Data Mining & Statistical Learning

November 27th, 2022

Course Project Proposal

Characterizing Factors affecting Animal Outcomes from the Austin Animal Center

Team Members

Sofia Laval **- ID:** 903299713 **Email:** slaval3@gatech.edu

Justin Olson **- ID:** 903558598 **Email:** jolson49@gatech.edu

Monica Singh **- ID:** 903655580 **Email:** msingh341@gatech.edu

Abstract

Over a million animals are brought into animal shelters a year resulting in overcrowded shelters. It's important to discover which factors could contribute to the outcome of an animal getting adopted or euthanized. Our team sought out to evaluate an animal shelter's animal outcome dataset and to discover which factors contribute to the fate of an animal. Utilizing popular ensemble techniques such as Random Forest and Boosting we found that some features such as the age of an animal and whether they are neutered or spayed can contribute to the likelihood that an animal gets adopted. We hope our observations can help animal shelters identify animals who are at risk of not being adopted and take steps to improve their chances.

Introduction

Austin Animal Center is a municipal animal shelter founded in 2011 and is in the city of Austin, Texas [1]. Their mission is to improve animal welfare by reuniting owners with lost pets, fostering, adopting out animals into permanent homes, providing a temporary home and medical care to animals, and educating people to prevent animals from being surrendered. They also have a Trap-Neuter-Return program, which means they neuter or spay feral animals free of charge and return them to their original location. They accept homeless and surrendered animals, regardless of their breed, age, health, and more.

A longstanding issue the Austin animal shelter (and many other underfunded shelters) experiences is overcrowding. This results in transferring animals to new shelters or rescue partner groups or, worst case, euthanizing the animals. Approximately 6.3 million animals are surrendered to animal shelters each year in the United States, with approximately 920,000 of those animals euthanized [2]. Around 58% of the animals put down are cats and 42% are dogs. In our research we aim to gain insight into which factors contribute to an animal's chances of getting adopted and how a shelter can increase adoptions rates.

Problem Statement/Data Sources

Kaggle hosted a competition in 2016 with the goal of improving the outcomes (adopted, died, euthanized, returned to owner, transferred) for cats and dogs at the Austin Animal Center. The data set, titled *Shelter Animal Outcomes,* is available at Kaggle.com under the *Competitions* tab. Alternatively one may use this link.

Outcome is defined as the status of the animal at the time of leaving the shelter. A cleaned training (26,730 observations) and testing (11,457 observations) data set are provided, with a random 70/30 split. The data was collected from October 2013 until March 2016. Each observation represents the outcome of either a cat or dog and information about the animal.

Exploratory Data Analysis

Below shows a list of variables given and their descriptions:

Variable	**Variable Description**
Outcome Type *(Y var.; multi-class)*	Adoption, death, euthanasia, return to owner, or transfer
Outcome Sub Type	Adopted through foster, partner transfer, etc.
Animal ID	Unique animal identifier
Animal Name	Name of animal
Date Time	Date and time animal *left* shelter
Animal Type	Cat or dog
Sex Upon Outcome	Neutered male, spayed female, intact male or female
Age Upon Outcome	Age when left shelter
Breed	Animal breed
Color	Animal color

*Appendix A.1 shows a snippet of the dataset

We performed some data cleaning on these variables since most of the variables were qualitative and included 5+ distinct values. We modified "Breed," which contained the exact breed of the animal, to "Breed Type" which tells us whether the breed of the animal is a mixed breed or purebred. We modified "Color" which contained many different values to "Color Number" which indicated how many distinct colors an animal had on their coat. We normalized "Sex Upon Outcome," which contained whether an animal was neutered/spayed and the gender of the animal, to two binary columns of "is male" (1 if the animal was male and 0 otherwise) and "is neutered" (1 if the animal is neutered/spayed and 0 otherwise). We also updated "Age Upon Outcome," which contained the age in weeks, months, and years, to only provide the age in weeks. Updating these variables allowed us to more easily utilize them when creating our models.

Train Data

Based on **Figure 1**, adoption was the most likely outcome for cats and dogs, with transfer being the second most likely outcome. The live release rate was 93.4%, which is the percentage of animals leaving the shelter alive.

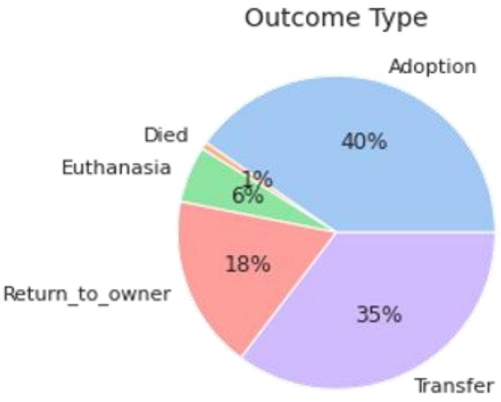


Figure 1. Animal Outcome Pie Chart

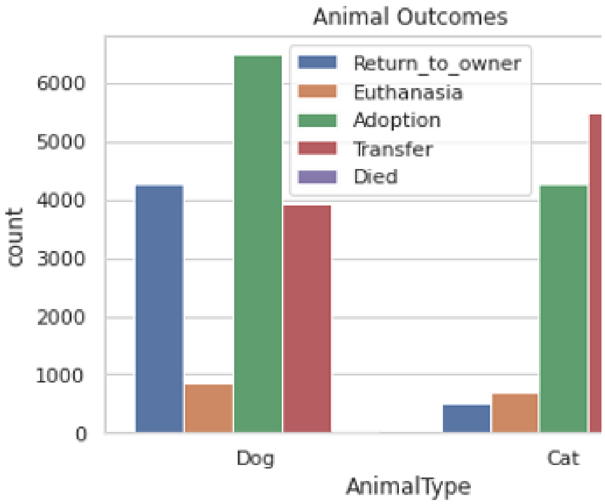


Figure 2. Animal Outcome Bar Chart

Figure 2 shows a more detailed breakdown based on animal type. Dogs are more likely to be returned to their owners compared to cats. This may be due to more dogs entering the shelter with a collar and identification tags. Cat owners may not be as likely to microchip or have a collar, especially if the cat is indoors. Another likely explanation is that there may not be many cats entering the shelter that are lost, as cats are more likely than dogs to find their way homes if they escape their house.

Even though both cats and dogs had high adoption counts, cats had a higher occurrence of transfers. **Figure 3** shows the outcome of cats based on their age. It seems that kittens get more adopted and transferred than middle aged or late life cats.

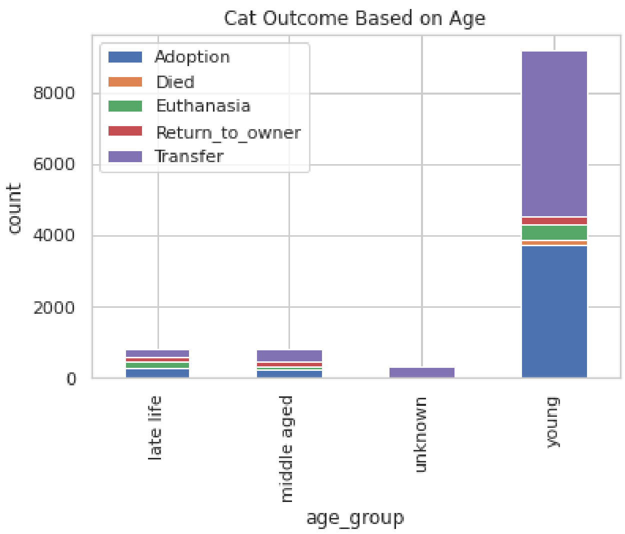


Figure 3. Outcomes for Cats Based on Age Group

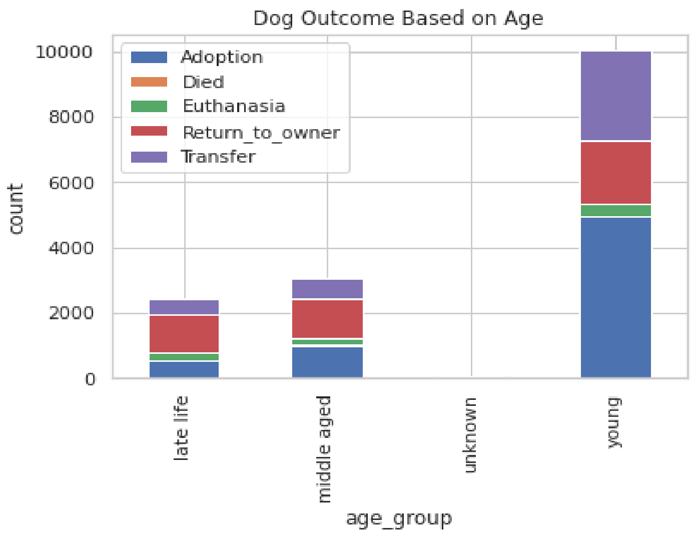


Figure 4. Outcome for Dogs Based on Age Group

Appendix A.2 and A.3 show a more detailed analysis of cats and dogs based on age in weeks. Kittens, ages 8-24 weeks, are more likely to be adopted and older cats are less likely. Newborns, from days old to 4 weeks, are more likely to be transferred. Dogs have similar results, as shown in **Figure 4**, except they are more likely to be returned to their owner, especially puppies.

Figure 5 displays a notched box plot of number of adoptions and euthanasia based on day of week. The notches represent a 95% confidence interval around the median count of adoptions or euthanasia. Hypothesis tests can be performed to evaluate if the median count differs or not when comparing multiple groups.

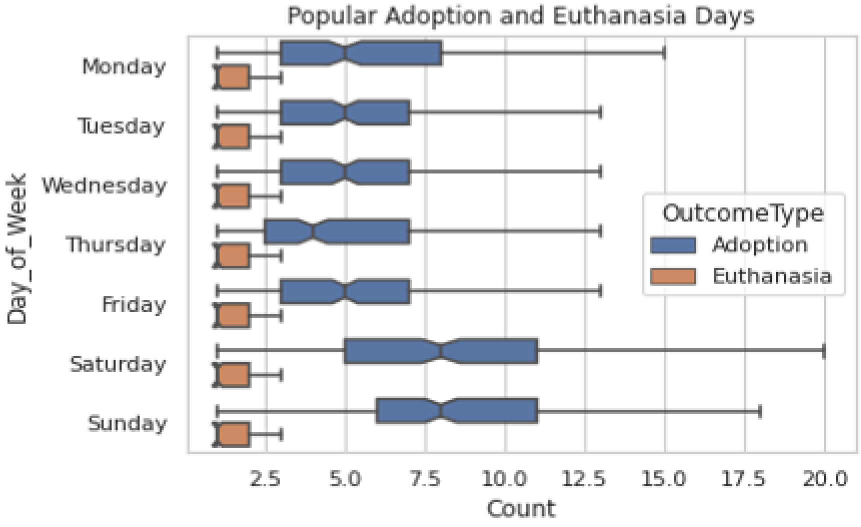


Figure 5. Adoption and Euthanasia Notched Boxplot Counts

For adoptions, Monday, Tuesday, Wednesday, and Friday overlap, meaning their median count of adoptions do not statistically differ. Thursday does not overlap with any days, and Saturday and Sunday do not overlap with Monday-Friday. This means that the null hypothesis can be rejected leading to the conclusion that their median adoption count is statistically different. It seems like Saturday and Sunday are more popular days to adopt and Thursday is the least popular. There does not seem to be a relationship with the number of animals euthanized or days of the week.

To get a more detailed explanation as to whether physical features of an animal influence adoptions, two bar plots were produced. **Figures 6** and **7** show the most popular adopted breeds and colors of animals adopted. For cats, there seems to be no strong preference for color, however, tricolored dogs seem to be more popular. The x axis represents the number of adopted dogs (or cats) divided by the total number of dogs (or cats) based on color. For tricolored dogs, more than 40% of tricolored dogs were adopted. Purebred cats and mixed dogs get adopted, which is an interesting find.

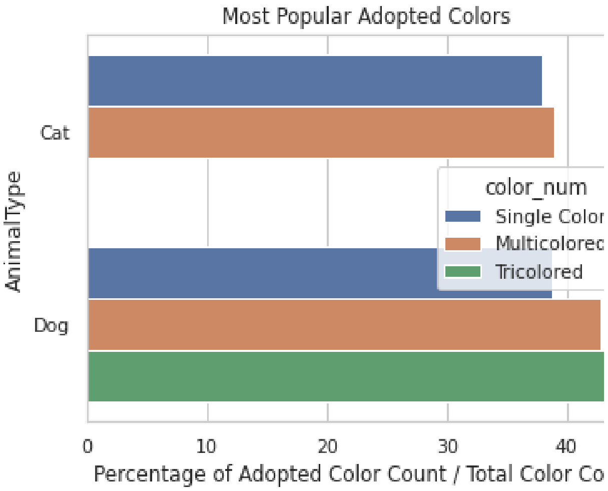


***Figure** 6. % of Adopted Animals Based on Color*

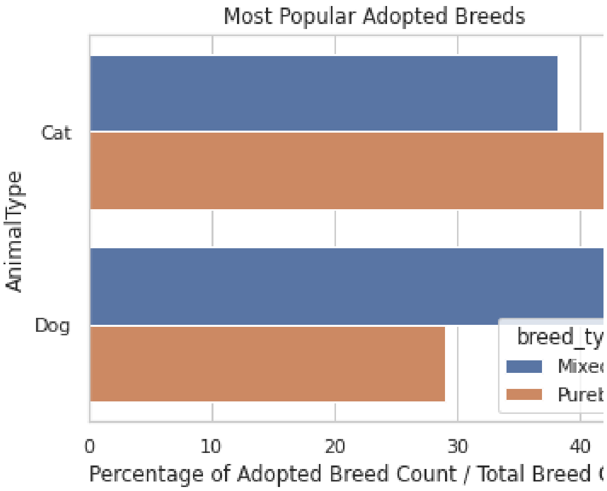


Figure 7. % of Adopted Breeds Based on Color

Charts in Appendix A.4 analyzes the data even further to see if other variables influence adoption and the bar chart in Appendix A.5 shows that one of the most common reasons for being put down is if the animal is suffering, with cats displaying higher counts of suffering. Kittens 4 weeks old and cats greater than 52 weeks old have a higher chance of being euthanized. Each graph seems to confirm that there is a relationship between the variables analyzed and outcome types. Therefore, all variables discussed above will be included in the models.

Proposed Methodology

A core objective of this project is to understand the nature of animal outcomes using predictive modeling. We chose to focus on a selection of three predictive models that capture a variety of modeling approaches.

The random forest (RF) model works by learning an ensemble of decision trees used in classification tasks. Extreme gradient boosting (XGBoost) is a version of gradient boosting that utilizes clever penalization of trees and proportional shrinking of leaf nodes to enhance its predictive capacity over standard boosting methods. Linear support vector machine for classification (Linear SVC) works by drawing a linear decision boundary in multi-dimensional space.

The linear SVC model was selected to test whether a linear decision boundary would serve as an effective predictive model without overfitting the data. Since we are assessing a relatively small number of features relative to the number of observations in the data set, we believe this is possible.

The RF and XGBoost models were selected to test the performance of two different ensemble modeling approaches, both of which generally perform very well on tabular data. These two different methods take different approaches to creating an ensemble of decision trees. The RF model focuses on building a collection of trees that work together to generate a correct prediction, where each tree makes a prediction and the tree with the most votes is selected. The XGBoost model builds trees in an additive manner (one after the other) so that a downstream tree is designed to address the limitations of the tree upstream from it.

While these two models are very similar in some respects, they demonstrate different performance characteristics depending on the predictive task and are thus deemed worthy of comparison.

Analysis and Results

Data was divided into training and test sets using an 85/15 split. Hyperparameter tuning was performed (to applicable models) using 3-fold cross-validation and a randomized grid search approach where 300 hyperparameter combinations were evaluated for both XGBoost and Random Forest models. For SVM, one-vs-one and one-vs-rest decision functions were evaluated.

Table 1 shows the selection of the top-performing hyperparameters for each model. For XGBoost and RF models, a similar number of estimators was selected (350 and 380, respectively). However, the best XGBoost model was of much larger depth than that of RF (60 vs 5, respectively).

For the SVC model, a one-vs-one decision function shape was the top-performer. In this approach, the multi-class problem is handled as a binary classification problem for each pair of classes, and the highest probability class is selected.

Hyperparameter	**XGBoost**	**Random Forest**	**SVM**
Learning Rate	0.55	NA	NA
Max Depth	60	5	NA
Max Features	sqrt	sqrt	NA
Min Samples per Leaf	0.05 (percent)	12 (count)	NA
Min Samples required for Split	0.525 (percent)	4 (count)	NA
N estimators	350	380	NA
Decision Function Shape	NA	NA	One vs One

Table 1: *Top-performing model hyperparameters.* Results captured across 3-fold cross-validation hyperparameter search using a randomized grid search approach on the training set.

Once tuned, the top performing models were evaluated on the training set using 10-fold cross validation. The results in **Figure 8** indicate that the XGBoost was the top-performing model with an accuracy score of 0.595. The XGBoost model outperformed the RF model, but not by much (acc. 0.595 vs 0.593, respectively). However, both models outperformed the SVM model (acc. 0.438).

These results suggest that an ensemble-based approach is superior to a simpler approach within the context of this prediction problem. It is likely that these outcomes are based on a complex combination of features and that SVM is not sufficiently flexible to make use of all available information.

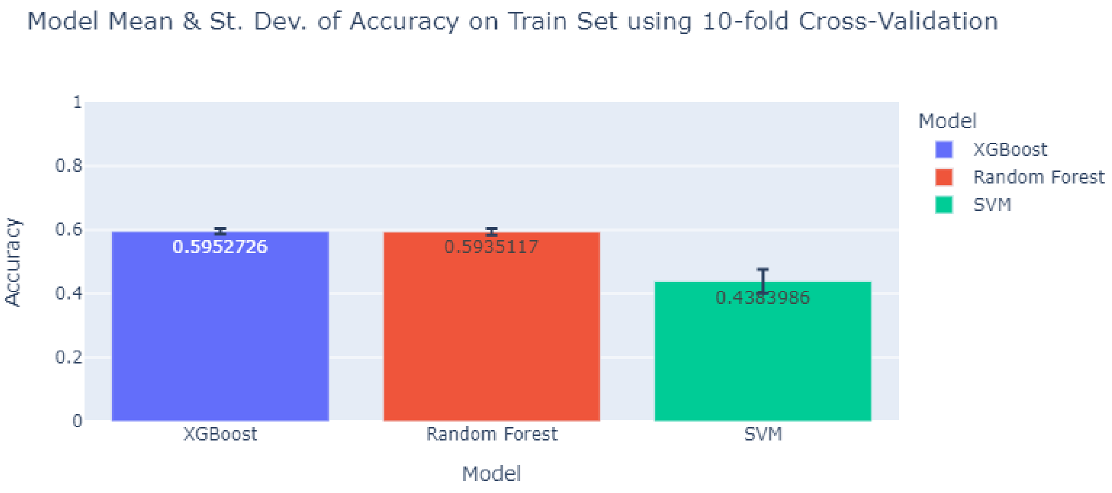


Figure 8: *Model Performance on Training Set.* Mean accuracy (large bars) & 1 standard deviation of accuracy (error bars) assessed over 10-fold cross validation of the training set.

The top-performing model, XGBoost, was selected for evaluation on the held-out test set. To provide some context as to how useful the model is, we also trained a naïve model to predict classes based on the most frequently observed class within the training set. **Figure 9** illustrates the XGBoost model provides superior predictive performance to a naïve approach (accuracy 0.591 vs 0.399, respectively). The ~0.2 accuracy difference suggests that there is significant predictive ability gained by employing the XGBoost model.

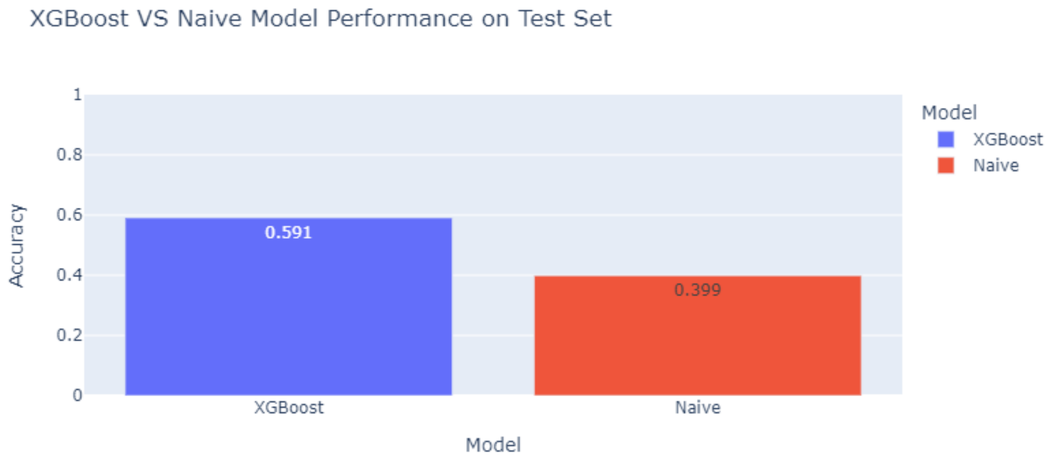


Figure 9: *XGBoost vs Naïve Model Performance on Test Set.*

In assessing the confusion matrix of our final model, we learn that the XGBoost model was proficient at classifying adoption and transfer cases, where correct classifications of these categories account for the majority of outcomes (30% and 20%, respectively). Perhaps not surprisingly, these two outcomes were also among the most confused. Transfer was incorrectly predicted as adoption in 11% of cases, and adoption was predicted as transfer in 4.5% of cases. Return to owner outcomes were also frequently (incorrectly) predicted as adoption (6.8% of cases).

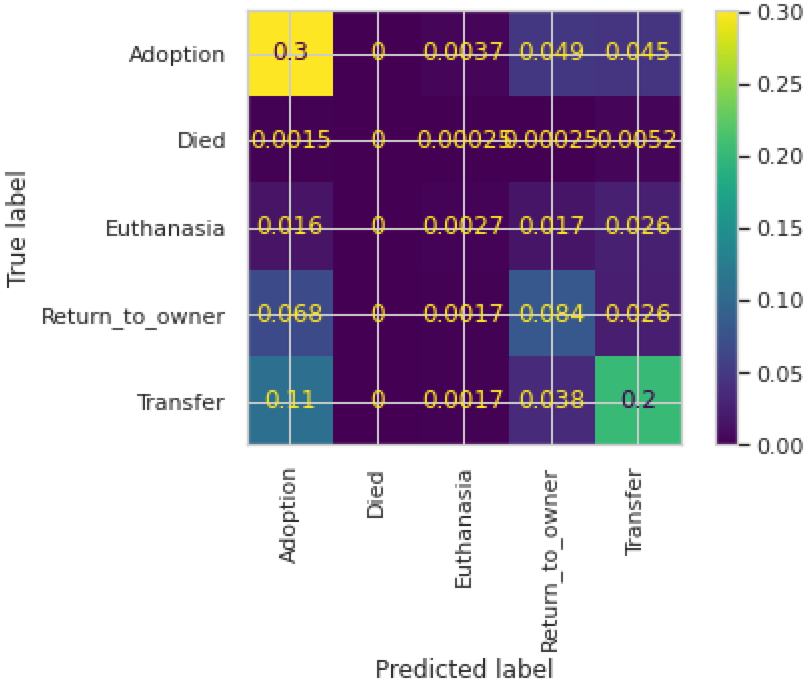


Figure 10: *Confusion Matrix of All Predictions for XGBoost Model (normalized values).*

The model feature information from the XGBoost model is shown in **Figure 11**. *Age in weeks* and *age group* are the top-2 features as measured by MDI. Together they account for 42% and 18% of the decrease in note impurity of the model. **Figure 11 (right)** indicates that the retrained model relied on *is neutered* and *age in weeks*, with each displaying a similar permutation importance value (approximately 0.11 each). This indicates that the retrained model performs poorly when either feature is dropped from the model.

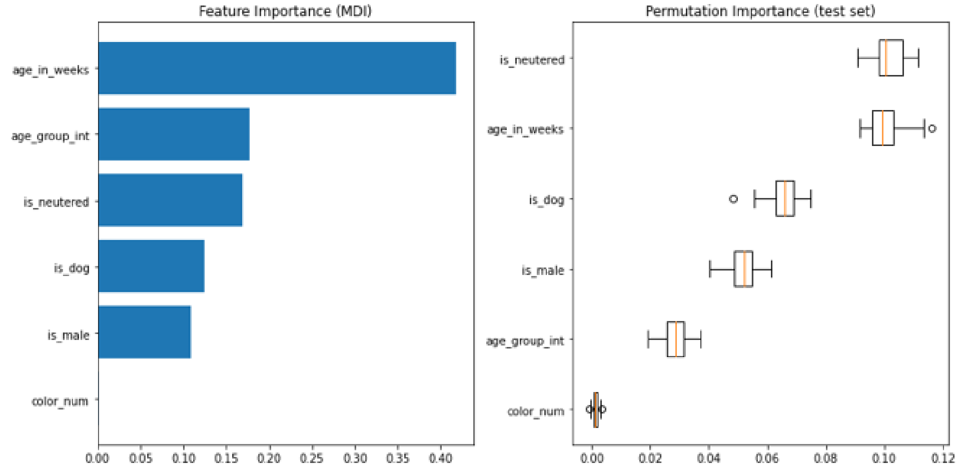


Figure 11: *Feature Information of XGBoost Model.* Feature importance (left) and Permutation importance (right). Feature importance is measured as the mean decrease in impurity (MDI). Permutation importance was generated by repeated sampling of the test set over *n* = 50 iterations.

Together, these results speak to the importance of two key features: *age in weeks* and *is neutered*. A*ge in weeks* is the most important variable in the model, with over twice the importance of the next-best feature and the second largest permutation importance value. The *is neutered* variable appears to be the second-most important feature overall. It may be possible the shelter has a mandatory spay or neuter adoption policy. It ranked third in terms of feature importance and was the most important feature in terms of permutation importance.

Conclusion

From our data exploration and modeling, we were able to see that physical traits such as color, breed, and age can affect cats' and dogs' chances of getting adopted. Although many physical traits of animals cannot be changed, animal shelters can identify specific animals which are at risk of not being adopted and try to showcase them more on their social media platforms or during high volume adoption days. We found that Saturday and Sundays are popular days for adoption, so keeping high risk animals in the front can increase their chances of being adopted. From our modeling we found that neutered and spayed animals have higher chances of being adopted. Animal shelters can focus on making sure that their animals are neutered and spayed to increase their likelihood of being adopted.

Further research can be performed on animal shelters to improve adoption rates. In our work we had few independent variables to work with and many of them were focused on physical attributes. Although these were helpful data points in understanding adoption rates for animals, they unfortunately were not all features that can be changed for some animals. Evaluating more data such as where in the animal shelter more animals are likely to be adopted, and which time of the year more animals are likely to be adopted can help animal shelters showcase high risk animals.

Lessons Learned

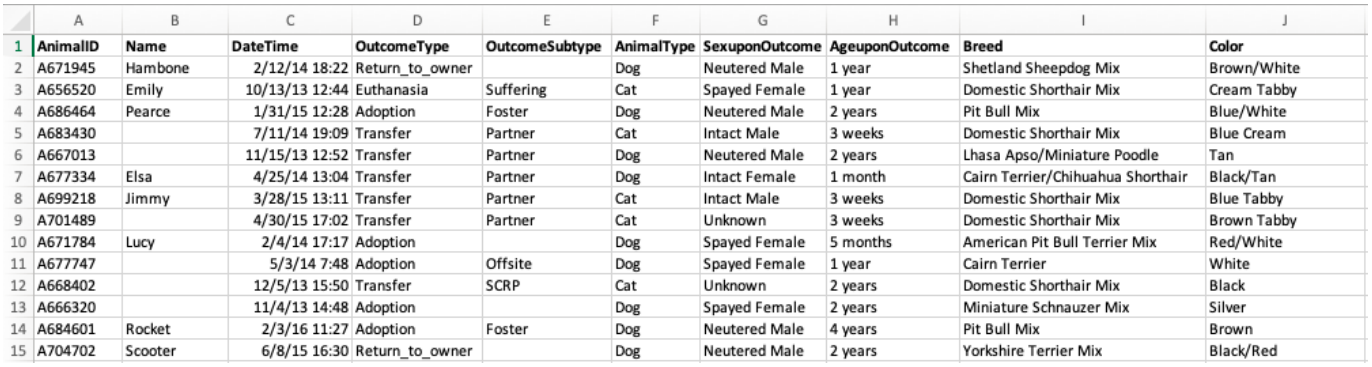
This course and project helped highlight the importance of data exploration such as understanding trends in the dataset, detecting multicollinearity, and observing relationships between the independent and dependent variables. We also learned the importance of being clear and concise in the paper, especially during the technical sections, and the importance of testing a variety of models. The performance of ensemble-based methods tends to perform well for tasks using tabular data. This project was an example of that and ensemble models like XGBoost and random forest should be considered when selecting models for similar tasks.

Resources

1. https://www.austintexas.gov/austin-animal-center
2. https://www.aspca.org/helping-people-pets/shelter-intake-and-surrender/pet-statistics
3. https://www.kaggle.com/competitions/shelter-animal-outcomes/overview/evaluation

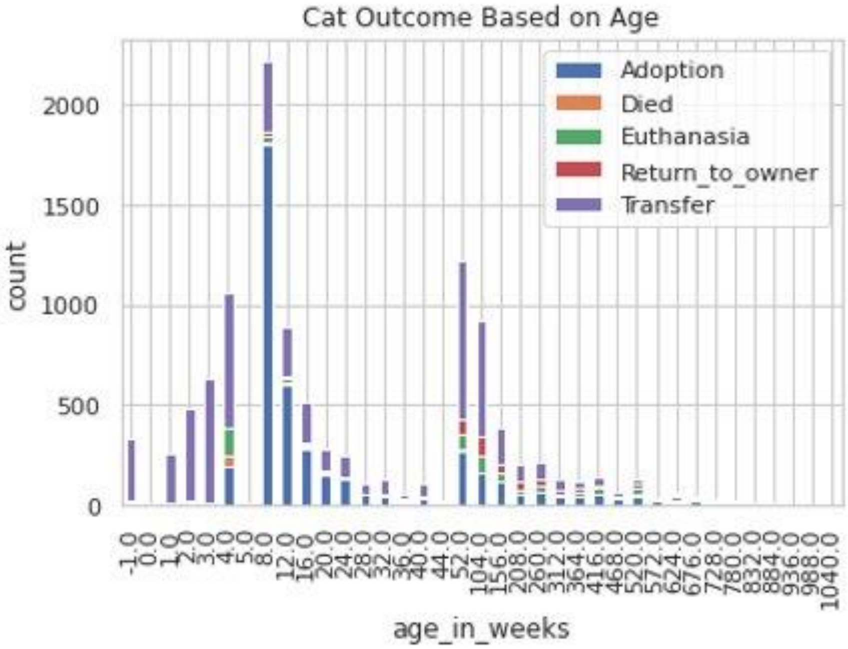
Appendix

A.1 Snippet of Dataset

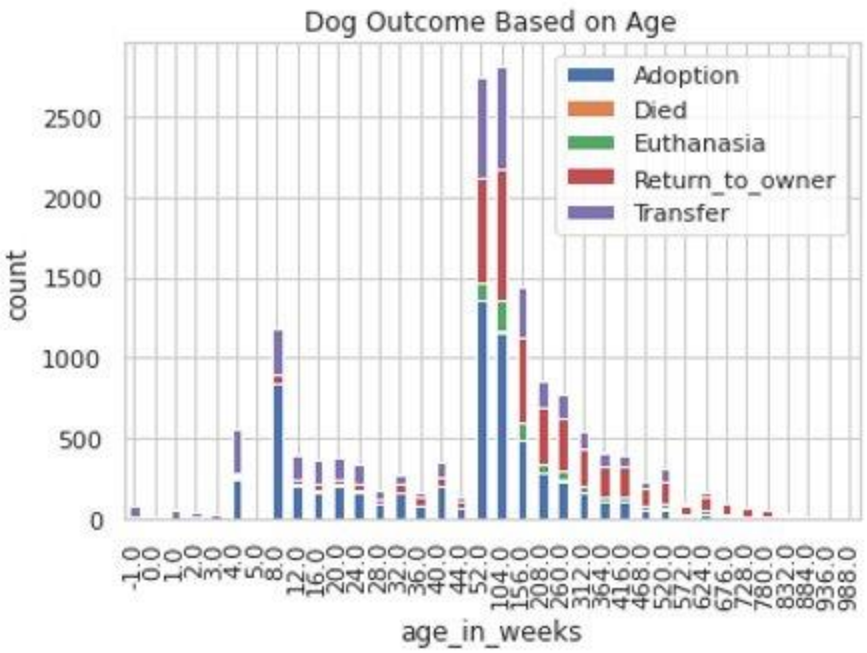


	A	B	C	D	E	F	G	H	I	J
1	**AnimalID**	**Name**	**DateTime**	**OutcomeType**	**OutcomeSubtype**	**AnimalType**	**SexuponOutcome**	**AgeuponOutcome**	**Breed**	**Color**
2	A671945	Hambone	2/12/14 18:22	Return_to_owner		Dog	Neutered Male	1 year	Shetland Sheepdog Mix	Brown/White
3	A656520	Emily	10/13/13 12:44	Euthanasia	Suffering	Cat	Spayed Female	1 year	Domestic Shorthair Mix	Cream Tabby
4	A686464	Pearce	1/31/15 12:28	Adoption	Foster	Dog	Neutered Male	2 years	Pit Bull Mix	Blue/White
5	A683430		7/11/14 19:09	Transfer	Partner	Cat	Intact Male	3 weeks	Domestic Shorthair Mix	Blue Cream
6	A667013		11/15/13 12:52	Transfer	Partner	Dog	Neutered Male	2 years	Lhasa Apso/Miniature Poodle	Tan
7	A677334	Elsa	4/25/14 13:04	Transfer	Partner	Dog	Intact Female	1 month	Cairn Terrier/Chihuahua Shorthair	Black/Tan
8	A699218	Jimmy	3/28/15 13:11	Transfer	Partner	Cat	Intact Male	3 weeks	Domestic Shorthair Mix	Blue Tabby
9	A701489		4/30/15 17:02	Transfer	Partner	Cat	Unknown	3 weeks	Domestic Shorthair Mix	Brown Tabby
10	A671784	Lucy	2/4/14 17:17	Adoption		Dog	Spayed Female	5 months	American Pit Bull Terrier Mix	Red/White
11	A677747		5/3/14 7:48	Adoption	Offsite	Dog	Spayed Female	1 year	Cairn Terrier	White
12	A668402		12/5/13 15:50	Transfer	SCRP	Cat	Unknown	2 years	Domestic Shorthair Mix	Black
13	A666320		11/4/13 14:48	Adoption		Dog	Spayed Female	2 years	Miniature Schnauzer Mix	Silver
14	A684601	Rocket	2/3/16 11:27	Adoption	Foster	Dog	Neutered Male	4 years	Pit Bull Mix	Brown
15	A704702	Scooter	6/8/15 16:30	Return_to_owner		Dog	Neutered Male	2 years	Yorkshire Terrier Mix	Black/Red

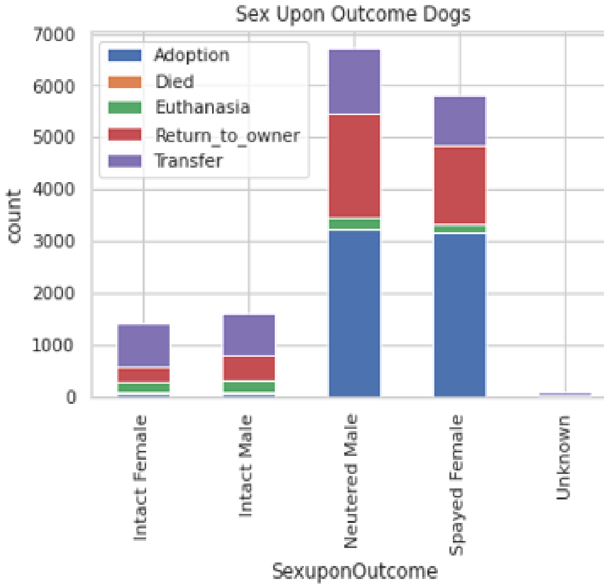
A.2 Detailed Breakdown of Cat Outcome Based on Age in Weeks

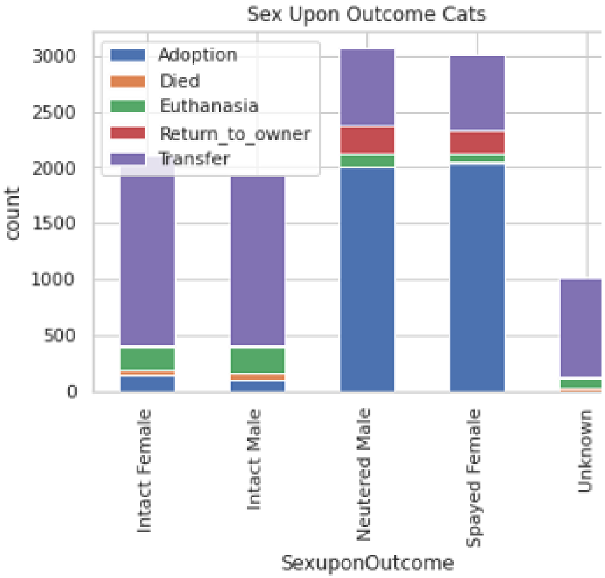


A.3 Detailed Breakdown of Dog Outcome Based on Age in Weeks



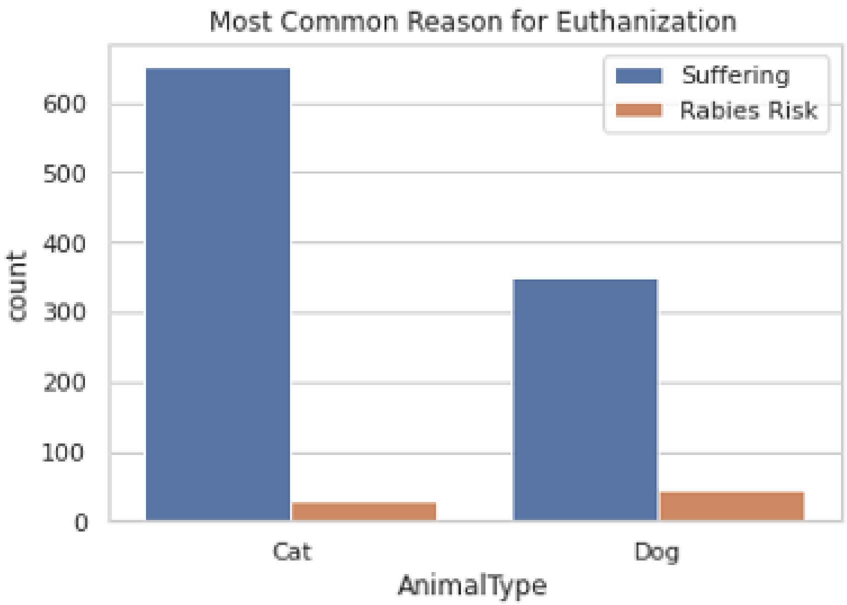
A.4 Animal Sex for Dogs and Cats





*Neutered or spayed animals get more adopted, but this is most likely due to mandatory spay and neuter policy on adoptions.

A.5 Reasons for Animal Euthanization Based on Animal Type


Most Common Reason for Euthanization
Suffering
Rabies Risk
count
600
500
400
300
200
100
0
Cat
Dog
AnimalType