Animal Shelter Adoption Rates Analysis

Final Report

Henry McGee, Joshua Gates, Malaya Reece

Data Science

CPSC 4300

Fall 2019

Introduction

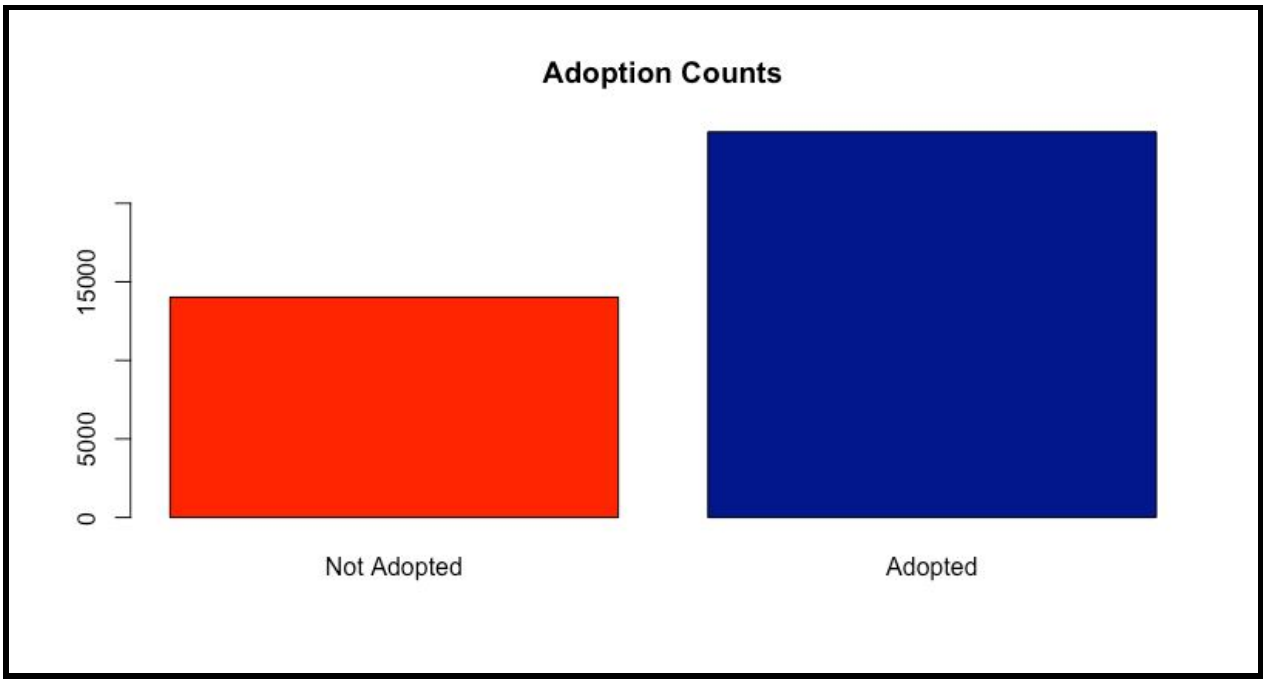
Each year, approximately 6.5 million animals enter shelter facilities nationwide, and approximately 1.5 million of them are euthanized. For many of them, the reason is that there is simply not enough room, and not enough resources to care for them, so it is critical that we continue to maximize adoptions as much as possible. Our project sought to determine what factors were most influential in determining adoption outcomes. To accomplish this goal, we analyzed a data set containing all of the animal intake data from the country's largest no-kill shelter, the Austin Animal Center. This set contained over 109,000 observations, covering a time span of about 6 years. By analyzing this data and developing prediction models, we hope to help shelters develop more strategies to place more animals into permanent homes.

Exploratory Data Analysis

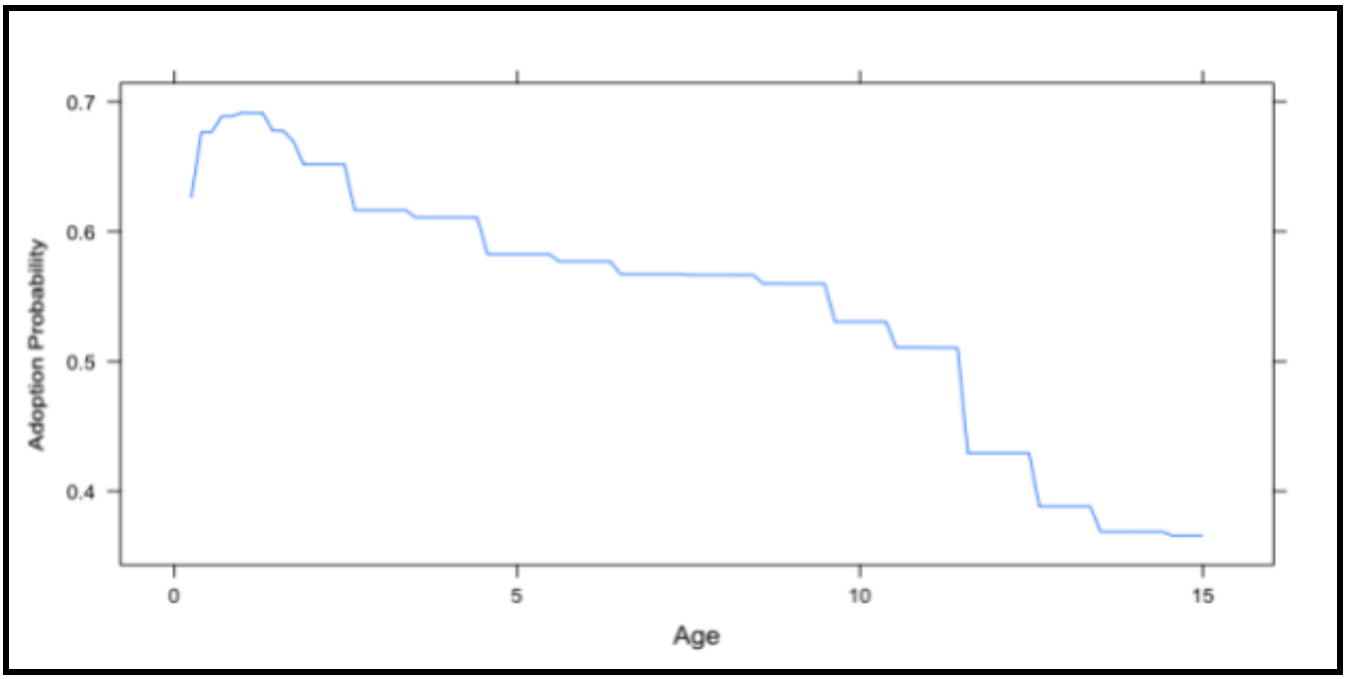
The data set spans a time period from October 1, 2013 through September 30, 2019, covering 6 years of observations. It includes several species of animals including dogs, cats, birds, and livestock. Every row in this set represented one animal that arrived to the shelter. For our particular case, we decided to only look at dogs, specifically, as they made up the majority of animals in the set.

Our unit of analysis was the adoption outcome, or whether or not an animal was adopted. We chose to study adoption since it allows us to compare how our predictors affect adoption rate, thus giving animal shelters an overview of which animals are most and least likely to be adopted. To evaluate this data, we created a function that separated animals into a binary classification, 1 for adopted and 0 for not adopted. To clean this predictor properly all animals that were not explicitly adopted were included in the not adopted section. We also removed runaway pets that were brought into the shelter but returned to their owners because they don't represent an actual adoption.

In the original data set of all animals, there were 109,682 observations. We only wanted to keep track of animals with unique IDs, however, so we refined the data to a set of 98,307 observations. Since we were only focusing on dogs, we further refined the set which narrowed the number of observations to 52,541. Now that we're down to just unique dogs, we visualized our main predictor with a bar graph since it's a binary outcome.



At the onset of our project, we thought the most important factor in predicting an animal's adoption would be its age. This seemed intuitive to us as most people desire puppies as opposed to older dogs. With that assumption, we plotted the adoption probability against age in the graph below. This regression line appears to be a good fit and shows a clear relationship between the single predictor and the outcome.



We also had to clean our data differently for the specific models we employed. For the first model, we selected a linear regression model. In order to do this, we narrowed our data to the most populous 24 breeds. This left us with 36,439 observations for our linear regression model. We capped the number of breed types at 24 because they make us roughly 50% of the observations and having any more breeds would make the linear regression model too complex.

For the second model we chose to employ bagging and boosting on a random forest model. In order to make use of this model we were limited to 32 features for each predictor, as the bagging process could not handle more due to technical limitations. This included Breed, Age, and Color. For Age we grouped with increments of .25 until age 2, then switched to whole numbers to an age limit of 15, and finally multiples of 2.5 years up to 30. 30 years of age was chosen as the max because the oldest living dog on record lived to be 29.5 years old. Most of our observations were located under age 2, which is why we used .25 increments only to age 2. Limiting our observations for Breed and Color was simple, since the count method in R allowed us to rank the features by number of appearances, and then we chose the top 32 for both of these predictors. After this pruning, we were left with 22,008 observations.

One last note is that our decision to use the most populated levels was to ensure we had a good pool of data to draw from. This was to try and avoid overfitting on a sample that was a rare occurrence.

Summary of Machine Learning Models

When considering the data, we thought the most important factor in predicting an animal's adoption would be its age. This seemed intuitive to us as most people desire puppies as opposed to older dogs. Thus, a multiple linear regression was chosen, firstly for its simplicity and ease of viewing, and secondly because it would allow us to show the relationship of the two predictors to age easily as the slope of the line would reflect adoption rate, the x-axis would be age, and each regression line would be a different breed. This data could be nicely presented on a multiple line graph.

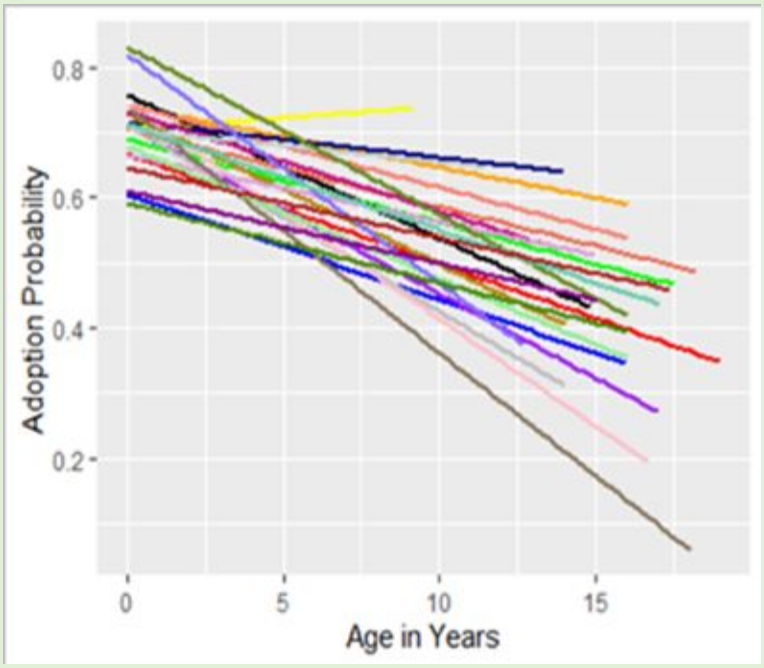
This, unfortunately, was a mistake on our part. The main problem with this approach was that we could not accurately test our model against real data since we gave a probability of an outcome but the actual outcome we were trying to predict was a binary between adopted or not. As we misunderstood this basic incompatibility, we tried to check its error rate based on an RSS of the probability instead of constructing a confusion matrix and grouping the probabilities into binaries based on a cutoff. We corrected this problem with our second model, however.

But to fully communicate our results, we did calculate this false error rate by using a K-fold method. Our K in this case ended up being 24 as we ended up checking each breeds adoption rate versus the rest of the set. This gave us an error rate of .1478, or about 15%.

However, as stated before, this was an incorrect error rate, based on a misunderstanding of the nature of the outcome we were trying to predict.

With this misunderstanding, although our model may represent how likely a specific dog is to be adopted, it does not measure the accuracy of the problem type and therefore does not fit this data very well.

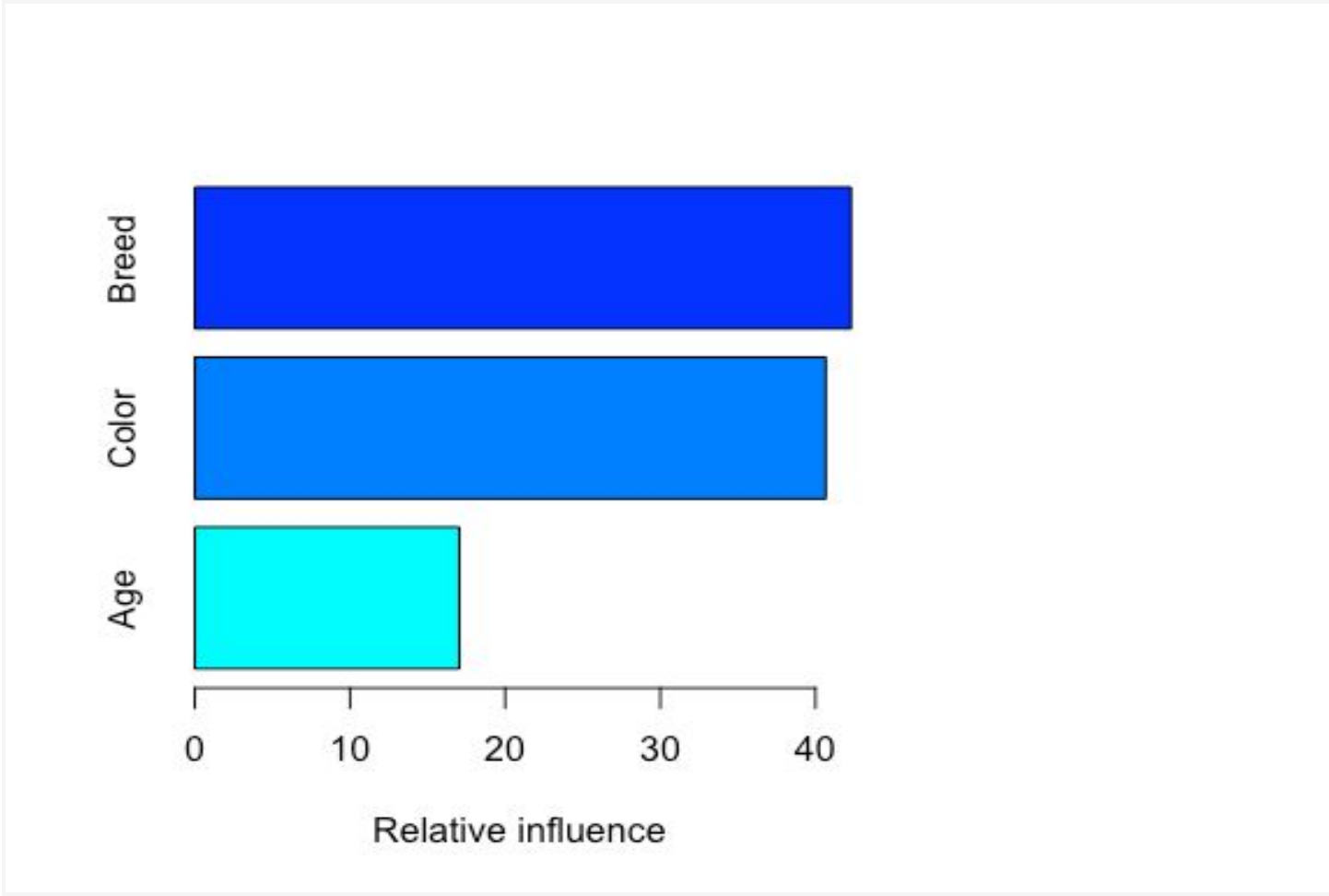
Legend:
Pit Bull Mix = blue
Chihuahua Shorthair Mix = red
Labrador Retriever Mix= green
Anatol Shepherd Mix = yellow
Australian Cattle Dog Mix = orange
Australian Shepherd Mix = purple
Border Collie Mix = black
Boxer Mix = gray
Cairn Terrier Mix = salmon
Catahoula Mix = navy
Chihuahua Longhair Mix = coral
Dachshund Mix = pink
German Shepherd = darkgoldenrod
Jack Russell Terrier Mix = medium violet red
Miniature Poodle Mix = plum
Miniature Schnauzer Mix = navajo white
Pointer Mix = azure
Rat Terrier Mix = light green
Siberian Husky Mix = light blue
Staffordshire Mix = olive drab
Yorkshire Terrier Mix = dark magenta
German Shepherd Mix = aquamarine
Labrador Retriever = chartreuse
Chihuahua Shorthair = firebrick



For our second model, we wanted to challenge our assumptions and decided to evaluate the importance of each predictor in relation to adoption outcomes, our core measurement. We started the process with a bagged tree model. After a lot of tweaking and testing, however, we ended up using a boosted model. We also now understood that due to the binary nature of our outcome, our results needed to be conveyed via a confusion matrix. But to actually understand how our predictors were affecting the data, we decided to use a variable importance plot to

demonstrate the importance of each predictor. This would be the best means of accomplishing this goal, because it ranks each predictor in terms of its impact on the outcome.

We tried many different values for this model's parameters; such as tree number, interaction depth, and shrinkage. After a lot of tinkering we decided on 10,000 trees, a depth of 49, and 0.0001 shrinkage value. Since the function used for boosting has a specific option to predict a binary classification, this model was perfect for us because we have a binary outcome of adopted or not. With this model, we also plotted the variable importance graph and generated a confusion matrix. Calculating the accuracy from our matrix, this model ended up being correct 63% of the time.



Overall, we found that this model fit the data relatively well. That's not to say it is perfect of course, as an accuracy rate of only 63% will surely tell you. One of the problems was that although dogs one year old or less held a large percentage of the overall adopted population, they also held the highest adoption rates. Due to these young dogs being so numerous and so often adopted, and older dogs in general making up a much smaller section of the data set, the model may have underestimated age's overall importance in determining adoption rate.

Finally, in regards to our confusion matrix, we were a bit disappointed. Regardless of how many times we ran the model and the tweaks we made to the prediction, we always ended up with an overwhelming probability of the dogs being adopted which led to a large number of false positives. We believe this may be a result of the data we have being mostly adoptions. Since the shelter is no-kill, most dogs have a higher chance of being adopted here than at other shelters because the length of time they stay is not an issue. Also since most of the data is comprised of young dogs, and this subset has such a high adoption rate, the random nature of boosting could have skewed our model to estimating higher adoption rates than may be suitable. Given these factors, however, we do think this is a relatively good model for this data set, and is certainly a better model than our first.

Confusion Matrix

OutcomeType	No	Yes	Sum
0	465	7519	7984
1	597	13427	14024
Sum	1062	20946	22008

Test Cases

Since boosting is a rather black box model, we have no direct data to plug our test cases into. But we can try to average the effect of each factor against adoption probability and calculate their chances of being adopted according to our mode in that way. Our test cases for this model are three dogs, each varying in some way for a control case dog.This can show the calculated difference in our model.

Control: Yellow, 5-year-old, Yorkshire Terrier Mix
Probability of Adoption: 60%

This particular case was selected because it represents an average rate for all three predictors.

Age Case: Yellow, 1-year-old, Yorkshire Terrier Mix
Probability of Adoption: 62%

This case shows a one-year-old, one of the highest adopted ages, and the gain of 2% probability reflected from that. This trend holds true and can help inform shelters to make decisions such as purchasing more beds for larger dogs than for puppies because of the high adoption rate younger dogs enjoy.

Color Case: White, 5-year-old, Yorkshire Terrier Mix
Probability of Adoption: 64%

This case illustrates the large impact color has on dog selection. Shelters can use this to their advantage by putting the less popular colors near the front, helping them to be seen first and more likely to be selected and by possibly sprinkling low numbers of the less popular colors among the more popular colors to make them stand out and attract potential adopter's attention.

Breed Case: Yellow, 5-year-old, Border Collie Mix
Probability of Adoption: 63.67%

Finally, this case shows the effect breed has on adoption rate. It is similar to color and the effect is stronger or weaker than color depending on the two colors or breeds in question. This particular predictor is very helpful for the same reasons as both of the predictors above. Different breeds have different sizes, so the bed logistics can be aided in the same way, and placing those breeds in strategic locations could also potentially boost their adoption rate.

Summary and Conclusion

When reflecting on the project as a whole, our team learned a lot; both about data science, and the particular problem we looked into. Our initial goal was to determine which factors led to an animal being adopted, and I believe we can answer that with some reliability. In contrast to our initial thoughts, we now believe aesthetics may be the primary factor that affects a dog being adopted. This would explain why breed and color of a dog had a much larger effect on the adoption outcome than did its age.

We also learned that fitting a model is a little more work than we initially thought. With the mistakes we made with our first model to all the fine tuning we did on our second, we definitely have a new appreciation for how delicate you have to be when approaching a data science problem. We also learned to make sure of a model's ability to deal with a qualitative versus quantitative outcome. This problem used mostly qualitative data and a binary outcome. This led to the problem with our first model were we treated the outcome as quantitative. It was definitely a valuable lesson for any data science we do going forward.

We can also give shelters a more reliable prediction of whether a dog will be adopted or not. Perhaps more important than that, however, is how these adoption probabilities can play into arming the shelters with the knowledge they need to plan for what resources they will need and perhaps how to increase a certain dog's chances of being adopted. Ultimately, we hope that this analysis could aid shelters in preventing euthanizations that would have otherwise occurred.

Finally, we wanted to talk about what could improve this project given more time and resources. The first thing that comes to mind is definitely data. Ideally, we would be able to gather more data from multiple shelters including one's that were not non-kill shelters. This could let us construct a model that could rule out variance on location and non-kill status. Further, if there was a way to somehow poll potential pet adopter's opinion on how cute they found the dog they were adopting versus those that they did not, we could make a more definitive statement regarding aesthetics with regard to adoption outcome, if only based on subjective feelings. Finally, we would have liked to develop separate models based on different locations to see if different parts of the country adopted different types of dogs. This could allow shelters to send certain types of dogs between shelters to maximize their adoption rate.