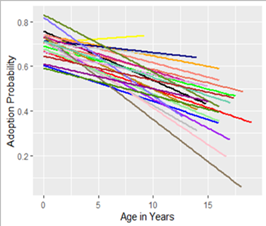
# **Variable Importance(Model 2)**

##### **Henry McGee, Joshua Gates, Malaya Reece**

DataScience

CPSC 4300



***Previous Work***

In the original model we wanted to show the relationship between age, breed, and adoption rate. We developed a multiple linear regression model to demonstrate how a dog’s breed and age affect their likelihood of being adopted.

This model showcased how adoption probability decreases with age. We found that this trend was similar across all breeds with the exception of one outlier, which we discussed in deeper detail in that report. The goal in utilizing linear regression was to measure the adoption rate across multiple breeds and be able to compare their slopes, thus giving an estimate of how age’s importance varies among breeds.

***Goals***

For our alternative 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. However, due to the nature of bagging, our results this time will need 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.

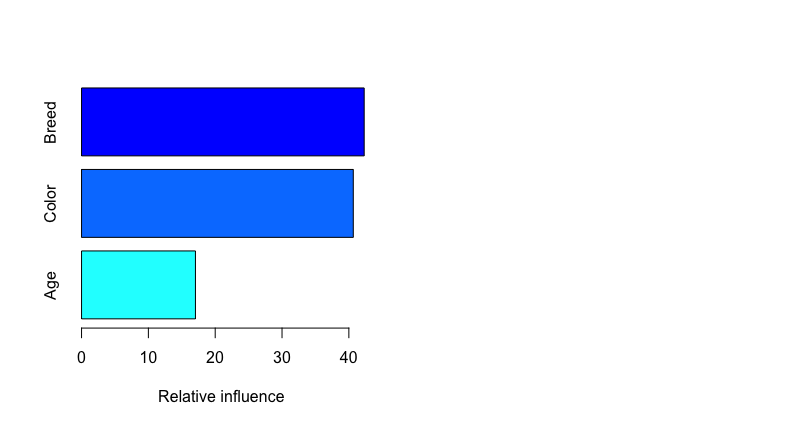
This model will hopefully be able to provide shelters with a solid prediction of which dogs will be adopted and which they should plan to take care of long term. This could save time, energy, and resources and reduce waste from assigning resources to a dog that will most likely not be with the shelter for very long.

***Data Cleaning***

Because of computer processing power limitations in bagging we decided to refine our data by limiting the predictors to the most populated 32 levels. This 32-level cutoff was done to be able to use the tree building functions themselves. Our decision to use the most populated levels was to ensure we had a good pool of data to draw from to try and avoid overfitting on a sample that was a rare occurrence. The three predictors in our data set we ended up being able to use were breed, age, and color.

***Model Creation***

After running through several preliminary models, we finally settled on Boosting. We tried many different values for the model, 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.



***Findings***

We found several interesting results in this project. One was Age seemed to not matter nearly as much as we had previously thought. In fact, our analysis showed the opposite. Breed and color actually have much more of an effect. We ran our models several times and these two factors kept switching back and forth (a function of the randomness involved) which indicates they are relatively equal in regards to affecting the adoption rate whereas age has a very low effect comparatively.

We do wish to note, however, that the breed may have slightly less effect than is shown in the model because of the Shih Tzu Mix Breed which was an outlier for the overall breed distribution.

However, given that some breeds are simply unpopular, this outlier should be considered in the scope of this project especially when put in the perspective of looking at data for running a no-kill shelter.

We noticed 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 young, 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.

**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.