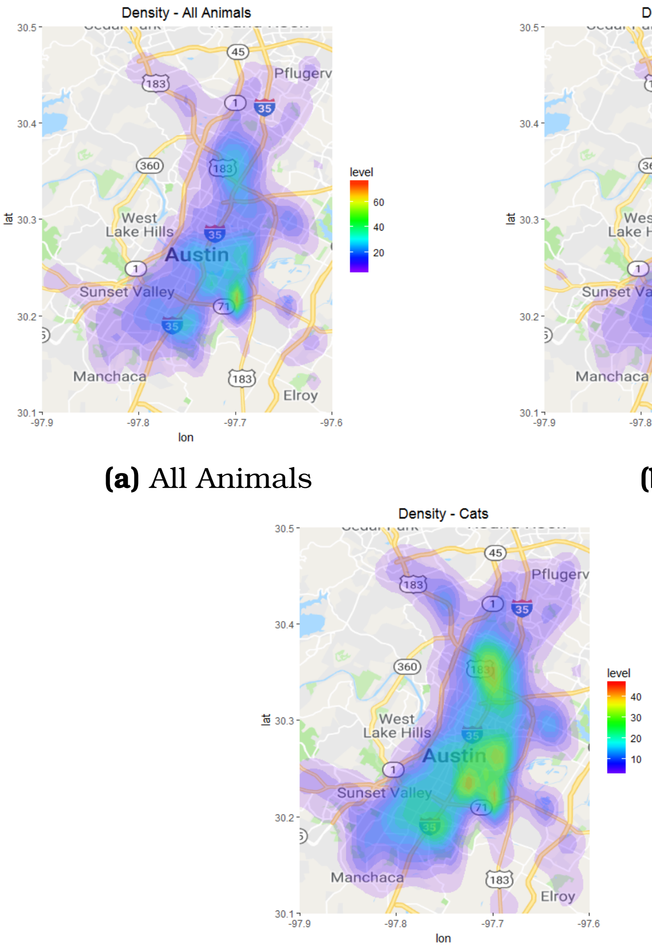


Figure 3: Correlation Plot

Besides the obvious correlations (age in days to age in years, and age to date of birth), we don't see many other strong correlations between the features.

2.2 Geocoding

One of the challenges with this dataset was the number of "categorical" features. Unfortunately, many of these features were actually just strings corresponding to the name, breed, or address. To remedy this, we used `ggmap`'s `geocode` function in conjunction with Google's



(a) All Animals

(b) Dogs

(c) Cats

Figure 4: Geographic Densities for Animal Intakes

geocoding API to convert addresses into latitude and longitude data. While this did not change the results of our models, it did allow us to look at the geographic density of animal intakes. Many of the addresses listed were simply "Austin (TX)," so we removed those observations before plotting. Furthermore, Google could not return the latitude and longitude for a given address with perfect accuracy, so we also removed insensible coordinates before plotting. We looked at all animals (Figure 4a), dogs only (Figure 4b), and cats only (Figure 4c).

3 Unsupervised Learning

Next we performed multi-dimensional scaling (MDS) on the dataset to see if any of trends we noted in our exploratory analysis could be confirmed by an unsupervised model. We performed MDS in R with 2% of the dataset (almost 1600 observations) to reduce processing time. The majority of the columns were included in the analysis, but many of the date/time columns were dropped because of duplicate information. Figure 5 shows the results.

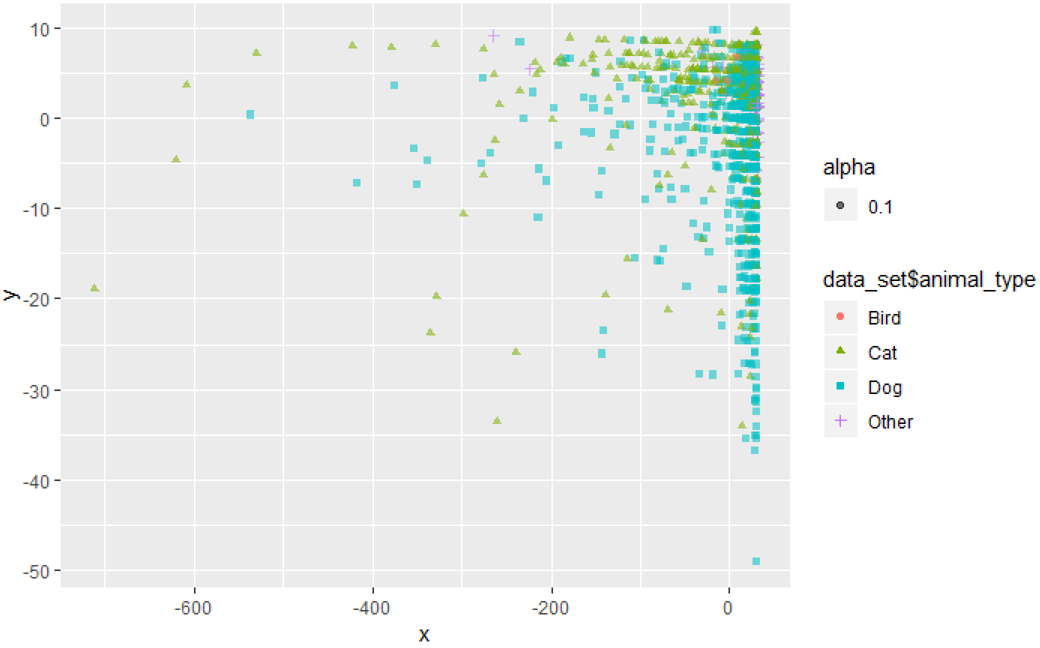


Figure 5: Multi-Dimensional Scaling

The animals are mostly aligned on one of the principal components (x) as demonstrated by the prominent vertical line. However, there is more variation along the second principal component (y), although we can still somewhat see “clouds” of the individual species. For instance, dogs appear to have the most variation overall, while others and birds seem to be more tightly packed in the upper corner. If we instead determine color/shape of the points by the outcome type, we see a similar picture (Figure 6), with adoption and return to owner being the most prominent.

Notably, we see adoption and return to owner clustering in a somewhat similar manner to cats and dogs in the previous plot. And additionally, euthanasia appears to be packed

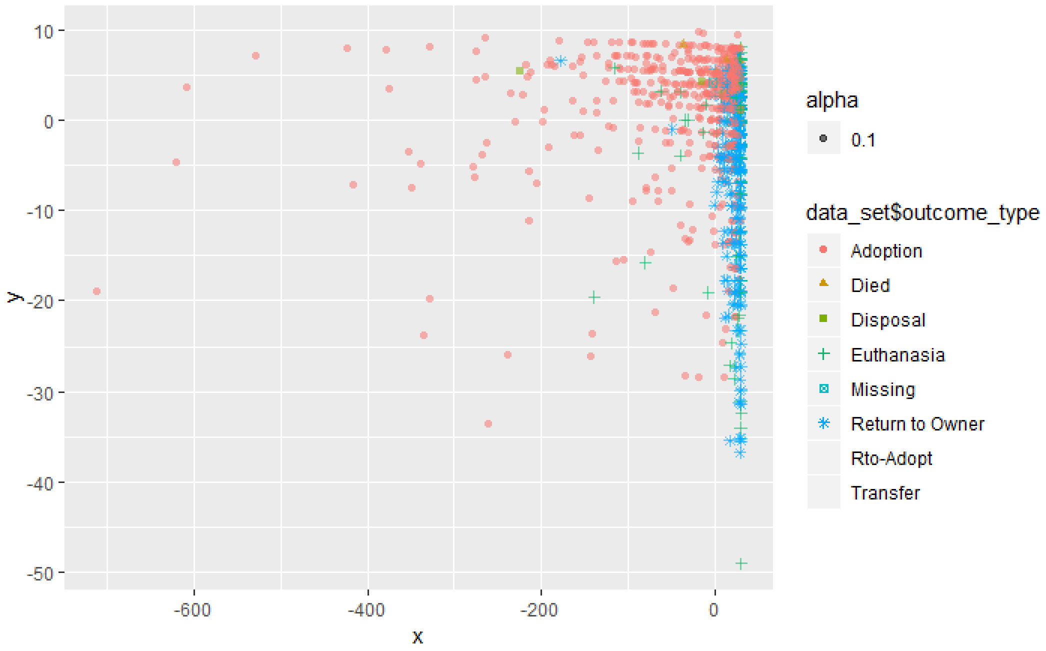


Figure 6: Multi-Dimensional Scaling

in the upper corner much like other was in the previous plot, indicating some correlation between the two, just a we found in our exploratory analysis.

4 Supervised Learning

As previously stated, one of the biggest challenges with this dataset was the number of categorical features. At the same time, however, this gives us plenty of opportunities to perform categorization. We used multiple different models to categorize animal species, animal outcome, and found location. Finally, we also used several models to predict the time in shelter. In each case, we summarize our accuracy for different models and, where applicable, give the baseline accuracy (guess majority class).

4.1 Predicting type

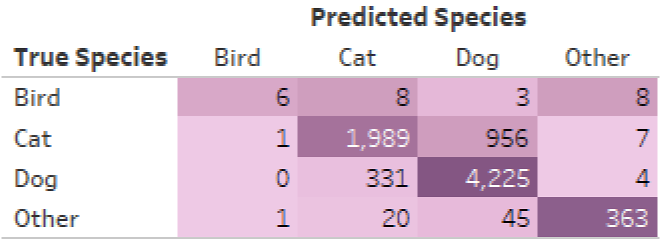
We tried three different models to predict the animal type: random forest, KNN classification, and logistic regression (cats vs dogs).

Model	Test Acc
Majority	0.569
Random Forest	0.826
KNN	0.733
Logistic	0.754

Table 1: Performance of Models Predicting Animal Type

For the random forest model, we used a 90/10 train/test split given the large relative size of the dataset (noting that a 10% test set still tests on around 7000-8000 observations). We considered only a subset of the features for the model (age upon outcome, outcome subtype, outcome type, sex upon outcome, intake condition, intake type, sex upon intake, time in shelter days). We did not include data-time data (redundant) or breed (too strongly predicts type). We initially included other features such as name and color, but these had negligible impact on model results and thus we removed them.

Worth noting in the random forest model is that not all classes performed equally well. The model performed by far the best in classifying dogs, with an within-class success rate of 92.61%, followed by others (84.38%), cats (67.22%), and trailed far behind by birds (24.00%). Figure 7 shows the model's confusion matrix as a heat map, with darker colors corresponding to a larger proportion of the count of predictions for that species (calculated row-wise). Species with a more accurate classification rate will have a distinct, darker cell per row.

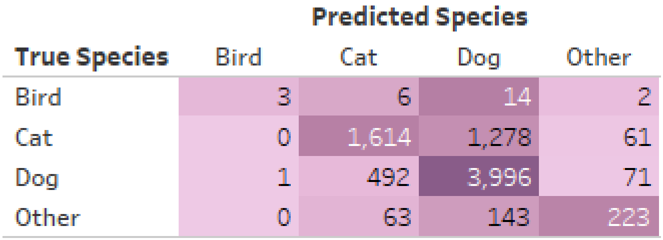


	Predicted Species			
True Species	Bird	Cat	Dog	Other
Bird	6	8	3	8
Cat	1	1,989	956	7
Dog	0	331	4,225	4
Other	1	20	45	363

Figure 7: Random Forest Predicted Type Confusion Matrix

One explanation for the poor result for birds is the relative lack of observations for birds in the dataset, with only 0.43% of the observations being birds. We see the random forest model predict others very well, which we anticipated due to its distinct features. The similarity of features in cats and dogs likely hindered the model's performance with cats.

For our KNN classification model, we included several categorical features by converting to binary columns; however, because this increases the width of the data frame (and thus could slow model processing), we could only reasonably include a few categorical features. Ultimately, we included outcome type, intake condition, and intake type due to their uniqueness and also relatively few categories. Continuous variables included were age upon outcome and time in shelter. After running several models, we chose a value of $n = 15$. The initial result with this model was a test accuracy of 61.27% (only marginally better than predicting the majority class). However, rerunning the model without including any of the categorical features (just including age upon outcome and time in shelter) produces a better result with a test accuracy of 73.25%, a significant improvement. Figure 8 shows the model's confusion matrix.



	Predicted Species			
True Species	Bird	Cat	Dog	Other
Bird	3	6	14	2
Cat	0	1,614	1,278	61
Dog	1	492	3,996	71
Other	0	63	143	223

Figure 8: KNN Predicted Type Confusion Matrix

It is apparent that others performed comparatively worse than in the random forest model. Within-class classification accuracy is then as follows: dogs 87.63%, others 51.98%, cats 54.65%, birds 56.00%. Compared with Random Forest, all of the species performed worse with the exception of birds which performed significantly better. Overall (and unsurprisingly), the random forest model performed better.

For our logistic regression model, we focused only on predicting the two most common animal types, cats and dogs. We used all the continuous features and any categorical feature with a reasonable number of categories (we ignored breed, color, outcome subtype, address, name, etc.). The model performed better than guessing the majority class (0.597), but it could not outclass a random forest model. Notably, the most important feature to the logistic model was age.

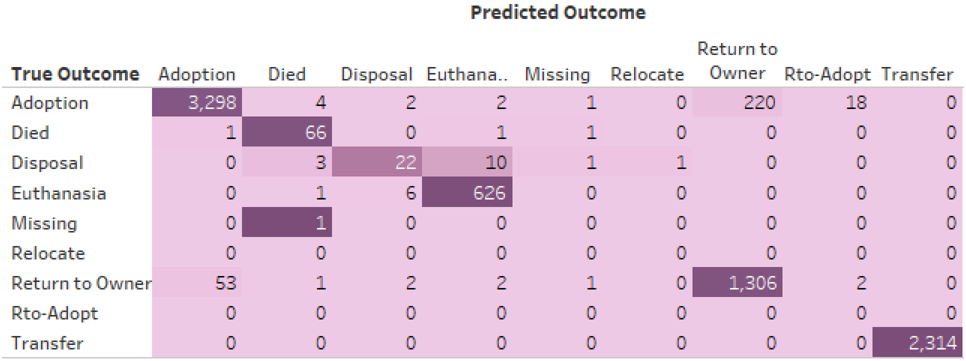
Model	Test Acc
Majority	0.422
Random Forest	0.795
KNN	0.649

Table 2: Performance of Models Predicting Animal Outcome (All)

4.2 Predicting Outcomes

Another, perhaps more important, feature in the dataset is the animal outcomes themselves. This is what we next attempted to predict using several different models. The dataset includes nine different outcome conditions, including adoption, transfer, return to owner, and euthanasia, among others. We first attempted to classify each animal by its outcome into one of these 9 categories.

The first model we used for this was a random forest model, which was essentially the same as that used for species, except we removed the outcome type and outcome subtype features and added the animal type feature. The resulting model had a 79.50% test accuracy rate. Figure 9 shows the model's confusion matrix. The outcomes that occurred very frequently such as return to owner and transfer seemed to perform the best.

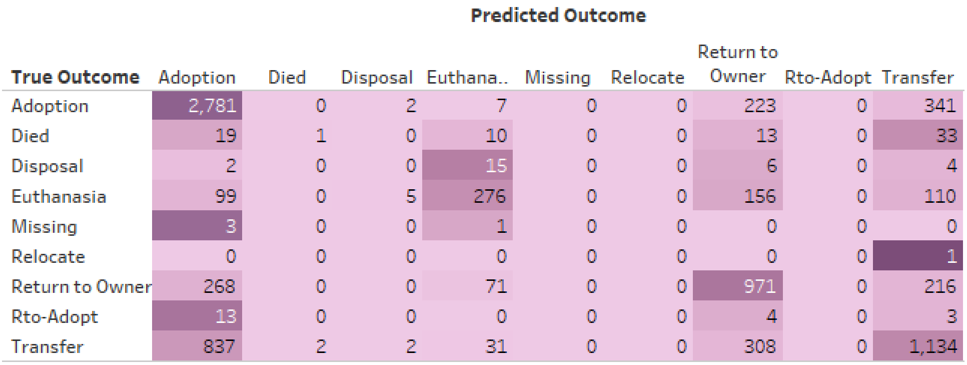


Predicted Outcome

True Outcome	Adoption	Died	Disposal	Euthana..	Missing	Relocate	Return to Owner	Rto-Adopt	Transfer
Adoption	3,298	4	2	2	1	0	220	18	0
Died	1	66	0	1	1	0	0	0	0
Disposal	0	3	22	10	1	1	0	0	0
Euthanasia	0	1	6	626	0	0	0	0	0
Missing	0	1	0	0	0	0	0	0	0
Relocate	0	0	0	0	0	0	0	0	0
Return to Owner	53	1	2	2	1	0	1,306	2	0
Rto-Adopt	0	0	0	0	0	0	0	0	0
Transfer	0	0	0	0	0	0	0	0	2,314

Figure 9: Random Forest Predicted Outcome Confusion Matrix

We next tried a KNN model again similar to that used for species classification, swapping animal type for outcome type. The test accuracy was again worse when including the categorical features, with 53.48% accuracy with these features included and 64.87% with them removed. Figure 10 shows the confusion matrix.



Predicted Outcome

True Outcome	Adoption	Died	Disposal	Euthana..	Missing	Relocate	Return to Owner	Rto-Adopt	Transfer
Adoption	2,781	0	2	7	0	0	223	0	341
Died	19	1	0	10	0	0	13	0	33
Disposal	2	0	0	15	0	0	6	0	4
Euthanasia	99	0	5	276	0	0	156	0	110
Missing	3	0	0	1	0	0	0	0	0
Relocate	0	0	0	0	0	0	0	0	1
Return to Owner	268	0	0	71	0	0	971	0	216
Rto-Adopt	13	0	0	0	0	0	4	0	3
Transfer	837	2	2	31	0	0	308	0	1,134

Figure 10: KNN Predicted Outcome Confusion Matrix

Model	Test Acc
Majority	0.908
Random Forest	0.960
KNN	0.945
Logistic	0.955
SVM	0.938

Table 3: Performance of Models Predicting Animal Outcome (Binary)

It is apparent that the KNN model mostly classified observations in the three most common outcomes: adoption, transfer, and return to owner. This makes sense given that the neighborhood of any given observation probably included observations in these outcome categories.

We also considered a logistic regression, limiting the classification to what seem to be the most critical functions of the shelter, adoption and euthanasia. This model was also very similar to the logistic model for animal type. We predicted with 95.2% accuracy whether an animal would be adopted or euthanised (majority class adopted 0.843). As above, the most important feature for the logistic model was age.

We figured we could get better accuracy by splitting the data into two outcome classes, so we divided the outcome categories into "life" and "death." We considered "life" as the outcomes adoption, relocate, return to owner, rto-adopt, and transfer, and "death" as all other outcomes. Life, the majority class, constituted 90.8% of the outcomes. The accuracies of our models are show in Table 3.

Interesting to note is that with binary classification task, the KNN model performed slightly better when including the categorical features, whereas previously it performed worse. Overall, the random forest, KNN, and logistic regression models performed very well at classifying the animals by “life” or “death” outcomes.

In addition to the models above, we used support vector machines (SVM) on a subset of the data. We only included nine features: income condition, income gender, income age, income type, animal type, and time in shelter. We selected these based on relevancy and correlation to each other. We used only one fourth of the total dataset to further split into our training set (two-thirds of the subset) and a test set (one-third of the subset). The SVM model gave a prediction accuracy of 93.77% using a radial basis kernel on the testing data when considering the binary outcome variable. All of our models tend to miscategorize life outcomes as death outcomes more often than they miscategorize death outcomes as life outcomes.

4.3 Predicting Found Location

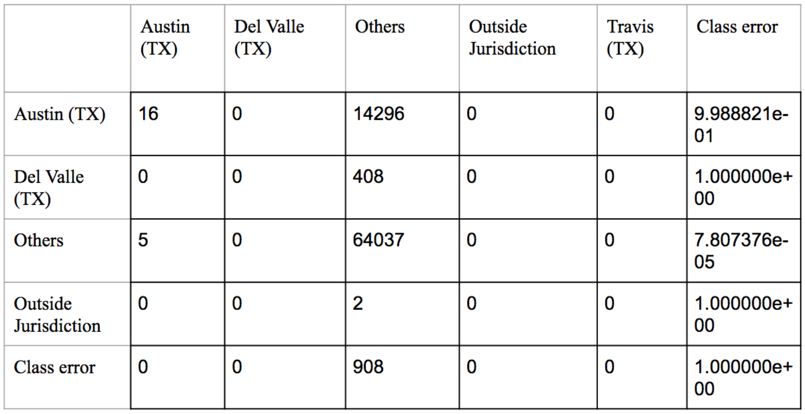
Another interesting question we considered was to determine the popular found location of animals—by accurately identifying these locations where animals are found frequently, shelters could enhance the efficiency as well as the success rate of finding animals. Since this is also a type of classification problem, random forest model was used as it runs efficiently on larger datasets compared to other models such as KNN.

Since a random forest model could only deal with factors with less than 53 levels, we omitted some features like breed and color. In addition, the outcome we are interested in, âĂIJfound locationâĂİ itself has 36576 “levels” (there are many unique addresses), so we first organize database into the desired format. We created a frequency table, classified any location with frequency less than 100 into the “Other” category. We ended up with 11 locations: Austin (TX), Travis (TX), Del Valle (TX), Outside Jurisdiction, 4434 Frontier Trl in Austin (TX), 7201 Levander Loop in Austin (TX), Manor (TX), Pflugerville (TX), 124 W Anderson Ln in Austin (TX), Leander (TX), and Other.

As mentioned above, only a subset of the covariates were used in the model (outcome type, sex upon outcome, intake condition, intake type, animal type) since some variables have more 53 levels. Further, we determined that other variables such as time and data are not strongly correlated with found location and have only a small effect.

The result of the random forest model was an 80.4% test classification accuracy with

19.6% OOB error rate. However, the model is not as good as the classification accuracy suggested. The majority class here was "Other" with 77.2% of observations. Thus, creating this new category might be inappropriate. Yet, almost all locations in the "Other" category have only appeared once in found location. This implies that, if we can capture the distribution of found locations geographically, we can still make meaningful predictions (see Figure 4a. Figure 11 shows this model's confusion matrix and error rate.



	Austin (TX)	Del Valle (TX)	Others	Outside Jurisdiction	Travis (TX)	Class error
Austin (TX)	16	0	14296	0	0	9.988821e-01
Del Valle (TX)	0	0	408	0	0	1.000000e+00
Others	5	0	64037	0	0	7.807376e-05
Outside Jurisdiction	0	0	2	0	0	1.000000e+00
Class error	0	0	908	0	0	1.000000e+00

Figure 11: Random Forest Predicted Found Location Confusion Matrix

4.4 Predicting Time in Shelter

By far the most important information for an animal shelter is how long animals will spend in the shelter. This represents the most expensive aspect of running a shelter—housing and feeding the animals. The overall average time that an animal spent in the Austin Animal Center was 16.75 days. To see if we could predict the amount of time that a particular animal spent in the shelter, we applied several linear regression models as well as a KNN regression model.

The first linear regression model included both income and outcome features. We excluded certain date, time, and string features due to the added complexity of the model. This model was extremely successful in predicting the time in shelter, with an RSS < 0.0001. However, one of the features included in the model was the animal's age upon outcome, which is linearly correlated with age upon intake and time spent in shelter, so this model cheated. In addition, it does not make sense to predict the time an animal would spend in a shelter if we already know the outcome of the animal.

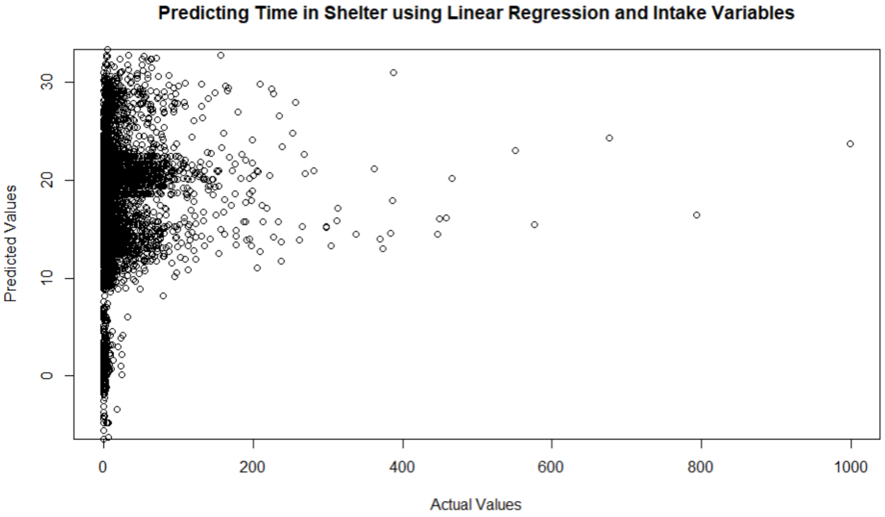


Figure 12: Linear Predicted Time in Shelter (Intake Features)

We created a linear regression model using only the intake features from the previous model, and we found it did significantly worse at predicting the time spent in the shelter, with an adjusted $r^2 = 0.02635$. Backwards stepwise regression limited the variates to animal type, intake type, gender, age at intake, and intake year, but produced similar results. Figure 12 shows the linear model with intake features only.

When using logistic regression to find the general outcome of the animal, the age of the animal appeared to be the most important feature. Although the age did not have a particularly high significance, we still found it to be significant in the previous linear models, and backwards stepwise regression did not exclude it. Based on the previous results, we decided to focus on the effect of the age of the animal on the time spent in shelter. When plotting the age of the animal in days against the time spent in shelter, we see that the relationship is not quite linear (Figure 13), which may explain the poor accuracy in the previous model. We created a linear model using only the age of the animal in days as a comparison, and we also created a KNN regression model with the hopes of reducing the error. As expected, the linear model did rather poorly, but the KNN regression model gave much more reasonable results as it is a more flexible model and the relationship is not linear. Figure 13 shows the linear regression model using only age as a predictor, and Figure 14 shows the KNN model.

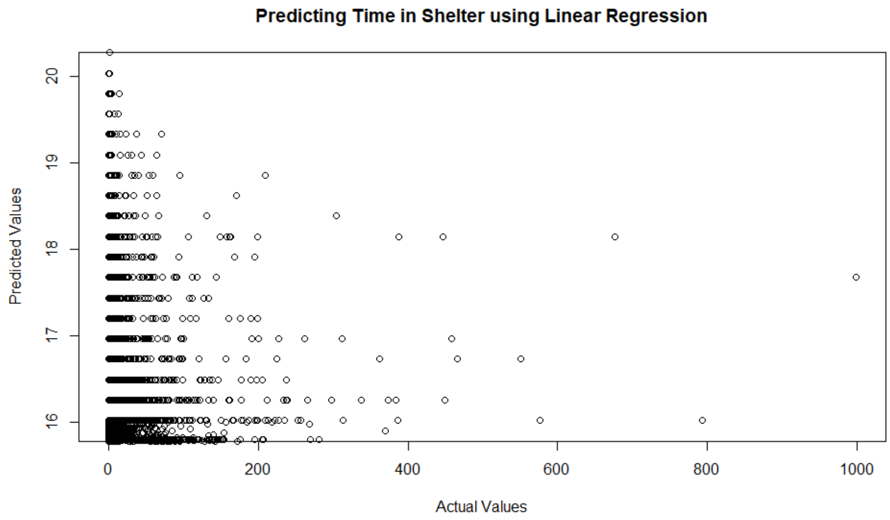


Figure 13: Linear Predicted Time in Shelter (Age Only)

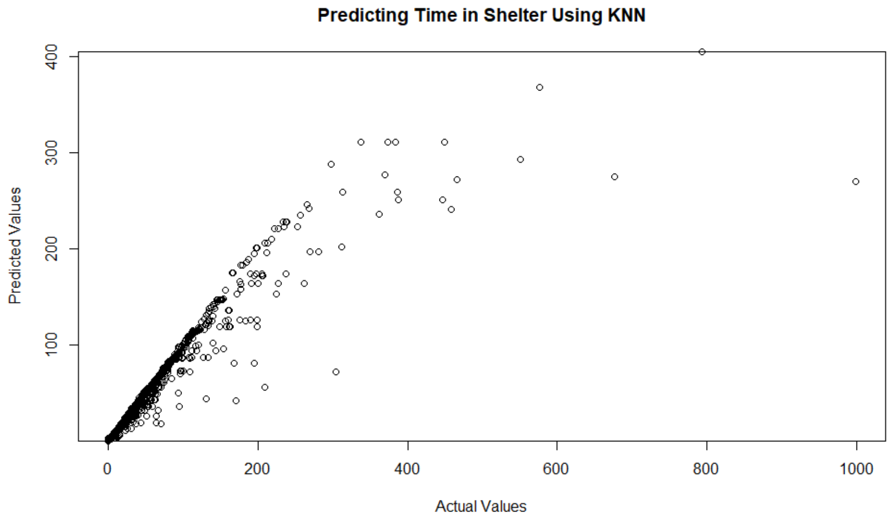


Figure 14: KNN Predicted Time in Shelter

5 An Ode to the Unnamed Animal

There is much power in a name. Unfortunately, some animals just don't have names. To remedy this, we used the over 50,000 named animals to give names to the 24,000 unnamed animals. To be sure, there is no practical benefit in doing this, but our group can now rest easy knowing we gave names to the nameless. Here is a random sample of 25 of the newly named animals:

Benji Button	Cat	Siamese Mix
James	Cat	Domestic Shorthair Mix
Drew Boy	Cat	Domestic Medium Hair Mix
Shenanigan	Dog	Border Collie Mix
Jinger	Cat	Domestic Shorthair Mix
Otter	Dog	Golden Retriever Mix
Kai	Cat	Domestic Shorthair Mix
Woody	Other	Bat
Junior	Other	Raccoon Mix
Major	Dog	Bernese Mountain Dog Mix
Moose	Cat	Himalayan Mix
Ziggy	Dog	Labrador Retriever Mix
Franziska	Dog	Boxer Mix
Pima	Cat	Domestic Shorthair Mix
Yuri	Cat	Domestic Shorthair Mix
Bingley	Cat	Domestic Medium Hair Mix
Ursa	Cat	Domestic Shorthair Mix
Alistair	Cat	Domestic Shorthair Mix
Soxy	Dog	Border Terrier Mix
Kit Kat	Cat	Domestic Shorthair Mix
Salem	Cat	Domestic Shorthair Mix
Alfie	Dog	Rat Terrier/Chihuahua Shorthair
Coconut	Cat	Domestic Shorthair Mix
Dulce	Cat	Domestic Shorthair Mix

6 Conclusion

Our goal in applying data mining techniques to the Austin Animal Center dataset was to see if there were any predictions we could make or patterns we could detect that improve the operations of the animal shelter or increase adoption rates for animals. Multi-dimensional scaling showed that most of the data tended to be similar with a few outliers. We also saw clustering by species and outcome type, as outcome should be relatively distinct for each species. Outcome types should also cluster, as the intake conditions and type of animal are likely to affect the outcome. Sick, injured, or feral animals are more likely to be euthanized while healthy animals, especially common household pet species like cats and dogs are more likely to be adopted or returned to their owners. Heatmaps showing the frequency of animals found in each location can assist the center by concentrating their efforts on the high frequency areas and saving resources by reducing efforts in low frequency areas.

One important thing we noticed as we were aggregating our findings was that many of our models used as many features as possible, including features about outcoming animals in models trying to predict things about outcomin animals. From the animal shelter's perspective, this doesn't makes sense, since the shelter would not have outcoming data on new animals, and the shelter would be most interested in predicting things about new animals. If we had more time, we would like to revise our models to adjust for this realization.

The amount of time that an animal would stay in the shelter can be predicted using only the age of the animal and KNN regression, but it is harder to predict when additional variables such as animal characteristics and intake conditions have an effect. Linear regression models did not capture the relationships between features well at all, as some of the relationships were non-linear. The ability to predict the residency of an animal in a shelter can help the shelter best prepare for each animal. If an animal is likely to have a long residency, the shelter could be able to better care for the animal and reserve spaces for it accordingly.

Predicting the outcome of an animal based on its intake data was much more successful. When the outcomes were split into two categories ("Death" versus "Life" outcomes), random forests, KNN classification, and support vector machines with a radial basis kernel were all able to determine the correct class with a 93% accuracy or higher. The random forest method had the highest classification accuracy of 95% accuracy. Predicting the outcome of an animal could also allow the shelter to best focus its efforts on beginning care procedures for animals, reducing the amount of time each animal would spend in the shelter and thus taking less resources. For more specific outcomes, KNN classification and random forests were applied, with a 65% and 80% classification accuracy, respectively. This is not as accurate as the binary classification models, but would give the animal shelter more information on exactly which procedures to begin, whether it be adoption, attempts to return to owner, or medical care.