Animal Shelter Analysis
Final Report

Katie Ritz
2020
github.com/krritz/Animal-Shelter-Exploratory-Analysis

1. Problem Statement

According to the ASPCA, approximately 6.5 million companion animals enter animal shelters in the US each year. Of these, only approximately half are adopted. This low adoption rate compounds into further issues: of the animals that are not adopted, approximately 1.5 million are euthanized[1]. Euthanization is an extremely controversial topic among the animal advocacy community. When faced with overcrowding and limited funding, shelters are left with the choice of whether to allow overcrowding by keeping animals in warehouse shelters with limited care, euthanization, or turning animals away[2]. As a consequence, one of the top reasons animals are turned away from shelters is due to overcrowding.

While there are likely many contributing factors to shelter overcrowding, including mass farming of animals such as in puppy mills and failure to spay and neuter pets, attempting to **maximize existing shelter funds** to improve adoption rates and therefore reduce shelter time is low-hanging fruit that can be completed by shelters without massive effort including buy-in from the public. This maximization may be completed by identifying high-risk adoptive pets that are less likely to be adopted in a timely fashion or by identifying high-adoption times of the year, week, and/or day. These data may ultimately be used for targeted marketing strategies or for reducing unnecessary effort in placing animals with foster families if they are likely to be adopted quickly. Additionally, identifying patterns in animals that are returned to their owners may provide insights into how to replicate success and provide educational opportunities for animals less likely to be returned to their owners.

[1] Data from the ASPCA
[2] PETA

2. Data Wrangling & Cleaning

2.1 Import and explore dataset

For this analysis, I used the Austin Animal Shelter dataset, which includes intake and outcome data for shelter animals from October 2013 to present. The Austin Animal Shelter is the largest no-kill shelter in the U.S.

1. Import dataset using python's Pandas package
2. Explore the shape, structure, and columns of the dataset.
    a. Each row represents a shelter animal, and the dataset contains many columns with descriptive data on each animal.
    b. I determined during this step that there was already a column that displayed number of days that each animal spent in the shelter; however, it needed to be converted to a datetime value.
3. Split the data into two subanalyses.
   I determined that the shelter times were distributed with an extremely long right tail: 95% of animals were adopted in <71 days, adn 95% of dogs were adopted in <65 days. To answer the question of what affects adoption time in *most* dogs, I used only dogs with adoption times of <65 days. To answer what affects adoption times in dogs that spend a lot of time in the shelter, I looked only at dogs with adoption times of >65 days.

2.2 Adoption times by animal type

4. I used slicing to determine that there were dogs, cats, birds, and 'other' in the dataset, with dog appearing most frequently.
5. I used a boxplot to briefly analyze each group.
    a. The vast majority of animals (75% for all categories) fall into a relatively short adoption time range of less than 20 days.
    b. All groups had a substantial number of outliers beyond the whiskers of the boxplot, making it difficult to interpret this group.
6. I sliced the animals into groups by each animal type.

2.3 Adoption times by dog breed

7. I first looked at dog breeds and determined that the formatting of this column created a lot of complexity due to the presence of mixed breeds. I created a column that returns 1 or 0 if the dog is a mixed breed or pure breed, respectively.
8. I split the breed column into a list of breeds and then used Pandas' get dummies function to return 1 or 0 if the list contained a specific breed.
9. I concatenated the dummies dataframe back to the dogs dataframe.

10. I then used slicing and a melt reshape the table to be long with fewer columns instead of wide, with a new column containing a breed name and a corresponding column with 1 or 0 if the dog contained that breed.
11. I used a violin plot and a boxplot to show adoption times by breed, split by purebred and mixed breed. Both plots were difficult to read due to the number of outliers, and for the boxplot, I hid outliers to get a better grasp on the majority of the data.
12. I found that due to the number of breeds, I had to create another column showing only the most common breeds and lumping the others together as “other”. I chose breeds that appear >400 times as the cutoff.

2.4 Dog adoption times by AKC Dog Group

13. Since there weren’t any clear patterns shown when I looked by breed, I pulled a table that shows breeds by AKC Dog Group in using Pandas and merged this dataframe with the dog breeds dataframe. Once again, notable patterns were difficult to determine with outliers present, so I hid them to see only the majority of the data.
14. I also found that many of the dog breeds were named differently between the shelter dataset. I identified differences and renamed breeds in the AKC dataset to match those in the shelter dataset wherever appropriate.

2.5 Purebred v. Mixed Breed Dogs

15. I looked at all dogs (with outliers hidden) using just purebreds vs mixed breeds and saw only a slight increase overall on adoption times for mixed breeds.

2.6 Adoption time by dog age

16. I also looked at age of dogs using a scatterplot to determine if age affects adoption times, but the age variable was not continuous (rounded by year) and created odd groupings. I decided to instead use age groups as a categorical to look at age as a bar chart.

2.7 Adoption time by coat color

17. I used a very similar analysis to the get dummies for this analysis, as described above.

2.8 Repeat analyses looking only at ‘outliers’

I repeated all the above analyses but sliced using only dogs with the highest 5% of adoption times (>65 days) to see if we could learn more about the outlier groups hidden from the initial analyses.

2.9 Analyses by adoption times

I looked at patterns in adoption times by season/time of year, time of day, and day of week.

3. Initial Findings

3.1 Common Trends in Adoption Times

For this analysis, I looked for trends in adoption times for most animals to make determinations about what the shelter can expect for the average animal.

a. Animal type
Dogs are the most common, but cats have longer shelter tenures.

b. Dog breed
Pit Bulls and Chihuahuas are the most common animal types in this shelter. Labrador Retrievers and German Shepherds are also very common. Of these 4, Labs and Pit Bulls have significantly longer adoption times; Chihuahuas actually have significantly shorter adoption times than the rest of the group.

c. Dog breed group (AKC)
The Working, Herding, Sporting, and Miscellaneous groups have significantly longer adoption times. Of note, the Sporting group includes Labs, which we know to be both more common and have longer adoption times.

d. Mixed breed v. purebred
Mixed breed dogs were found to have significantly longer adoption times than purebred dogs.

e. Dog age
The youngest dog group actually had the longest adoption times; this shelter seems to have pretty rapid turnover, since the average adoption time for dogs is only 9 days. I would speculate that this result is due to the overall composition of dogs in the shelter being young and the overall adoption time being fast.

f. Black dog syndrome
Blue, brown brindle, fawn, and white colored dogs have significantly longer adoption times. Additionally, black dogs did have longer adoption times, but not by a large margin - the mean adoption time was 8.73 days for black dogs and 9.02 for all other coat colors ($p = 0.008$).

g. Coat color
I looked at whether being single colored or multicolored had an effect on adoption time and found that dogs with multicolor coats do have significantly longer adoption times.

3.2 Trends for animals with long adoption times

For this analysis, I looked at animals with the longest adoption times (the top 5th percentile). The goal is to identify trends in these animals that are more likely to have long adoption times.

a. **Animal type**
   In this extreme group, dogs had significantly longer adoption times and cats had significantly shorter adoption times. Both cats and dogs were nearly equally well-represented in this group.

b. **Dog breed**
   In this group, only American Bulldogs had significantly longer adoption times, with a mean adoption time of 200 days. Additionally, Pit Bulls were by far the most common breed in this group at over 30%.

c. **Dog breed group (AKC)**
   There weren't extreme trends in breed groups for this analysis, although the Hound group did have slightly longer adoption times (mean 180 days, $p=0.03$).

d. **Mixed breed v. purebreed**
   There was no effect of mixed breed v. purebred on adoption times in this extreme group.

e. **Dog age**
   In a reversal from the majority, young dogs had significantly shorter adoption times and dogs 2.5-5 years and senior-aged dogs (15 years+) had significantly longer adoption times.

f. **Black dog syndrome**
   Again, there was no support for black dog syndrome at this shelter. Brown Brindle dogs were the only coat color with a significantly different adoption time: they had a shorter adoption time.

g. **Coat color**
   There was no significant effect of multicolor v single color coats on adoption times in this group.

3.3 Temporal patterns in adoption times

Here, I looked at patterns in when animals are adopted to gain insight into optimal times for the shelter.

a. **Season**

Overall, there was a clear seasonal trend in adoptions, with more adoptions during the summer. Interestingly, when I parsed out dogs and cats, there was no seasonal trend for dogs, but we did see a seasonal trend for cats.

b. Time of day

There was a clear peak in adoption times flanking a standard business day - between 8-10 AM and 4-6 PM. This trend was more pronounced in dogs than cats, which had a pretty steady adoption time throughout the day.

c. Day of week

I looked at average adoptions per day and found that weekends had significantly higher adoption rates, while late-week days had significantly lower adoption rates (Wednesday-Friday).

4. Modeling of Shelter Times

4.1 Modeling Overview

The goal of this project was to gain a better understanding of what makes a specific animal more likely to have a longer adoption time and propose strategies for resource optimization for a shelter. Throughout this analysis, I have focused specifically on dogs, so my goal here was to develop a model to determine time in shelter based on available data at intake.

I tested numerous regression models and was able to achieve a maximum R^2 value of XXX% and therefore only explain about XXXth of the variance in shelter times. However, I think this model could still be useful for shelters in gaining a general idea of animals at a high risk of having a longer adoption time upon intake.

4.2 Data Preparation

To prepare the data for building and optimizing a model, I used all data that would be available at animal intake and transformed it as needed.

1. Breed (categorical). To transform this variable to a usable, numerical form, I used the get dummies function to create a binary column for each breed.
2. Purebreed v. mixed breed (binary). I created a binary column that returns 1 or 0 if the animal is mixed breed or purebreed, respectively.
3. Intake condition (categorical). I used the same transformation as I used for breed for this column.
4. Intake type (categorical). I used the same transformation as I used for breed for this column.
5. Intake sex (categorical). I used the same transformation as I used for breed for this column.
6. Coat color (categorical). I used the same transformation as I used for breed for this column.
7. Multicolor v. single color coat (binary). I created a binary column that returns 1 or 0 if the animal is multicolor or single colored, respectively.
8. Age at intake (days). Although this column is presented in days, it is actually years -

Additionally, time in shelter is exponentially distributed, so I used a log transformation of that for the model's target.

4.3 Ordinary Least Squares

I used a linear regression using Ordinary Least Squares to develop a model to predict shelter times. This method was not particularly effective, and I was only able to obtain a model with an R^2 of only 10.86%. I used both scikit-learn and statsmodels to try to optimize the regression.

Feature Selection

For feature selection, I searched for the model that would explain the most variance in the data (highest R^2). I examined root mean squared error as well (standard deviation of the residuals) to select the optimal model.

First, I examined the top correlated variables to see if R^2 would improve. I looped through all possible numbers n of top correlated variables and determined that the highest R^2 value, 10.67%, was obtained with the top 114 correlated features. The root mean squared error was relatively high, at ~1.7; however, this model could still be useful in getting a general estimate of shelter time for a given dog.

Examination of Residuals

I examined the distribution of residuals as well as comparing the predicted and observed y values. They are mostly normally distributed, although they are skewed right. There are also lines in the distribution, which I would hypothesize is a result of the data being highly binary or shelter times being limited at 0 days (recall the plot below is showing the log of shelter time).

4.4 Ridge Regression

I created another model using Ridge Regression. I thought this model would be particularly appropriate for this dataset since there is so much variation and not a lot of available data: by introducing a bias, we might see an improvement in ability to predict new data instead of overfitting to the available data.

I included all features in this model since Ridge regression will shrink features that do not contribute to the model's predictive power.

I found this model performed similarly well to OLS, with a maximum R^2 of 10.55% and an rmse of 1.7.

Parameter Optimization

scikit-learn has a built-in parameter optimization for alpha for the Ridge regression model, so I used that to optimize alpha.

Examination of Residuals

Again, I examined the distribution of residuals as well as comparing the predicted and observed y values. The distribution and patterns were nearly identical to those for the linear regression, although the predicted and observed values appear (slightly) more linear than they do for the linear regression.

4.5 Support Vector Regression

As a result of the data being sparse with many features, I chose to use a Support Vector Regression (SVR). SVM models are known to be effective in high dimensional spaces. Additionally, only certain training data are used in the decision function, which I hoped would help with the stochastic nature of this dataset.

I continued to use R^2 to evaluate the efficacy of this model, in addition to RMSE. After scaling the data and optimizing parameters, I was able to obtain an R^2 of 14.06 and an RMSE of ~0.94, the best values yet.

Data Preprocessing

I used scikit-learn's preprocessor to scale the data, as recommended by the documentation, because SVM algorithms are not scale invariant.

Of note, one of my goals is to develop a model that can be used on new data. This requires that new data be scaled using the StandardScalar and that results be “inverse transformed” to get an interpretable outcome.

Feature Selection

Because SVR is more computationally expensive, I only examined the performance of the model on the top 114 features compared to all features. I found that the model performed best using the top 114 features.

Parameter Optimization

I used GridSearchCV to test values of C and epsilon from 0.01 to 100. The results found that the default values of 1.0 and 0.1, respectively, were the optimal parameters.

Examination of Residuals

The distribution of residuals looked more normal for this model, although the lines in the scatterplot were still present. The distribution of residuals was also shifted closer to 0, and the relationship between predicted and actual y values appeared (slightly) more linear than that for Ridge regression, another slight improvement.

4.6 Ensemble Methods

After working to optimize several different models, I decided to try a few ensemble methods to see if combining model results would improve overall predictive power and model performance.

Averaging

I used averaging of the OLS, Ridge, and SVR models (and combinations of the three) to see if the predictive power could be improved. I found that using only Ridge and SVR resulted in a slightly higher R^2 but also a higher error rate.

None of the models seemed extraordinarily better than the SVR model on its own; OLS and SVR together performed moderately.

Gradient Boosting Regression

I also included a gradient boosting model to see if this strategy could help improve the model. Boosting is the development of many sequential models, where each model attempts to correct errors made in the previous model. The final model is the weighted mean of many weaker models. Because all of my model attempts have been mediocre, I hoped that this type of ensemble method could help improve predictive power. Gradient Boosting is highly resistant to overfitting, so I used 1,000 estimators.

This model had a much higher R^2 value of 18.26% and a similar RMSE to those in previous models (~1.63).

Examination of Residuals

I examined the residuals for the gradient boosting model and found they looked (again) highly similar to those for the other models - somewhat normally distributed but skewed right with some lines. The predicted v. acutal values looked best of all the models and somewhat linear for a greater range.

4.7 Model Testing

I created several test animals to predict their shelter times and demonstrate the potential use of these models.

4.8 Challenges

One of the main challenges with developing a model with this data was that the vast majority of the available data were categorical; age at intake was the only numerical variable available, and even that was not purely linear: it appears that the ages are age in integer years or half-years multiplied by 365 to generate a loose estimate of age in days. I created dummy columns using categorical data, e.g., for each breed, there is a column that returns 1 if that dog is that breed or 0 if not. This resulted in a sparse, high-dimensional dataset.

Additionally, SVR was highly computationally expensive and very slow to run. It made it difficult to optimize and test different options.

Finally, I believe this data has a lot of unknowns, including why people adopt certain animals. Most animals at this shelter are adopted within a week, which does not really lend itself to a lot of predictable variation. In the end, I think these models could be useful for indicating a trend for adoption times but are likely not dependable for operational decisions.

5. Final Recommendations

1 Adoptions aren't random.

These analyses showed that most information available at animal intake impacts adoption time, including animal type, breed, color, age, intake condition, etc. Using modeling to determine what animals are at a higher risk of having a longer adoption time can be used for proactive advertising and outreach. The following recommendations could be made based on these findings:

- Labs and Pit Bulls are some of the most common dog breeds and also have significantly longer adoption times: targeted efforts for these breeds could be beneficial. Pit Bulls and American Bulldogs are more likely to have extraordinarily long adoption times.
- Mixed breeds are more likely to have long adoption times; however, purebred dogs are known to have more health issues due to poor breeding practices[3]. Perhaps some targeted messaging here could benefit this group.
- High-energy dogs have longer adoption times, such as those in the Working, Herding, Sporting, and Miscellaneous groups. Support for adopting a high-energy dog, such as recommendations for exercise and stimulation, might be helpful here.

2 Adoption times can be leveraged for shelter operations.

I found that there are effects of season, time of day, and day of week on adoption times. For this shelter in particular, the following recommendations might be utilized to maximize business hours and events:

- Adoption events are likely to be more productive during the summer for *cats*, but not dogs. In general, cats have longer shelter times than dogs, too, so some cat-specific events might be rewarding for adoption rates.
- A split business hours schedule of morning and evening operations to facilitate adoptions flanking the standard 9-5 business day could optimize employee time.
- If needed, reduced hours Wednesday/Thursday are least likely to disrupt current adoption patterns.

[3] *Although Purebred Dogs Can Be Best in Show, Are They Worst in Health?* by Claire Maldarelli, Scienceline; February 21, 2014