Analysis of Frederick County Animal Control Data: Understanding and Predicting Animal Shelter Outcomes

Kaitlyn Longo

Loyola University of Maryland

May 5, 2020

Abstract

Objective – This report includes a thorough data analysis for Frederick County Animal Control (FCAC) in effort to identify important trends and to determine the factors that have the most impact on predicting an animal's outcome (adoption/return to owner or euthanasia) at FCAC.

Approach – The data used for the analysis covered October 2011 through September 2019 and included roughly 30,000 dogs and cats. The first phase of the project involved conducting thorough exploratory data analysis to understand different patterns within the data. The second phase of the project involved implementing different supervised machine learning algorithms to predict if a dog or cat would be adopted/returned to owner or euthanized. Separate models were created for dogs and cats in order to compare the factors that had the most predictive power.

Results – The exploratory data analysis phase led to many important findings. The 21702, 21701, and 21703 zip-codes were identified as top regions for abandoned animals. Thurmont (21788) was also identified as a top region for stray and seized cats. FCAC leadership can use this information to help promote spay/neutering programs in these areas or even work with veterinarians to lower costs in those areas. Additionally, an analysis of adoption times showed that Wednesday evenings were the most popular day for adoptions, followed by Saturday afternoons. FCAC staff can use this information to determine when to schedule the most convenient adoption times or even extend their hours. While analyzing the length of time dogs and cats spend at the shelter, an important finding was that cats wait a much longer amount of time to get adopted than dogs. Cats waited a median of 42 days, while dogs waited a median of 15 days. In particular, juvenile dogs are the quickest to get adopted, and adult cats wait the longest to get adopted. Dynamic visualizations showing additional plots and charts were compiled into an interactive dashboard via Tableau.

The three different algorithms used for machine learning were random forest, XGBoost, and logistic regression with lasso regularization. For both cats and dogs, the three algorithms performed nearly the same. For cats, the accuracy, F1 score, and AUC were roughly 0.76, 0.74, and 0.82 respectively. For dogs, the accuracy, F1 score, and AUC were roughly 0.83, 0.76, and 0.86. These numbers are close to 1 (meaning a perfect model) but show that there is room for improvement. Based on the random forest models, the top predictive features for cats were the age of the cat, the condition the cat is in, the intake type, and whether or not it was pre-altered. For dogs, the top features were the intake type, the condition the dog is in, the age of the dog, and the gender of the dog. It is evident that both the age and the condition of both dogs and cats are key factors when it comes to their outcome at FCAC.

Conclusions – Overall, supervised machine learning techniques were able to predict an animal's shelter outcome with some success. The biggest challenge was due to a limited amount of data in both number of records and variables. An increase in data (i.e. from other animal shelters) could provide more generalizable results and a model that can be used as a decision-making tool. Although the models should never be used as a mechanism for deciding if an animal should be euthanized, it could be used as a tool to identify animals that have a higher chance of being euthanized. If euthanasia is avoidable, these animals could potentially be promoted more often on social media and given extra attention (i.e. behavioral interventions).

Keywords: animal shelter, machine learning, random forest, XGBoost, logistic regression

Table of Contents

List of Figures

List of Tables

1. Introduction

1.1 Background

According to the American Society for the Prevention of Cruelty to Animals (ASPCA), approximately 6.5 million animals enter animal shelters in the United States each year, roughly half dogs and half cats. Only half of these animals get adopted, and about 25% of them are euthanized every year. (The remainder are either returned to their owners or deceased.) Although some animals may be euthanized because they are dangerous, there are many innocent animals that lose their lives because they could not find a home. It is up to humans to save these potentially adoptable animals. The analysis outlined in this report was conducted for Frederick County Animal Control (FCAC), and their mission is to prevent animal cruelty, abuse, and neglect. They shelter animals without homes and help them find suitable “forever” homes. In addition, their goal is to educate the public on animal issues to create a more aware community. As it is located in one of the largest counties in Maryland, FCAC intakes thousands of animals each year. Since there are so many, it is nearly impossible to find homes for all pets in need.

1.2 Objectives

FCAC records thorough data about every animal they take in. They use this data to understand general trends, such as how many animals are taken in each month, how many are adopted, how many are euthanized, etc. However, FCAC is interested in a more thorough data analysis that dives deeper into their data and understanding different relationships between variables. Thus, the overarching goal of this project is to provide an extensive analysis of FCAC’s data in order to help them understand how they can potentially improve animal adoption rates and highlight important trends. Specifically, a large portion of the project is aimed at predicting animal shelter outcomes (adoption, return to owner, euthanasia, etc.) and highlighting the factors that have the most influence on those outcomes. The goals for this project are categorized as descriptive and predictive, which are outlined below:

Predictive Goals:

- Determine if machine learning classifiers can be used to predict a dog/cat’s shelter outcome at the FCAC
- Identify the top determinants for predicting dog/cat shelter outcomes
- Determine if the top determinants for predicting the shelter outcome differs for dogs and cats

Descriptive Goals:

- Identify trends in the amount of time dogs/cats stay at the shelter
- Determine if there is a relationship between the age of the dog/cat and its outcome
- Identify areas of Maryland/Frederick County where the most dogs/cats are coming from

Although the analysis will not solve the problem entirely, it can help provide insights about animals that have a higher risk of getting euthanized. If these high-risk animals are identified, this could also potentially decrease costs incurred by length of stay prior to either adoption or euthanasia. Additionally, if shared with the community, the analysis could be used to raise awareness and lower the rate of animal intakes. Lastly, it could also be leveraged by shelters across the country to identify ways to help animals in need, improve daily operations, or even inspire them to collect more detailed data on their animals.

Predicting an animal’s shelter outcome is a challenging task for a variety of reasons. There is likely a lot of variation in the data that could potentially lead to biases in the model. For example, the majority of golden retrievers are well-behaved, but there could always be a one-off case of an ill-behaved retriever. Situations like this could be hard for a model to successfully predict the animal's outcome because it is used to seeing the typical cases rather than very unique cases (unless there is highly detailed information on that particular animal's behavior). Additionally, humans themselves are hard to predict. Although we are predicting the animal's outcome, we are really predicting a choice that a human would make. Lastly, there is limited data both in volume and in variables. The data only contains certain characteristics about each animal, which may or may not be enough to predict their outcome.

2. Research platform

2.1 Software resources

All of the analysis was conducted using the Python programming language in a Jupyter Notebook environment. Jupyter Notebooks are useful because they allow for step-by-step code, mathematical equations, visualizations, and text explanations. Separate Jupyter Notebooks were created for data cleaning/feature engineering and exploratory data analysis (EDA) and predictive modeling. Since the data set was not too large, a local machine was sufficient for data analysis and model development. The exploratory data analysis phase involved using common data science Python packages such as Pandas, NumPy, and Matplotlib for quick visualizations. The predictive portion of the project involved implementing machine learning algorithms via the Scikit-Learn Python package as well as a separate package called XGBoost. XGBoost is a package commonly used in Kaggle competitions that provides an implementation of gradient boosted machine learning. Lastly, since the client's mission involves informing the community, it is important to have aesthetically pleasing and informative visualizations. Tableau, which is a powerful and interactive data visualization software, was used to create an interactive visualization dashboard that provides key insights in one place.

2.2 Dataset

FCAC gathers and records data on all of their animal intakes and their associated outcomes, which includes information on dogs, cats, and any other animals they receive. Animal intake data is separate from the animal outcome data, but they can be combined based on the unique animal ID. The animal intake data contains information regarding the type of animal, primary/secondary breed, age/age group, gender, primary/secondary color, name, when they were taken in, whether they were strays/surrenders/etc., and many other fields. The outcomes data set includes the outcome date, what the outcome was (adoption, euthanasia, return to owner, transfer etc.), and reason for euthanasia if that was the outcome. The data that was provided covers October 2011 through September 2019 and contains roughly 34,000 unique animals (~9,300 dogs and ~20,300 cats). The dataset contains a few inconsistencies since many fields have missing data. To alleviate this issue, these fields were either omitted during the analysis or new features were created to replace missing data; this will be discussed further in a separate section. A sample of the data is depicted below:

Table 1. Animal intakes example

Animal ID	Animal Name	Species	…	Operation Type	Age Group	…
A12167351	NYLA	Dog		Stray	Adult (1yr-4yrs)	
A12183835	MAZDA	Cat		Return	Juvenile (8wks-11mon)	
…						

Table 2. Animal outcomes example

Animal ID	Animal Name	Species	…	Outcome Type	Outcome Subtype	…
A12167351	NYLA	Dog		Return to Owner/Guardian	Stray Reclaim	
A12183835	MAZDA	Cat		Euthanasia	Temperament	
…						

The technical aspects of the project were broken up into different phases, which include data cleaning, feature engineering, EDA, model development, and model assessment. The process and results of each phase will be discussed in Section 4 of this document.

3. Related work

3.1 Lin's finding forever homes

This project was directly inspired by a similar effort conducted by Ms. Joanne Lin as a part of Thinkful's data science bootcamp. Ms. Lin leveraged a dataset found on Kaggle from the Austin, Texas Animal Center to determine if she could predict whether a pet gets adopted. The data set included fields such as the animal name, date of the outcome, type of outcome, type of animal, gender, spayed/neutered, age, breed, and color. The dataset is fairly large, with 44K dogs and 30K cats. She started by conducting exploratory data analysis to understand the different variables and how they might relate to the different outcomes (Return to owner, Adoption, Transfer, Euthanasia, or Died). She then performed feature engineering to include features such as the coat color and coat pattern.

Ms. Lin used the Scikit-Learn Python library to apply two different machine learning algorithms: random forest and support vector machine. She found that for dogs, the most predictive features were the breed of the dog, whether it is spayed/neutered, and the dog's age. For cats, the top features were whether the cats were spayed/neutered, whether that cat had a name, and the cat's age. The accuracy for her models (separate models for dogs and cats) was around 0.82, so she was able to correctly predict the animal's outcome 82% of the time.

Ms. Lin also conducted analysis to identify the most "adoptable" names, which was Ginger for dogs and Cookie for cats. She also found that the majority of adoptions occurred on the weekend for both dogs and cats (Levenson, 2019). Lin's work was the driving motivation for this project, and the similarities/differences will be discussed in Section 5 of this report.

3.2 Dyché's making a difference with animal shelters

Another interesting study was conducted by Ms. Jill Dyché who is focused on improving the shelter adoption processes. Although this study is more focused on the processes employed by different shelters, it includes important and relevant findings. She found that the peak time for animal shelters is during the summer because people often send their animals to shelters before going on vacations in order to avoid paying kennel/pet sitting fees. She also discovered that at least 90% of dogs are mislabeled as the wrong type of breed (usually pit bulls who have higher euthanasia rates). Lastly, she discovered that videos of animals at the shelter leads to a higher chance of the animal being adopted as opposed to static images (Underwood, 2016).

4. Analysis

4.1 Data Pre-Processing

4.1.1 Data Cleaning

The first phase of data pre-processing involved joining the intakes and outcomes data in order to obtain one final dataset. A left join was first performed on the outcomes data with the intakes data by the Animal ID, however, this led to an incorrect number of total records. By digging deeper into the dataset, some duplicate records were found. This is due to the fact that some animals were taken into the shelter more than once (i.e. returns). To alleviate this, the join was done on both Animal ID and the Intake Date. In the machine learning portion of the project, these duplicate records were removed because they could potentially skew the results since the same animal could have more than one outcome.

As stated previously, many of the fields had missing data. Additionally, there were some fields that did not add value to the analysis; therefore, these fields were dropped. They include: Distinguishing Markings, Date of Birth, ARN, Danger, Danger Reason, Pet ID, Pet ID Type, Site Name, Injury Type, Cause, Age in Months Intake, Age (Age Group was used instead), Source, Intake Reason, Outcome Reason, Length Owned, Unit, Agency Name, and Asilomar Status.

The dataset was already mostly clean as only some of the fields needed further processing in order to be consistent. For instance, the Animal Name field needed to be adjusted to ensure that there were no special characters and that the names were in a uniform format. Additionally, the Jurisdiction field needed to be adjusted in order to contain a clean 5-digit zip-code for further analysis.

A table containing the fields used for analysis and their descriptions is presented below:

Table 3. List of fields and descriptions

Field	Description
Animal ID	Unique ID for each animal
Animal Name	Name given to animal
Species	Type of animal (dog, cat, bat, etc.)
Primary Breed	Dominant breed of animal (pit bull, dachshund, etc.)
Secondary Breed	Second most dominant breed of animal
Age Group	Senior (5yrs and older), Adult (1yr-4yrs), Juvenile (8wks-11mon), Unweaned (less than 8wks)
Gender	M, F, or U (unknown)
Pre-Altered	If the animal was altered in some way at intake; Y, N, or U (unknown)
Spayed Neutered	Y, N, or U (unknown)
Intake Type	Return, Surrender, Stray, etc.
Intake Date/Time	Date/Time of intake
Operation Type	Same as Intake Type (redundant field)
Operation Sub Type	Further explanation for Intake Type (e.g. public drop-off, surrendered for adoption etc.)
Jurisdiction	County animal came from

Condition	Appears normal, sick, aggressive, etc.
DOA	(Dead on Arrival) True or False
Location	Where the animal resides in shelter
Sub Location	Further explanation for location
Outcome Date	Date/Time of outcome
Outcome Type	Animal's outcome (adoption, euthanasia, return to owner, transfer, etc.)
Outcome Subtype	Reason for outcome type (e.g. temperament, owner request)

4.1.2 Feature Engineering

The next phase involved feature engineering, which is the process of transforming the original data into new features that may enhance the analysis. The features that were added and a description of each are presented in the table below:

Table 4. List of new features and descriptions

New Feature	**Description**
Has Name	1 if the animal has a name, 0 otherwise
Is Mix	1 if the animal is a mixed breed, 0 otherwise
Is Black	1 if the primary color is black, 0 otherwise
Is Multicolor	1 if the animal is multi-colored, 0 otherwise
Is Top Dog	1 if the primary breed is in the top 20, 0 otherwise
Time in Shelter	Total time the animal spent in the shelter in days
Intake Day of Week	Day of week of intake
Intake Month	Month of intake
Intake Hour	Hour of day of intake
Intake Year	Year of intake
Outcome Day of Week	Day of week of outcome
Outcome Month	Month of outcome
Outcome Hour	Hour of day of outcome
Outcome Year	Year of outcome
Dog Breed Group	For dogs only (herding, working, terrier, etc.)
Energy Level	For the primary breed only – low, medium, high; note that this is not entirely accurate since research was based on the primary breed only
Shedding Level	For the primary breed only – heavy, minimal, some; note that this is not entirely accurate since research was based on the primary breed only

4.2 Exploratory Data Analysis

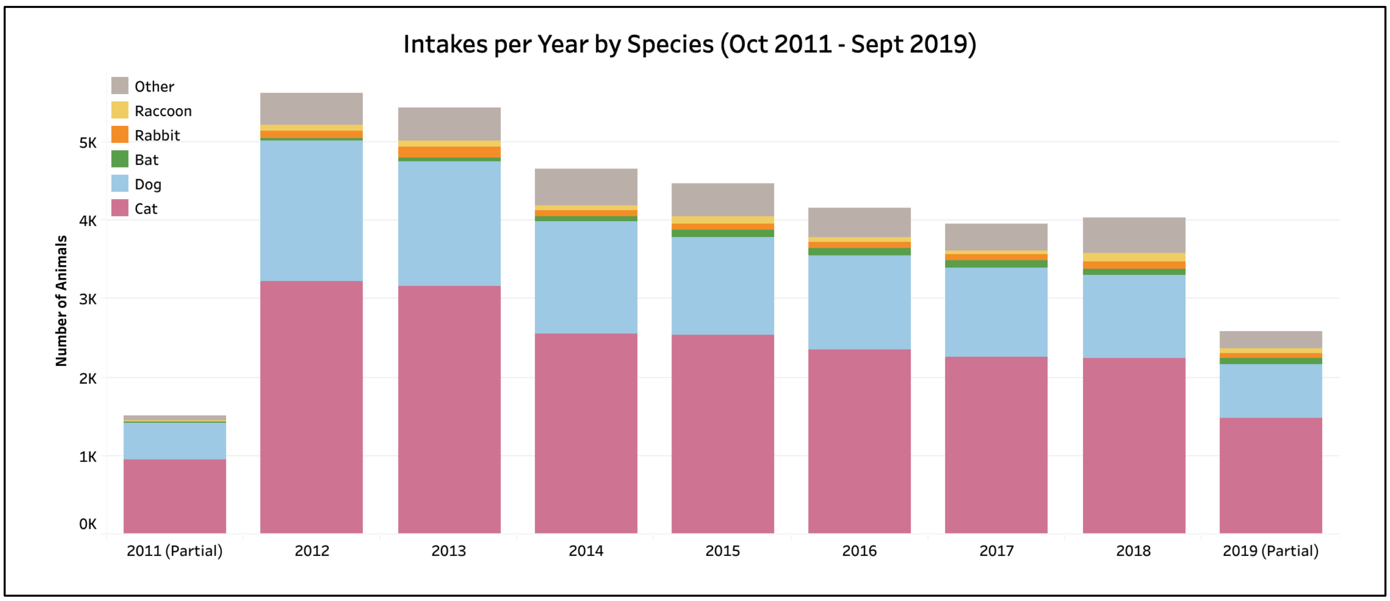
This section will include an analysis of the various different fields within the dataset to answer the descriptive analytic goals and to provide a better understanding of different trends in the data.

4.2.1 Intakes

4.2.1.1 Intakes per year and month

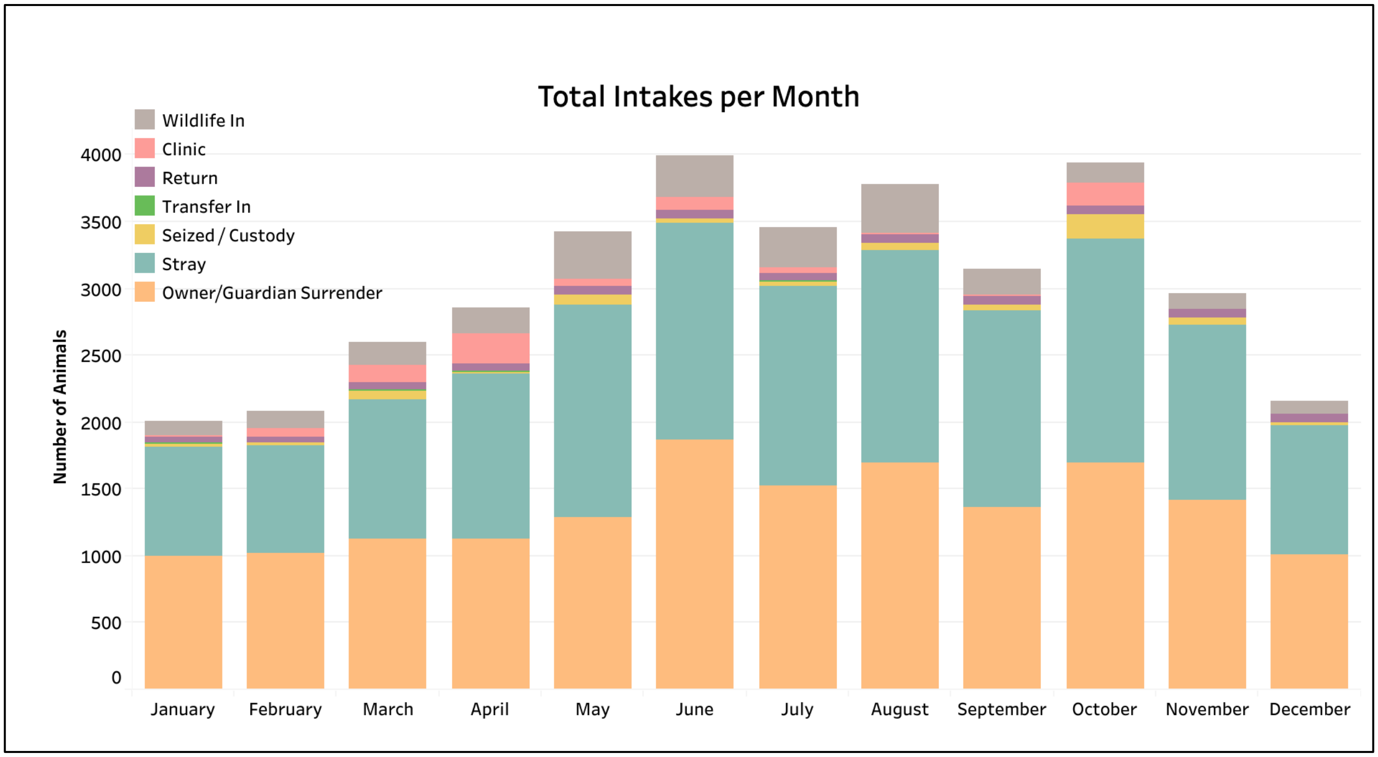
Figure 1 below depicts the number of animals taken in each year at FCAC broken out by the species. (Recall that the data covers October 2011 – September 2019.) It appears that there has been a slow decline in the total number of intakes between 2012 and 2018. Additionally, cats make up the largest number of intakes at over 2,000 per year, followed by dogs at over 1,000 per year.

Figure 1. Intakes per year by species



Next, we can examine trends in intakes per month, as shown in Figure 2 below.

Figure 2. Intakes per month by type



Intakes are highest in the summer months (primarily June) as well as October. The spikes in June and October are driven by owner/guardian surrenders and stray animals. This aligns with Ms. Dyché's study; people might be surrendering their animals to the shelter in order to go on vacation without having to pay for a pet sitter or boarding at a kennel. Additionally, in the United States, "kitten season" is between April and October each year (MacPete, 2015). Thus, FCAC may see a spike in kittens during this time. October is also when people are preparing for holidays and travel, so they might relinquish their pets to get more money or to free up their time. A recommendation is to lower or waive adoption in fees in June and October to promote adoptions during times of heavy intakes. Additionally, FCAC should make sure they have enough staff/volunteers during these busy months.

4.2.1.2 Analyzing where animals came from

One of the descriptive analytic goals for this project involved identifying locations in Frederick County that contribute to the most intakes. Focusing on dogs and cats, Figures 3 and 4 below depict heat maps of Maryland. The darker shades correlate to areas that contribute to a higher number of animals.

Figure 3. Heatmap of Frederick (dog intakes)

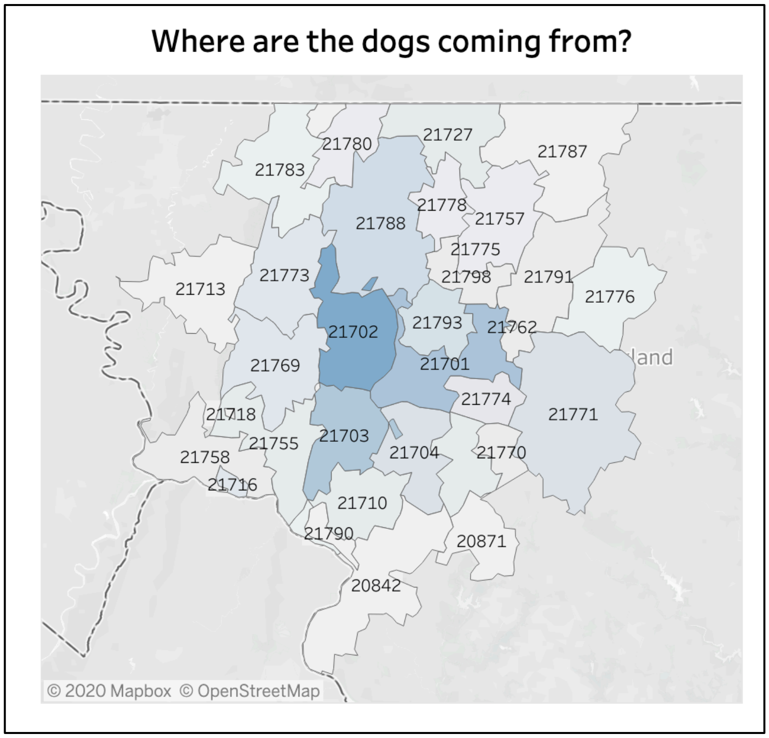
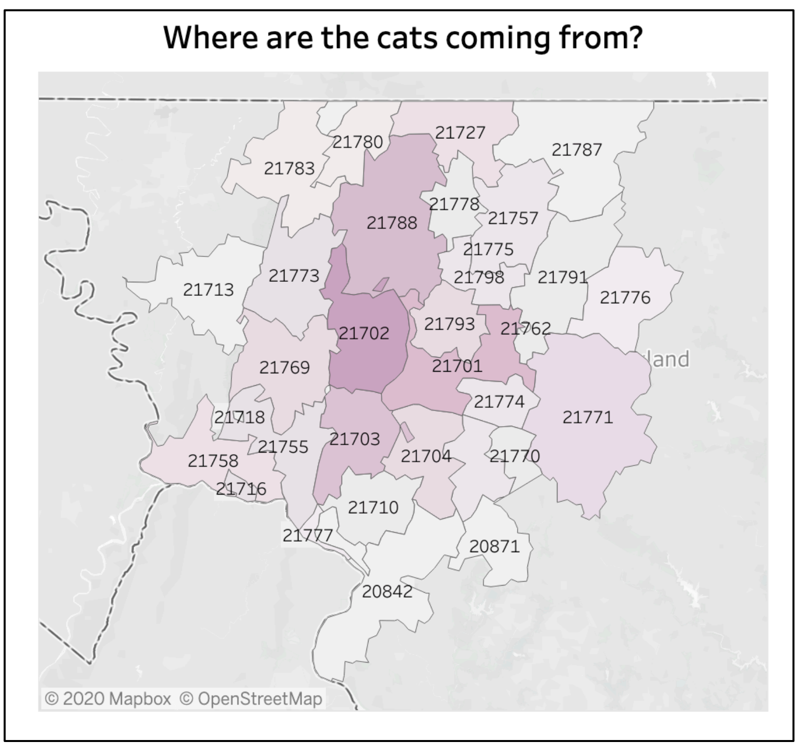


Figure 4. Heatmap of Frederick (cat intakes)



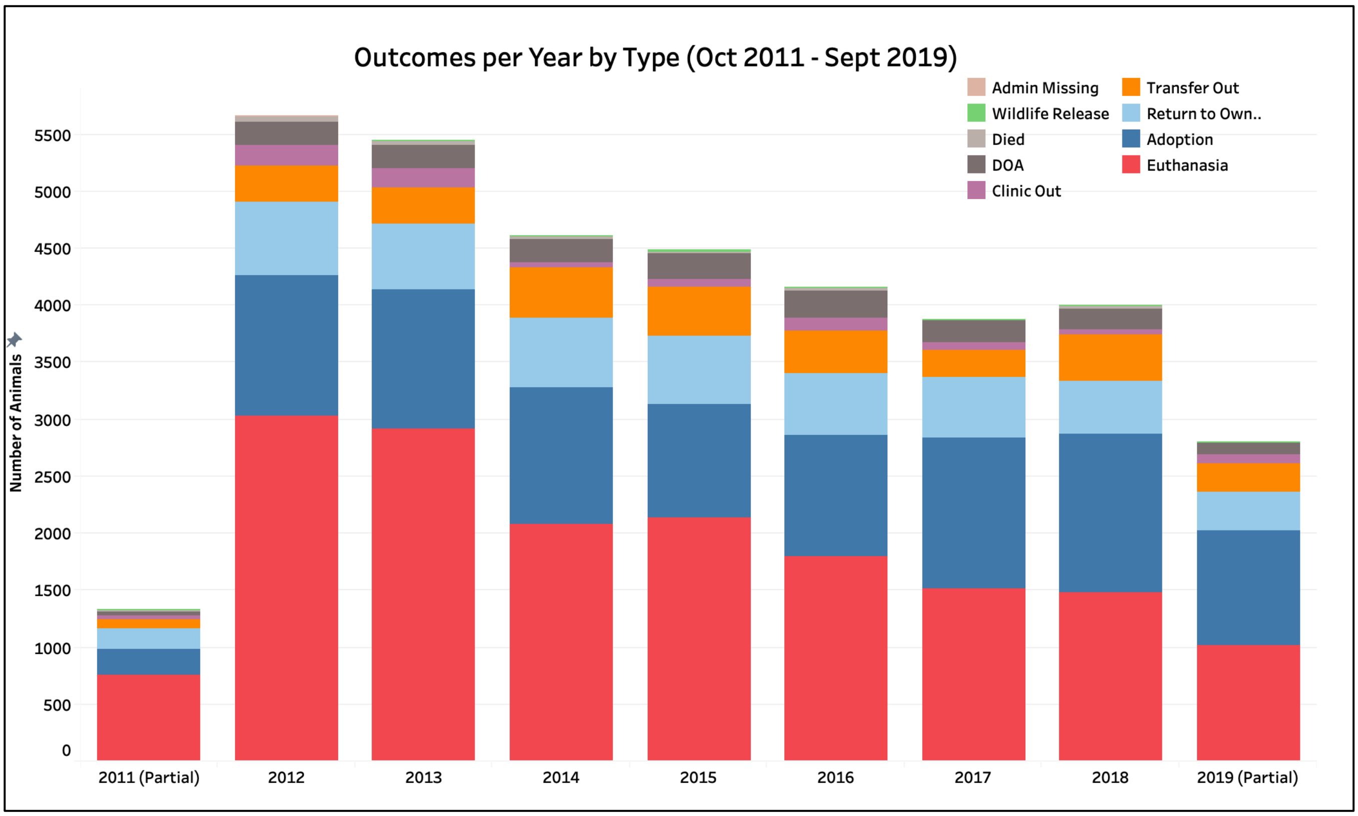
For both cats and dogs, the majority come from 21702, 21701, and 21703, which correspond to cities of Frederick. This makes sense since the shelter is located in the 21702 zip-code, so it is expected that most of the animals are coming from the surrounding area. However, 21788 (Thurmont), is a top zip-code for cats. In fact, Thurmont is the second largest city for cat strays and the number one largest city for seized cats. Thurmont is a very rural area of Frederick County with many farms, which could be a contributing factor to the number of cats that come from this area. A recommendation is to implement spay/neuter programs in Thurmont to help control the number of cats.

4.2.2 Outcomes

4.2.2.1 Trends in animal outcomes over time

Figure 5 portrays the animal outcomes per year broken out by the type of outcome. This chart shows that the number of euthanized animals has decreased between 2012-2018, and adoptions have increased. In 2018, the total number of adoptions and returns to owner outweighs the number of euthanized animals, which is a very hopeful trend for animals at FCAC.

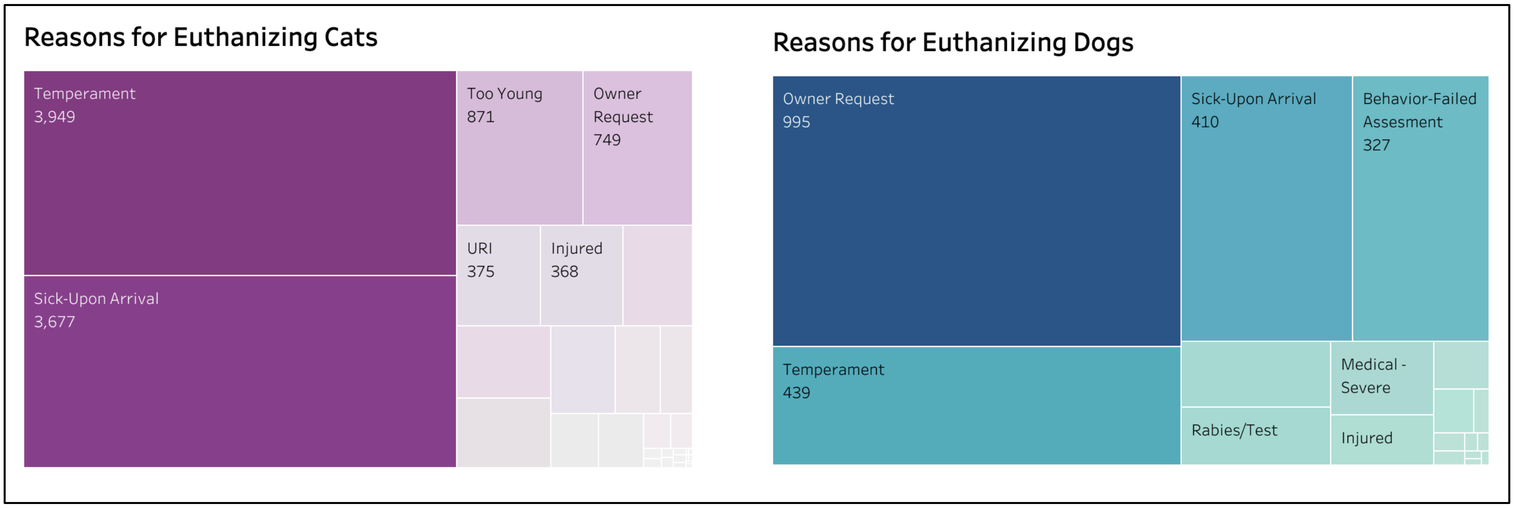
Figure 5. Outcomes per year by type



4.2.2.2 Reasons for euthanizing animals

Euthanasia might not always be avoidable for some animals. Figure 6 depicts the top reasons for euthanizing dogs and cats.

Figure 6. Top reasons for euthanasia

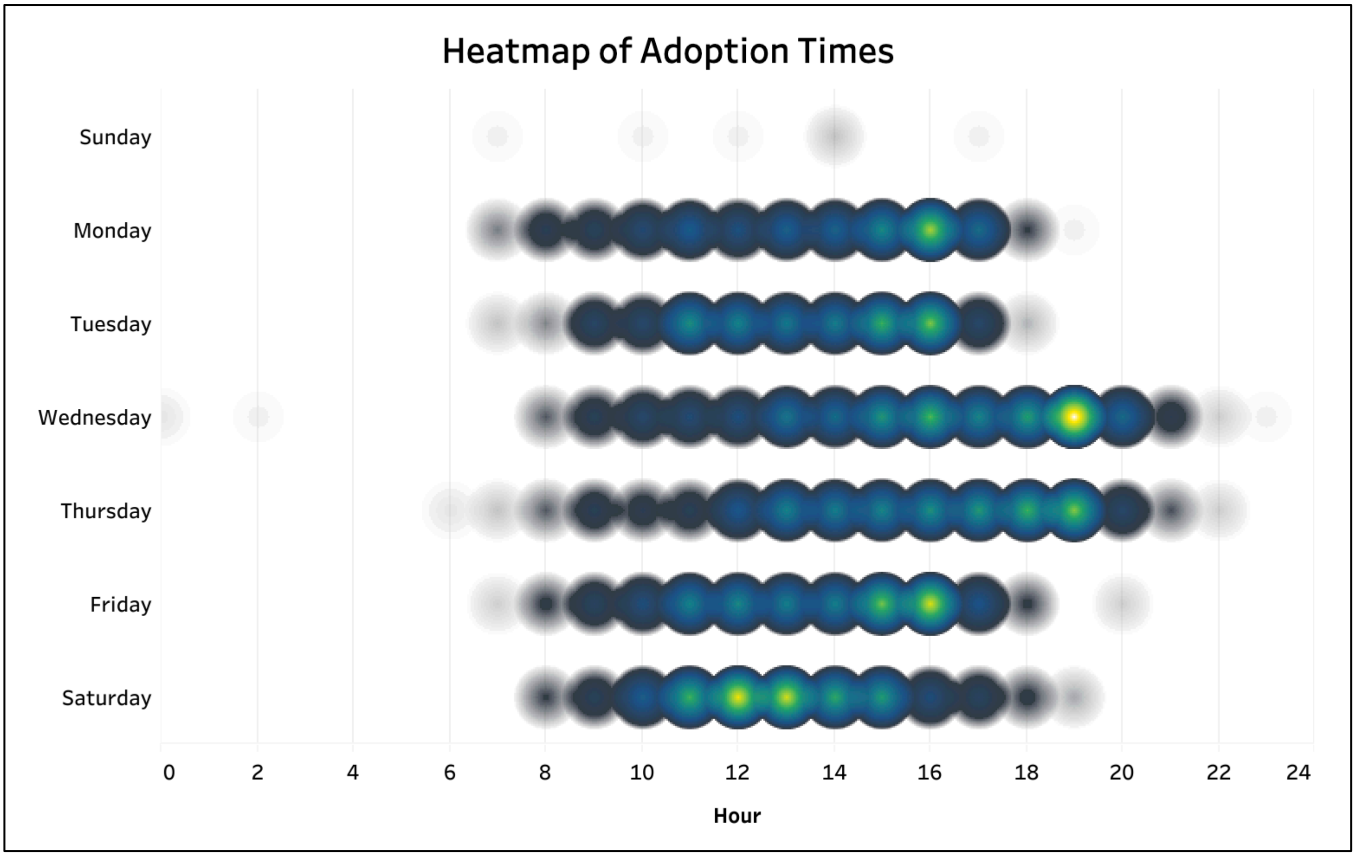


For cats, the top two reasons for euthanasia are temperament and sick upon arrival. For dogs, the top two reasons for euthanasia are owner request and temperament. The cases where "owner request" is the reason might be harder for a machine learning model to predict, since we do not know the underlying reason for the request from the data. As mentioned above, it is important to remember that euthanasia might be the best option for an animal if they are very sick or very dangerous.

4.2.2.3 Analyzing adoption times

Adoptions may happen throughout any time of day, but when do they happen the most? Answering this question can help FCAC determine when to schedule ideal adoption times, determine when to have the most staff present, or even determine when to extend their hours. Figure 7 is a heat map that displays the day of week on the y-axis and the hour of day on the x-axis. The yellow color shows that the most adoptions happen during these times, and the darker colors (blue/black) mean that less adoptions happen during these times.

Figure 7. Heatmap of adoption times



The heatmap shows that weekday evenings, specifically Wednesdays at around 5:00 PM, are when the most adoptions occur. Additionally, Saturdays between 11:00 AM and 1:00 PM are popular as well. This makes sense as most people that are working cannot make appointments until later in the day on weekdays, and they have more free time during the weekend.

4.2.2.3 Analyzing time spent at FCAC

Another goal involved examining trends in the amount of time different groups of animals spend at the shelter. First, we will examine the distribution of time spent in the shelter for the top three outcomes: adoption, return to owner, and euthanasia for dogs and cats. The distributions are depicted in the box and whisker plots below and the statistics are depicted in the tables below.

Figure 8. Distribution of time spent in shelter by outcome

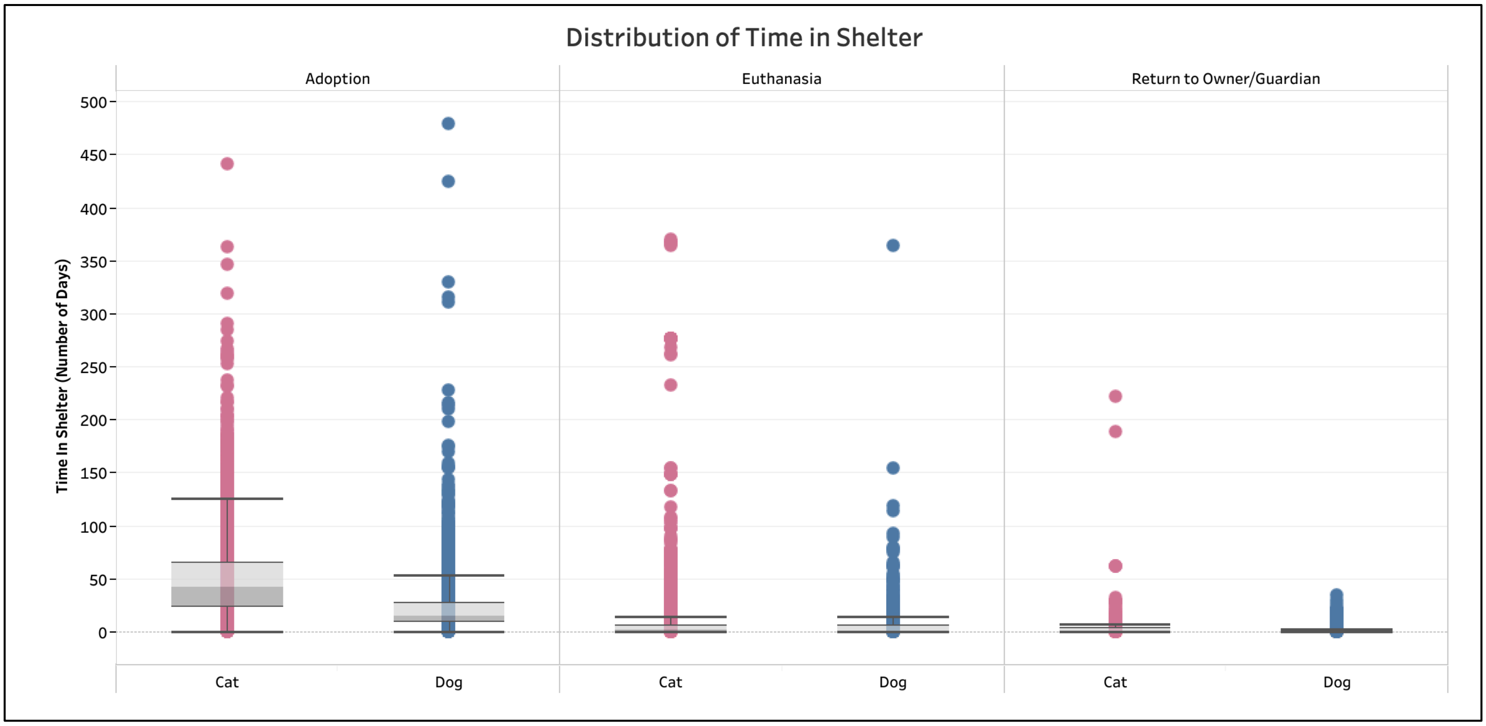


Table 5. Time spent in shelter statistics (cats)

Cats			
Outcome	**Adoption**	**Euthanasia**	**Return to Owner**
Minimum	Less than a day	Less than a day	Less than a day
Lower quartile	24 days	Less than a day	Less than a day
Median	42 days	2 days	1 day
Upper Quartile	65 days	6 days	3 days
Maximum	442 days	370 days	222 days
Average	49 days	5 days	4 days

Table 6. Time spent in shelter statistics (dogs)

Dogs			
Outcome	**Adoption**	**Euthanasia**	**Return to Owner**
Minimum	Less than a day	Less than a day	Less than a day
Lower quartile	9 days	Less than a day	Less than a day
Median	15 days	1 day	Less than a day
Upper Quartile	27 days	6 days	1 day
Maximum	480 days	365 days	35 days
Average	22 days	4 days	1 day

There are many interesting take-aways that can be derived from this distribution analysis. For both dogs and cats, there is a much larger spread in the time in shelter for adoptions. Additionally, it is clear that cats typically wait longer to be adopted than dogs, as the median time for cats is 42 days whereas it is only 15 days for dogs. A recommendation is to promote cats more often on social media in order to potentially encourage adoptions.

Additionally, for both dogs and cats, euthanasia happens fairly quickly. Although there are some outliers that wait months before they are euthanized, the majority wait one or two days. This suggests that the decision to euthanize happens very quickly for dogs and cats, which could cause additional difficulties in attempting to predict an animal's outcome. Lastly, lost dogs and cats are returned to their owners very quickly at FCAC, with the median being less than a day for dogs and one day for cats.

4.2.2.4 Relationship between age group and time until adoption

Next, we can assess if the age of the animal impacts how long an animal waits to be adopted. The box and whisker plots and tables with statistics are displayed below. (Note that the less than eight-week age group was removed since there were so few up for adoption.)

Figure 9. Distribution of time spent at FCAC waiting for adoptions by age group

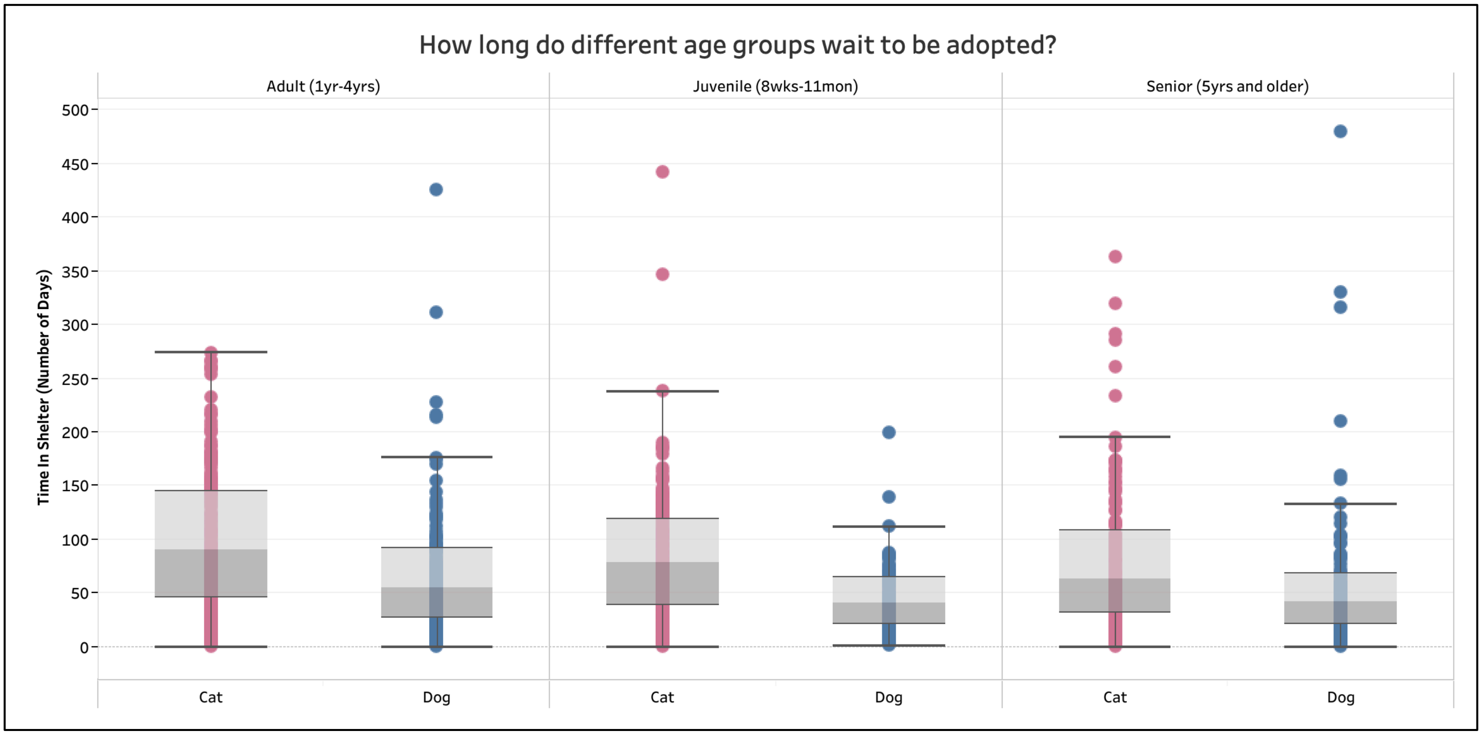


Table 7. Age groups waiting for adoption statistics (cats)

Cats			
Age Group	**Adult (1yr-4yrs)**	**Juvenile (8wks-11mon)**	**Senior (5yrs and older)**
Minimum	Less than a day	Less than a day	Less than a day
Lower quartile	46 days	39 days	32 days
Median	91 days	79 days	64 days
Upper Quartile	145 days	119 days	109 days
Maximum	274 days	442 days	363 days
Average	49 days	49 days	48 days

Table 8. Age groups waiting for adoption statistics (dogs)

Dogs			
Age Group	**Adult (1yr-4yrs)**	**Juvenile (8wks-11mon)**	**Senior (5yrs and older)**
Minimum	Less than a day	Less than a day	Less than a day
Lower quartile	27 days	21 days	21 days
Median	55 days	41 days	42 days
Upper Quartile	92 days	65 days	69 days
Maximum	425 days	199 days	480 days
Average	23 days	21 days	25 days

Based on the box and whisker plots, it is again clear that there is a distinct difference in adoption times between dogs and cats. However, it is not clear that age group makes a significant difference. A hypothesis test was conducted to see if there is a significant difference in adoption times between age groups. The Kruskal Wallis statistical hypothesis test is used to determine if more than two groups have different distributions. It does not require that the data is normally distributed, which is why it was used. The null hypothesis is that the distributions are the same, whereas the alternate hypothesis is that at least one of the distributions is different.

For cats, the chi-squared statistic was 44.04 with a p-value of 2.73e-10. For dogs, the chi-squared statistic was 24.70 with a p-value of 4.32e-06. For both dogs and cats, the null hypothesis is rejected at an alpha level of 0.05. This suggests that at least one of the distributions is different.

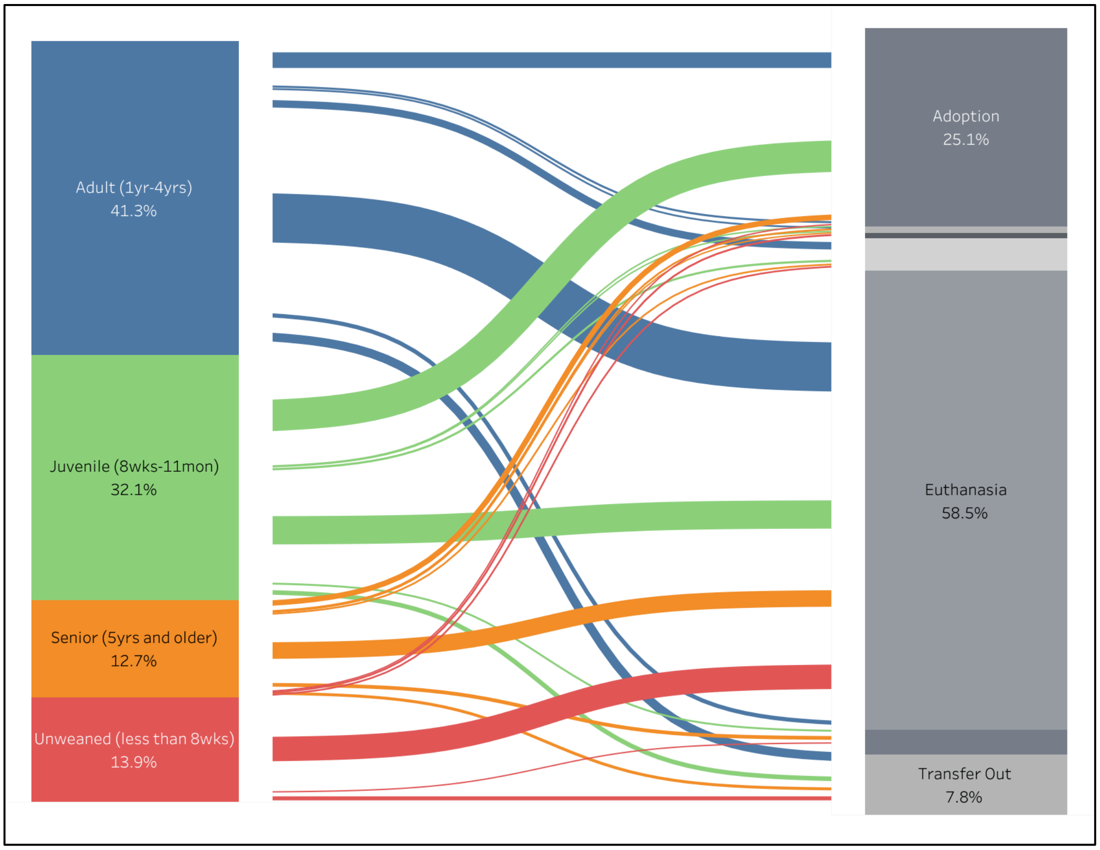
For cats, seniors seem to be adopted the fastest with a median of 64 days. For dogs, both juveniles and seniors are adopted at a median of ~41 days. However, juvenile dogs appear to be adopted the quickest overall.

4.2.2.5 Relationship of age group and outcome

Next, we can look at the relationship between the age group of the animal and the outcome. This can answer questions like: “Do most puppies/kittens get adopted?” or “Are most senior cats/dogs euthanized?” The plots below are called Sankey plots, and can be interpreted the same way as a stacked bar chart. On the left side of the chart, there is a breakout by age group. On the right side is a breakout by shelter outcome. The lines in the middle show the relationship between age group and outcome; the thicker the line, the larger the number of animals.

The relationship between age group and outcome is shown in Figure 10 for cats.

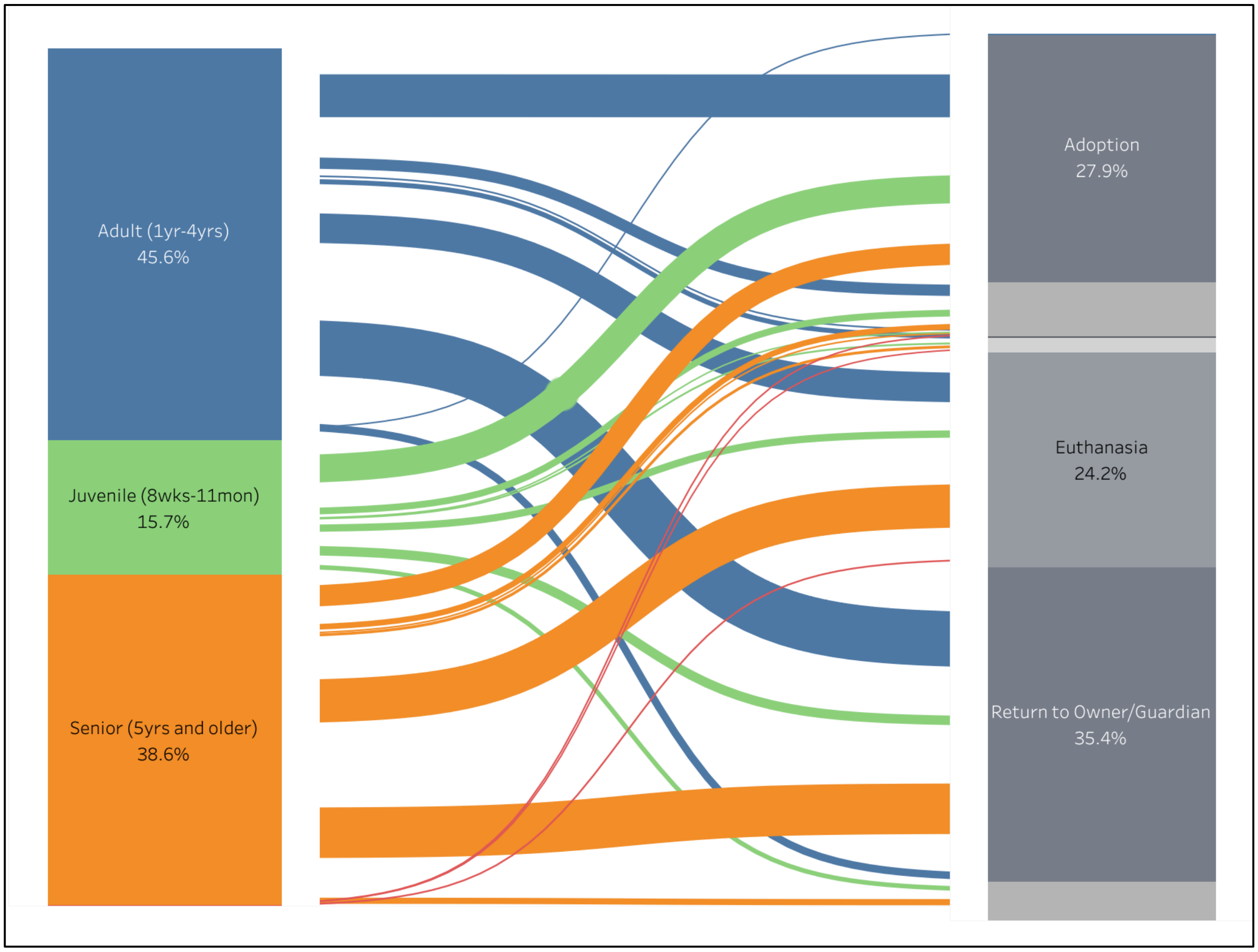
Figure 10. Relationship between age group and outcome for cats



There are a few key take-aways from this plot. For instance, the majority of adoptions are comprised of juvenile cats, followed by adult cats. Additionally, the majority of euthanized cats are adults, but almost all of the less than eight-week-old kittens are euthanized. Another important thing to note is that very few cats are actually returned to their owners. (Not visible in the chart above.)

The relationship between age group and outcome for dogs is shown in Figure 11 below.

Figure 11. Relationship between age group and outcome for dogs



From this plot, it is clear that the majority of dogs returned to their owner are seniors and adults. Additionally, seniors make up a large proportion of euthanized dogs. (Senior dogs are often requested to be euthanized by their owners for health reasons.) Lastly, adoptions are largely comprised of adults, but juveniles make up a large proportion as well.

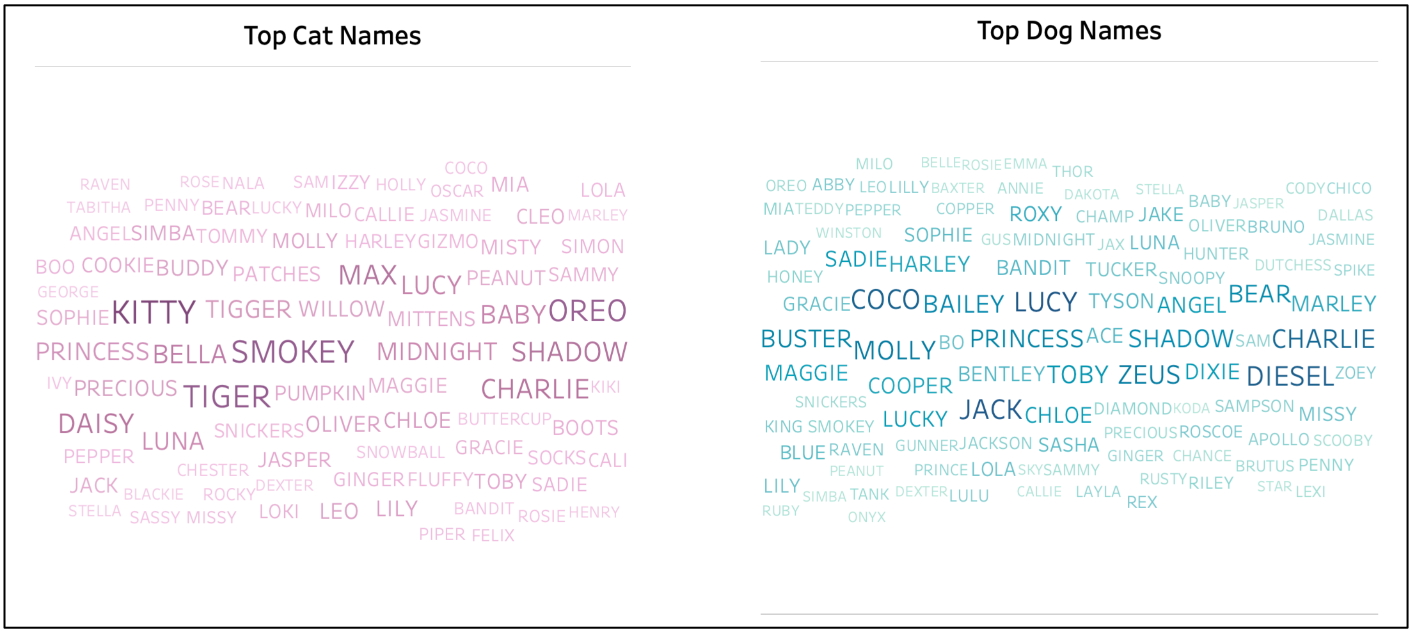
These key trends suggest that age group will be a powerful feature when it comes to predicting an animal's outcome.

4.2.2.6 Most "appealing" names and breeds

In this section, we will look at the most common names and breeds in the dataset and identify what names are most successful, where success is defined as either an adoption or return to owner.

First, we will look at the most common dog and cat names in the entire dataset, as shown in the word clouds below.

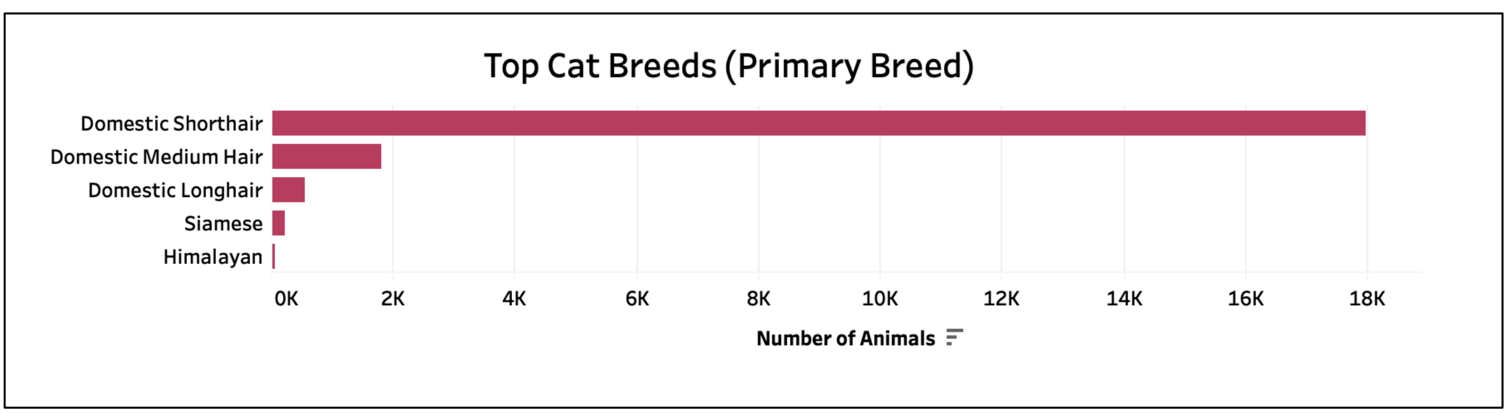
Figure 12. Word clouds of animal names



The top names for cats are Kitty, Smokey, and Tiger. The top names for dogs are Lucy, Jack, and Diesel. Of the top 25 top cat names, the most "successful" names were Willow, Oliver, and Leo. Cats named Willow were either adopted or returned to their owner 86% of the time. Of the top 25 top dog names, the most "successful" names were Lucy, Charlie, and Dixie. Dogs named Lucy were either adopted or returned to their owner 78% of the time.

Next, we will assess the top breeds and the ones that are most appealing. For the purposes of this analysis, the "primary breed" field is used. Note that these may or may not be purebreds. The top five primary breeds for cats are shown in Figure 13, since cats are almost always labeled as domestic shorthair. Domestic shorthair cats make up around 87% of the dataset.

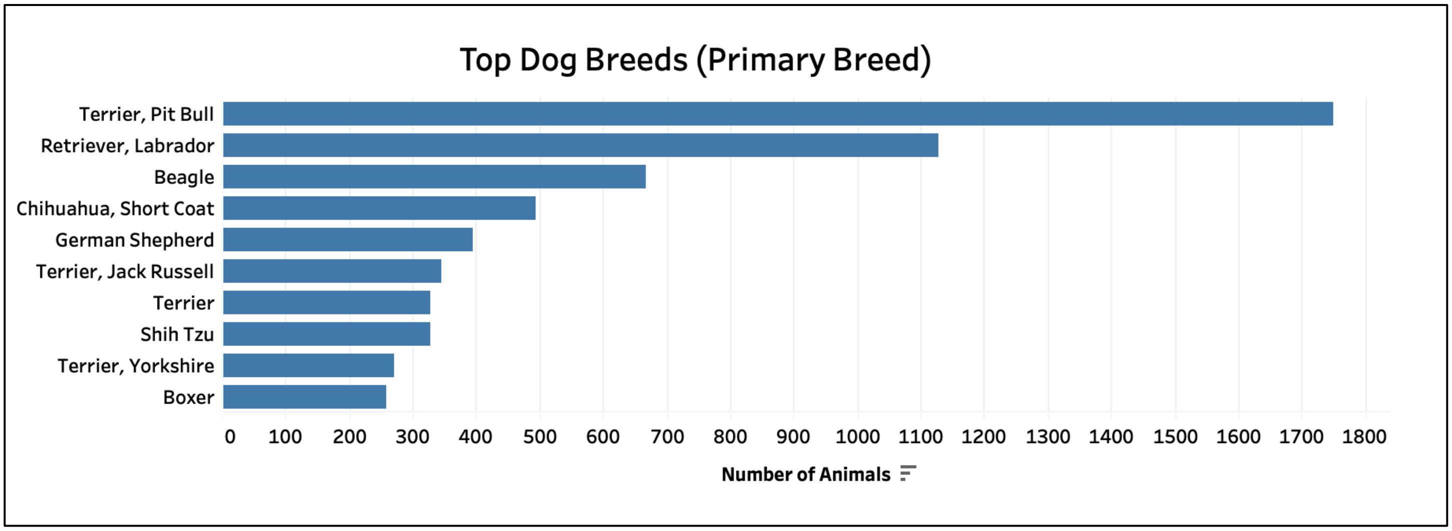
Figure 13. Top cat breeds



Are there certain breeds that are more likely to be adopted or returned to their owner? Considering breeds with more than 20 records in the data, the most successful primary breed for cats is the Maine Coon (not shown in top five above). However, this breed was only adopted or returned to its owner 54% of the time. If we look at breeds that are very rare, such as British Shorthair, Turkish Van, Havana Brown, etc., these were returned/adopted 100% of the time. This suggests that rare breeds might be more desirable, whereas very common mixed cat breeds are not.

The top 10 primary breeds for dogs are shown below. Pit bull is the most common primary breed, followed by Labrador retrievers.

Figure 14. Top dog breeds



Since there is a larger spread in dog breeds than cat breeds, we will consider breeds with more than 100 records. Based on the data, the most "successful" primary breeds for dogs are the Siberian husky, hound dog, and beagle. These breeds were either adopted or returned to their owner over 70% of the time. Similar to cats, very rare dog breeds such as coonhounds, Finnish Spitz, Carolina dog, etc. were almost always adopted/returned.

4.3 Machine Learning

4.3.1 Assumptions and hypotheses

As previously stated, the major goal of the project is to provide extensive analysis of FCAC's data and to determine if machine learning techniques can be used to predict an animal's outcome at FCAC. Specifically, machine learning was used to identify the factors that have the most impact on these outcomes, and thus have the most predictive power. This section will focus on the machine learning portion of the project.

Dogs and cats make up the majority of the dataset, therefore they were the focus of this effort. Separate models were created for cats and dogs in order to identify similarities and differences in the most predictive features. Additionally, there are around eight different outcomes a dog/cat could have in the dataset: adoption, return to owner, euthanasia, transfer, clinic out, DOA (death on arrival), or died. The majority of the animals are either adopted, returned to their owner, or euthanized (13% of dogs and 13% of cats had other outcomes). The primary goal of this project is to understand why some animals get adopted/returned and some do not. Because of this, the model will attempt to predict a "successful" outcome, where success is defined as either an adoption or return to owner, versus the euthanasia outcome. This is an important assumption to keep in mind since the model is assuming only two possible outcomes (e.g. a binary problem), which are euthanasia vs. adoption/return. Although we are referring to adoptions/returns as successful outcomes, it is important to note that sometimes, euthanasia cannot be avoided and is often the best solution for the animal (e.g. dangerous, sick, etc.).

Based on the exploratory analysis and the related work by Ms. Joanne Lin, some features were hypothesized to be more predictive than others. These features include: the age group, whether or not the animal had a name, whether or not the animal was spayed or neutered, and the breed. However, there was an issue with the "has name" and "spayed neutered" features, as shown in the figures below.

Figure 15. Relationship of spayed/neutered to target variable

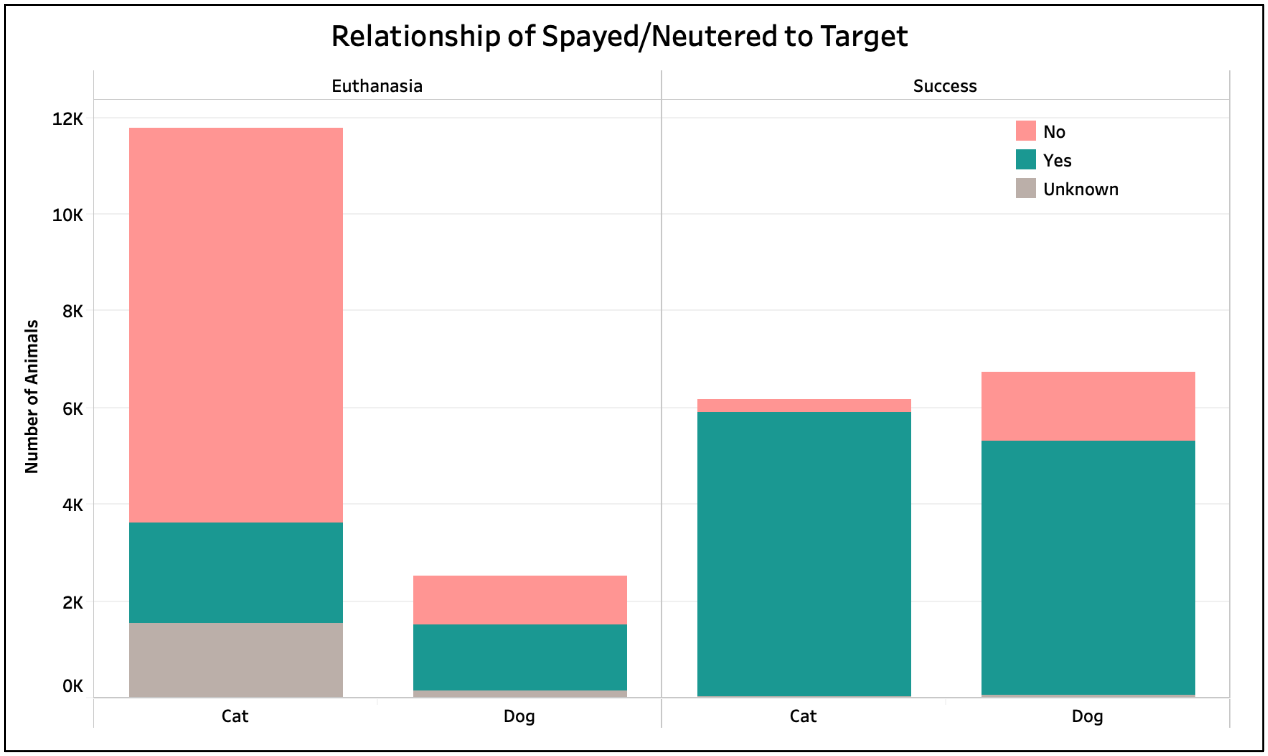
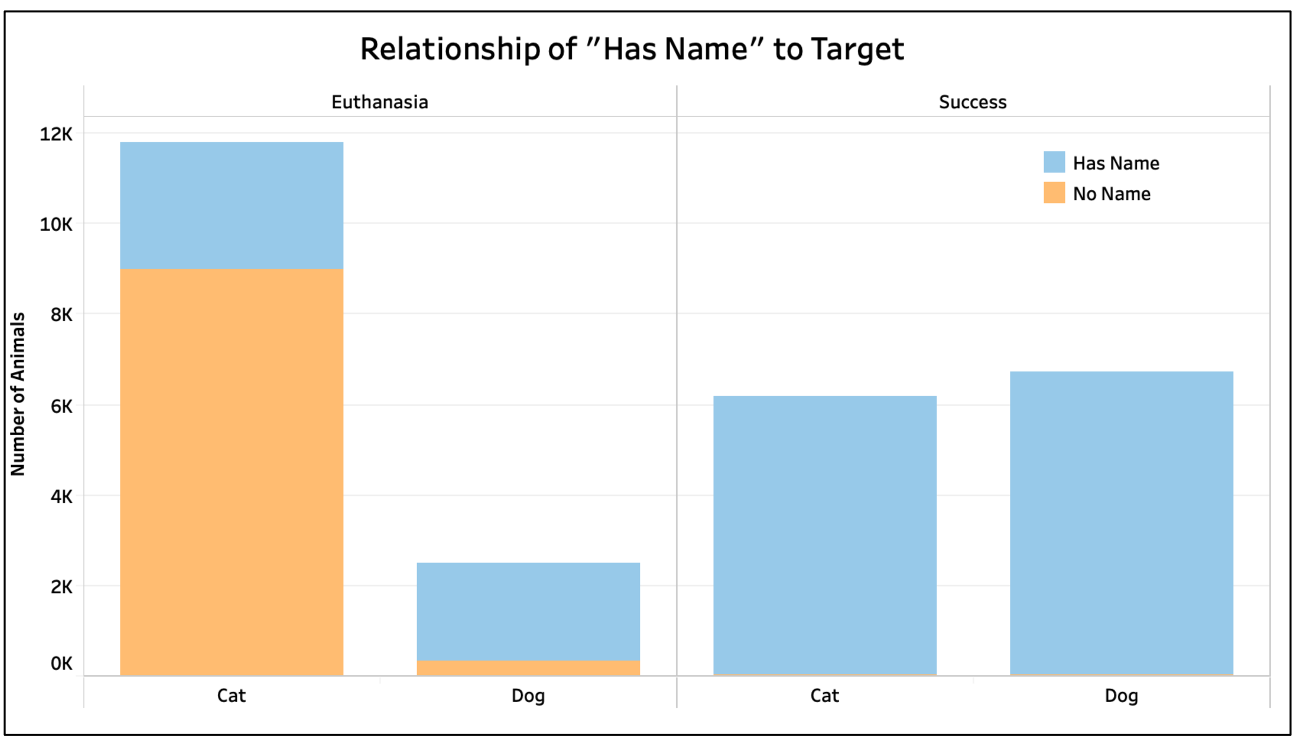


Figure 16. Relationship of "has name" to target variable



It turns out that at FCAC, every animal appropriate for adoption gets spayed/neutered. Additionally, every animal that is up for adoption is given a name, and it is likely that if an animal got lost and their owner is looking for it, they had a name as well. One could say that the animal wasn't given a name or wasn't spayed/neutered because it was euthanized. Because these features could potentially lead to bias in the model, they were not used. Any feature that involved the name of the animal was removed. Fortunately, the pre-altered feature could be used in place of the spayed/neutered feature. This feature refers to if an animal was altered in some way when they were taken in, which could include spayed/neutered etc. The final features that were used in the models for dogs and cats will be discussed below.

4.3.2 Experimental plan

Since the outcome is defined in the dataset, this is a supervised machine learning problem. When choosing a machine learning algorithm to use, there is always a trade-off between performance and explain-ability to keep in mind. Being able to explain the models and the features that contribute to the predicted outcomes was the most important factor in this case. Because of this, three different supervised machine learning algorithms were used: Random Forest, Gradient Boosting via XGBoost, and Logistic Regression (using Lasso regularization). Each algorithm will be briefly explained at a high level in the sections below.

For model selection, k-fold cross validation with a grid parameter search was used. K-fold cross validation is a method used to avoid overfitting and bias in a machine learning model. In this process, the data set is split into different sections (e.g. 5 folds). During the first iteration, the first fold used to test the model's performance, and the remaining data is used to train the model. During the second iteration, the second fold is used to test the model's performance and the remaining data is used to train the model. This is repeated until K number of folds have been used as the test set. This process ensures that every observation in the data set will show up in the test or train set. Grid parameter search was used in combination with K-fold cross validation. Every machine learning model has different hyper-parameters that can be tuned in order to improve performance. Grid search is a method used to iterate through different combination of values of hyper-parameters in order to optimize performance. The data was split into a train and validation set; the train set was used in K-fold cross validation, and the validation set will be used as a final set that was never seen in training the model in order to test the performance.

There are many different metrics that can be used to assess the performance of each model. The most common metric is accuracy, which refers to the proportion of correct predictions out of the total number of observations. However, this metric is flawed when the classes do not have an equal number of samples. As shown in Figure 14 and 15, the two classes (adoption/return (success) vs. euthanasia) are imbalanced for both dogs and cats. Cats have a higher number of euthanasia than adoptions/returns, and dogs have a higher number of adoptions/returns than euthanasia. Because of this, other metrics will be used to assess performance. The metrics that will be used are the Area Under the Curve (AUC) and the F1 score.

AUC is specifically for binary classification problems. It is equal to the probability that a classifier will rank a randomly chosen positive example higher than a randomly chosen negative example. The AUC comes from the area under the Receiver Operating Characteristic Curve (ROC Curve), which plots the true positive rate against the false positive rate for different threshold values. AUC is essentially a measure that identifies how well a classifier can distinguish between the two outcomes. An AUC of 0.5 means the model is no better than random guessing, and an AUC of 1 is a perfect model.

The F1 score is referred to as the harmonic mean between precision and recall. Precision is the number of correct positive results divided by the total number of positive results predicted by the classifier. Recall is the number of correct positive results divided by the total number of positive samples. F1 score is a balance of these two measures, and the closer the F1 score is to 1, the better the model.

4.3.3 Additional pre-processing

The following features were used in the machine learning models:

- Intake Type
- Primary Breed
- Gender
- Primary Color
- Pre-Altered
- Size
- Condition
- Age Group
- Is Mix
- Is Black
- Is Top Dog (dog model only)
- Is Multicolor
- Breed Group (dog model only)
- Energy
- Shedding

Before implementing the machine learning algorithms, additional pre-processing was required. The Gender variable had some "unknown" values for a few of the euthanized cats; these were removed in order to avoid bias. Additionally, the categorical features needed to be converted into numerical dummy variables. A dummy variable takes on two values, 0 or 1. Typically, 1 represents the presence of an attribute and 0 represents the absence. For example, to represent gender, one would define 1 as male and 0 as non-male. There are typically k-1 dummy variables for a categorical variable that can assume k different values.

As stated above, the data was split into training and validation sets. For cats, 40% of the data was kept for validation. For dogs, 30% of the data was kept for validation since the dataset was much smaller. The table below depicts the size of each data set for cats and dogs.

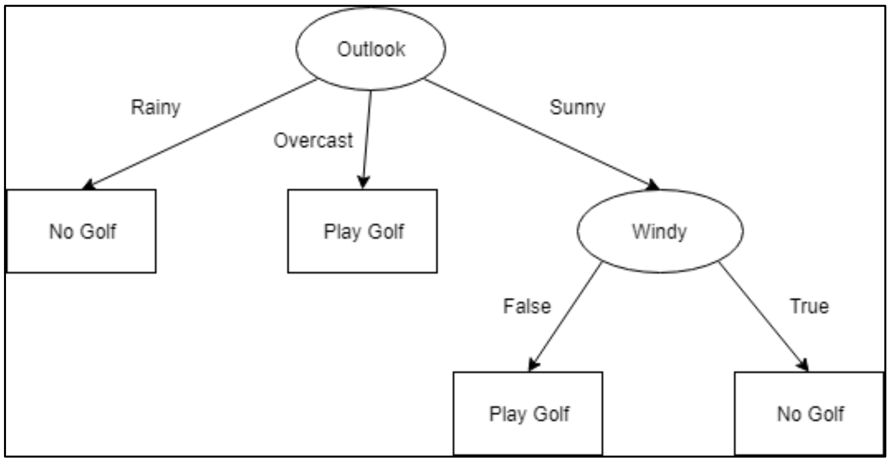
Table 9. Size of datasets

	# Adoption/Return (Success)	**# Euthanasia**	**Total**
Cat Dataset	5,959	9,807	15,757
Dog Dataset	6,510	2,418	8,928

4.3.4 Random forest

The first model that was used was a random forest. In order to understand the random forest machine learning classifier, it is important to understand a decision tree. A decision tree is a type of classification model, which is interpreted like a flow chart. It is thought of as a series of cascading questions in the form of a tree. The trees are made up of decision nodes, leaf nodes, and branches. An example of a simple decision tree is below:

Figure 17. Decision tree example



(Ganegedara, 2018)

A random forest is a little less intuitive than a simple decision tree. It is actually a collection or ensemble of decision trees, and the prediction is obtained based on the majority vote. A random forest is referred to as random because when the trees are built, they are based off of a random sample of the observations as well as a random subset of the features. This helps to avoid over-fitting the data.

All algorithms have advantages and disadvantages. The advantages of a random forest are that it is fairly straightforward, avoids overfitting, works well out of the box, works well with unbalanced classes, and provides insight into the importance of each feature. The disadvantages of a random forest are that training can take a long time and predictions are slower, but these issues are not a concern for this project.

4.3.4.1 Predicting cat outcomes

The table below shows the different parameters that were tuned during cross validation.

Table 10. Hyper-parameters for random forest

Hyper-Parameter	**Description**	**Values**
Max_depth	Maximum depth of the tree	10, 20, 30, 40
Max_features	The number of features to consider when looking for the best split	5, 10, 20
N_estimators	Number of trees in the forest	100, 200, 300, 1000

As mentioned previously, K-fold cross validation was used to identify the optimal parameters. The cat data was split into three different folds. The optimal parameters were 10, 20 and 100 for the max_depth, max_features, and n_estimators respectively. The F1 score was used as the metric to optimize during cross validation. The average macro F1 score for cross validation was 0.73. Recall that an F1 score close to 1 is ideal.

Once the optimal parameters were found, the model was tested on the validation set that was held out in the beginning of the process. The figures below show the classification report and confusion matrix. The classification report shows various performance metrics for the model, and the confusion matrix shows the number of correctly and incorrectly classified observations. **(Note that class 0 = Euthanasia and class 1 = Adoption/Return.)**

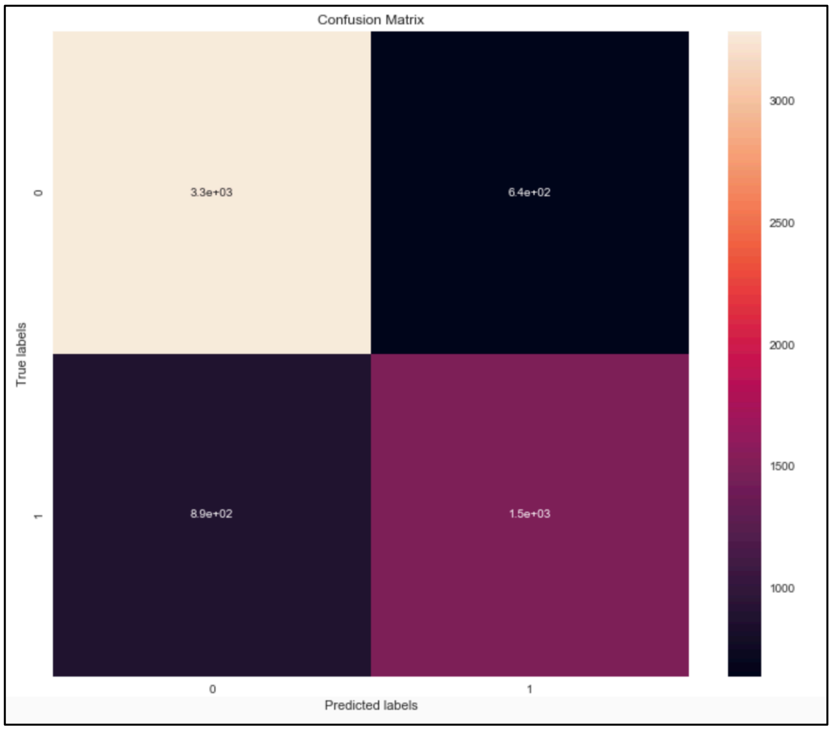
Figure 18. Classification report (cat random forest)

```
=== Classification Report ===
              precision    recall  f1-score   support

           0       0.79      0.84      0.81      3923
           1       0.70      0.63      0.66      2380

    accuracy                           0.76      6303
   macro avg       0.74      0.73      0.74      6303
weighted avg       0.75      0.76      0.76      6303
```

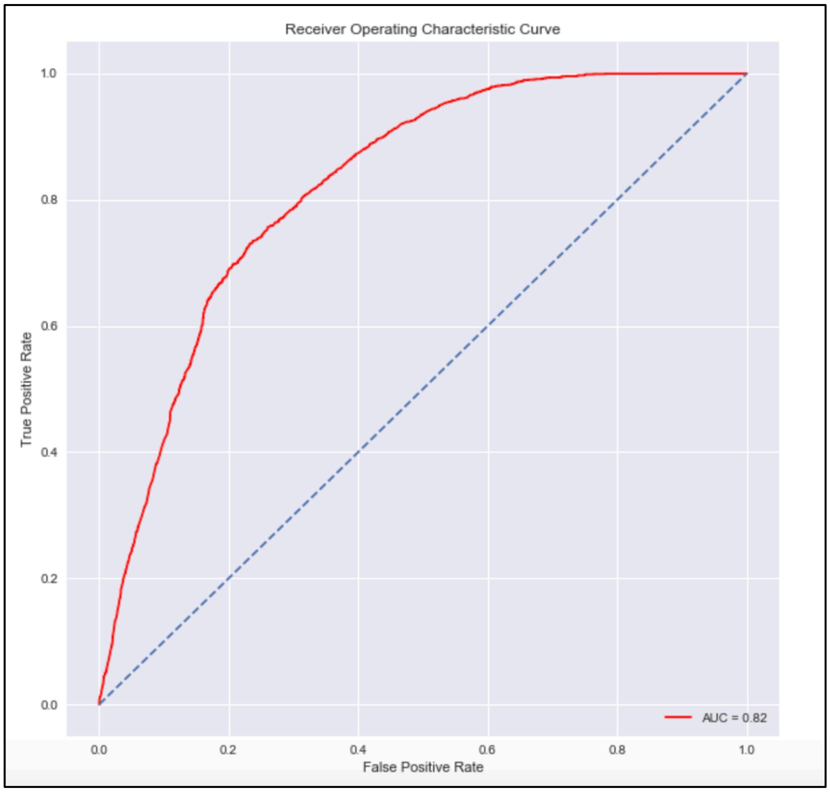
Figure 19. Confusion matrix (cat random forest)



The model is much better at predicting euthanasia than it is at adoptions/returns. This makes sense since the cat dataset has many more euthanasia observations than adoptions/returns. The F1 score for the euthanasia class (class 0) is 0.81, which is very close to 1. However, the F1 score for the adoptions/returns class (class 1) is 0.66. This appears to be because recall is fairly low. Consequently, the macro F1 score is 0.74 and the accuracy is 0.76.

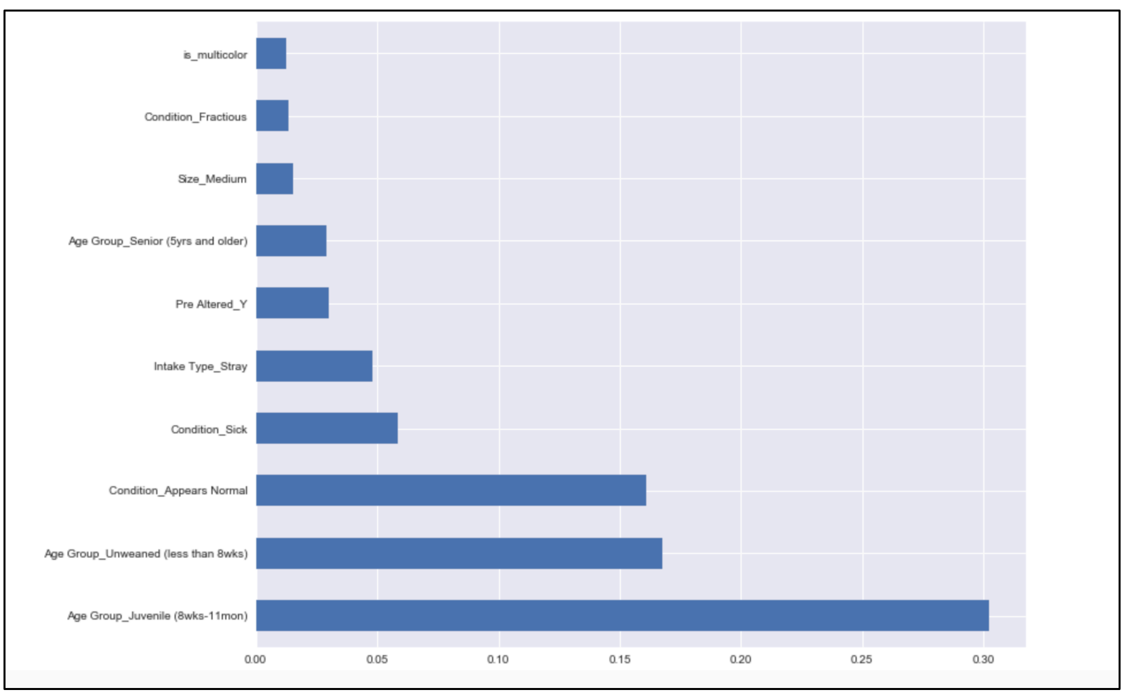
Below is the ROC curve, which shows the curve in red and a reference curve in dotted blue for a "random guessing" model. The resulting AUC is 0.82, which is close to 1, so the model is fairly good at distinguishing between the two classes.

Figure 20. ROC curve (cat random forest)



Lastly, we can examine the feature importance plot. Feature importance helps us understand the most impactful features that define the predictions. In a random forest model, the most important features are the ones that are best at partitioning the data so that a prediction can eventually be made.

Figure 21. Feature importance plot (cat random forest)



The most important features for predicting a cat's outcome are the age of the cat, the condition the cat is in (e.g. "appears normal", sick, etc.), the intake type (e.g. stray), and if it was pre-altered.

4.3.4.2 Predicting dog outcomes

Next, we used a random forest to predict dog outcomes. The same hyper-parameters that were tuned in the cat model were also tuned in the dog model.

The dog data was also split into three different folds during cross validation. The optimal parameters were 20, 20, and 1000 for the max_depth, max_features, and n_estimators respectively. Again, the F1 score was used as the metric to optimize during cross validation. The average macro F1 score during cross validation was 0.76.

Once the optimal parameters were found, the model was tested on the validation set that was held out in the beginning of the process. The figures below show the classification report and confusion matrix. **(Note that class 0 = Euthanasia and class 1 = Adoption/Return.)**

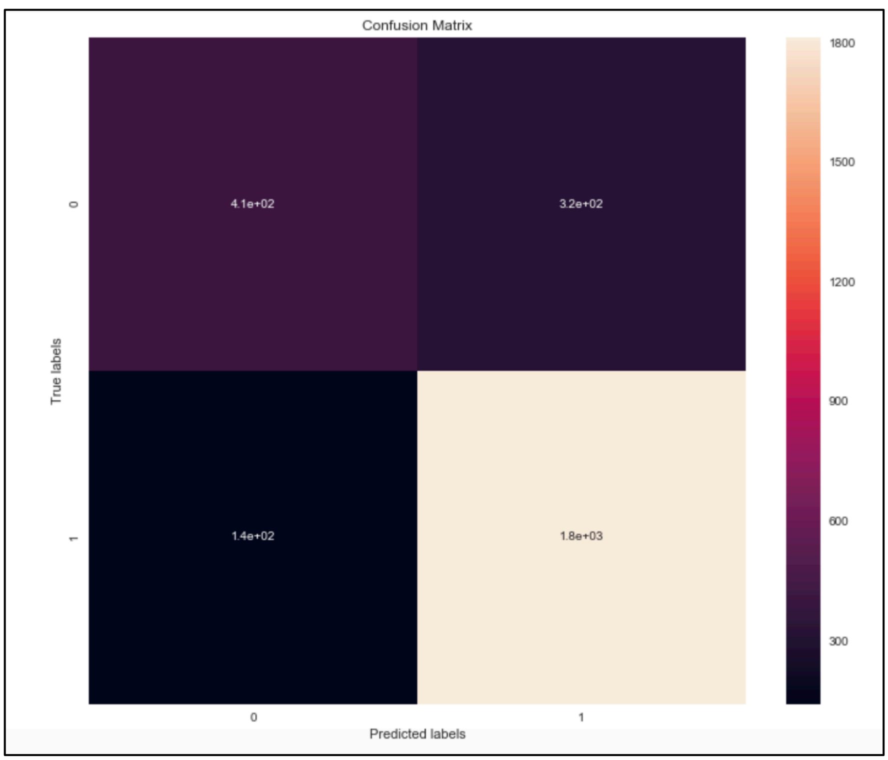
Figure 22. Classification report (dog random forest)

```
=== Classification Report ===
              precision    recall  f1-score   support

           0       0.74      0.56      0.64       726
           1       0.85      0.93      0.89      1953

    accuracy                           0.83      2679
   macro avg       0.80      0.74      0.76      2679
weighted avg       0.82      0.83      0.82      2679
```

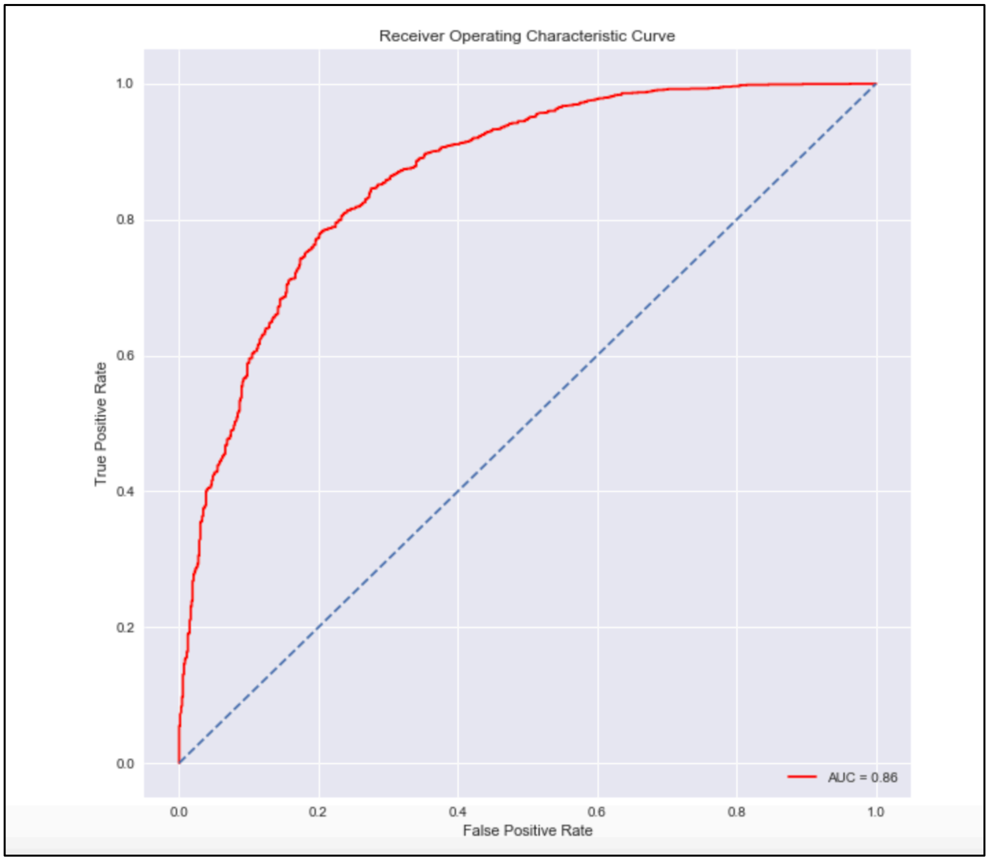
Figure 23. Confusion matrix (dog random forest)



Recall that the cat random forest was much better at predicting euthanasia than it was at prediction adoptions/returns. We see that the opposite is true for the dog model, which makes sense since there are more adoption/return observations in this dataset. The F1 score for the euthanasia class (class 0) is 0.64, which is not ideal. This is due to the low recall of 0.56. However, the F1 score for the adoptions/returns class (class 1) is 0.89, which is very close to 1. Consequently, the macro F1 score is 0.76 and the accuracy is 0.83.

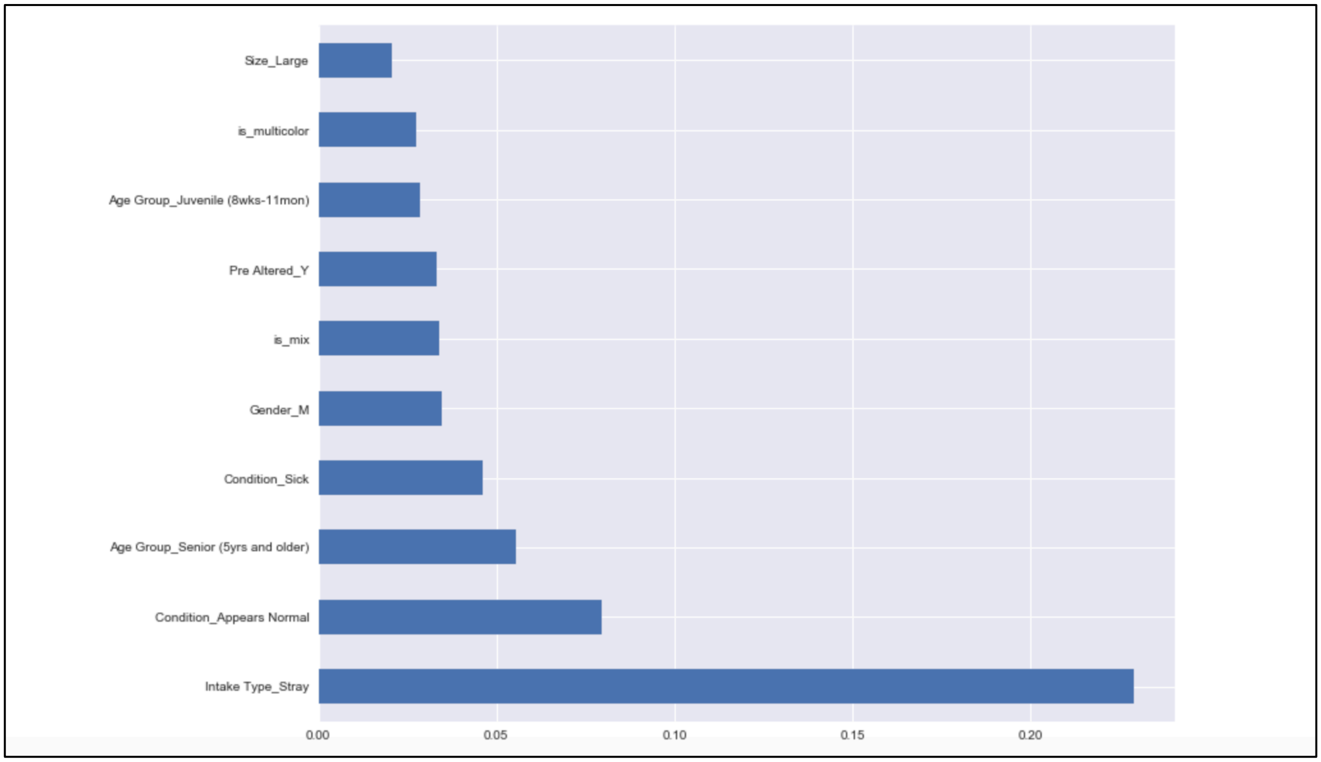
Below is the ROC curve, which shows the curve in red and a reference curve in dotted blue for a "random guessing" model. The resulting AUC is 0.86, which is close to 1, so the model is fairly good at distinguishing between the two classes.

Figure 24. ROC curve (dog random forest)



Like we did for the cat model, we can examine the feature importance plot below.

Figure 25. Feature importance plot (dog random forest)



The most important features for predicting a dog's outcome are the intake type (e.g. stray), the condition the dog is in (e.g. "appears normal", sick, etc.), age of the dog, and the gender.

4.3.5 Gradient boosting

The next algorithm used was gradient boosting via the XGBoost Python package. XGBoost is very popular in data science competitions like Kaggle because of its speed and ability to produce models with strong performance. XGBoost is another decision tree-based ensemble machine learning algorithm like a random forest. Gradient boosting trains models in a sequential manner; the idea is that subsequent models learn from the mistakes of the previous models. The goal is to minimize the error of the overall model.

The advantages of XGBoost are similar to those of a random forest. It provides insight into important features, it is fast, provides good model performance, it is less prone to overfitting, and it is not sensitive to outliers. Its disadvantages are that it is not as easy to interpret the algorithm and it is harder to tune since there are more hyper-parameters.

4.3.5.1 Predicting cat outcomes

The table below shows the different parameters that were tried during cross validation with XGBoost.

Table 11. Hyper-parameters for random forest

Hyper-Parameter	**Description**	**Values**
Objective	logistic regression for binary classification, returns a predicted probability	Binary:logistic
Learning_rate	Step size shrinkage to prevent overfitting	.01, 0.05, .1
Max_depth	Maximum depth of a tree	5, 10, 20
Colsample_bytree	Fraction of the features to be randomly sampled for each tree	0.7
n_estimators	Number of trees to use	5, 50, 100, 1000

Similar to the random forest, K-fold cross validation was used to identify the optimal parameters. The cat data was split into three different folds. The optimal parameters for the learning rate, max depth, and number of estimators were 0.01, 20, and 100 respectively. The F1 score was again used as the metric to optimize during cross validation. The average macro F1 score for cross validation was 0.73, which was the same as the random forest.

Once the optimal parameters were found, the model was tested on the validation set that was held out in the beginning of the process. The figures below show the classification report and confusion matrix. **(Note that class 0 = Euthanasia and class 1 = Adoption/Return.)**

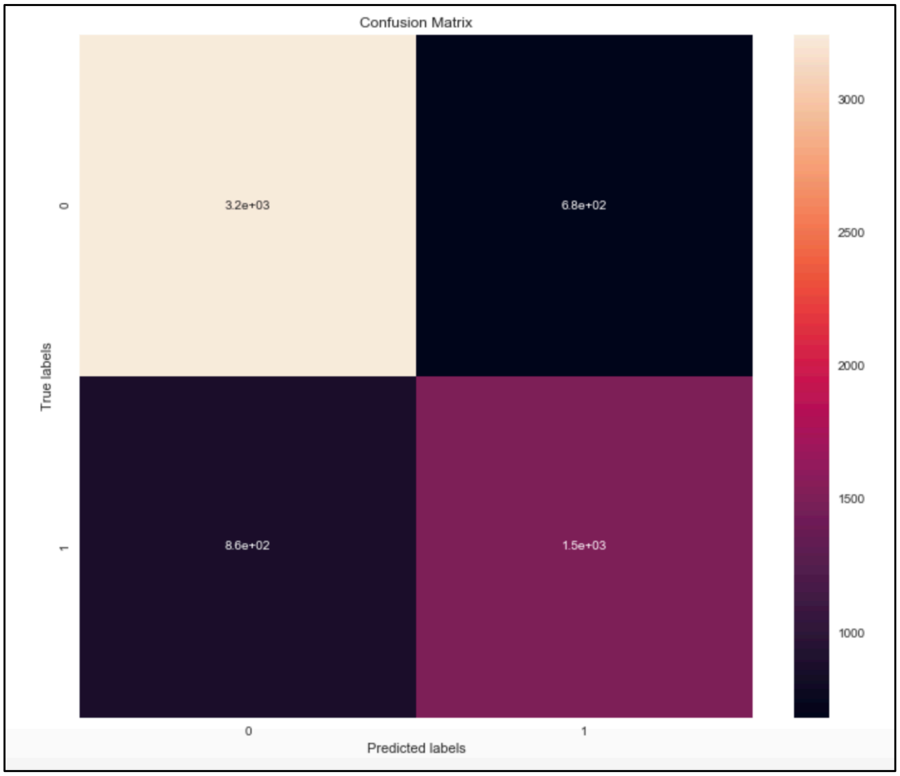
Figure 26. Classification report (cat xgboost)

```
=== Classification Report ===
              precision    recall  f1-score   support

           0       0.79      0.83      0.81      3923
           1       0.69      0.64      0.66      2380

    accuracy                           0.76      6303
   macro avg       0.74      0.73      0.74      6303
weighted avg       0.75      0.76      0.75      6303
```

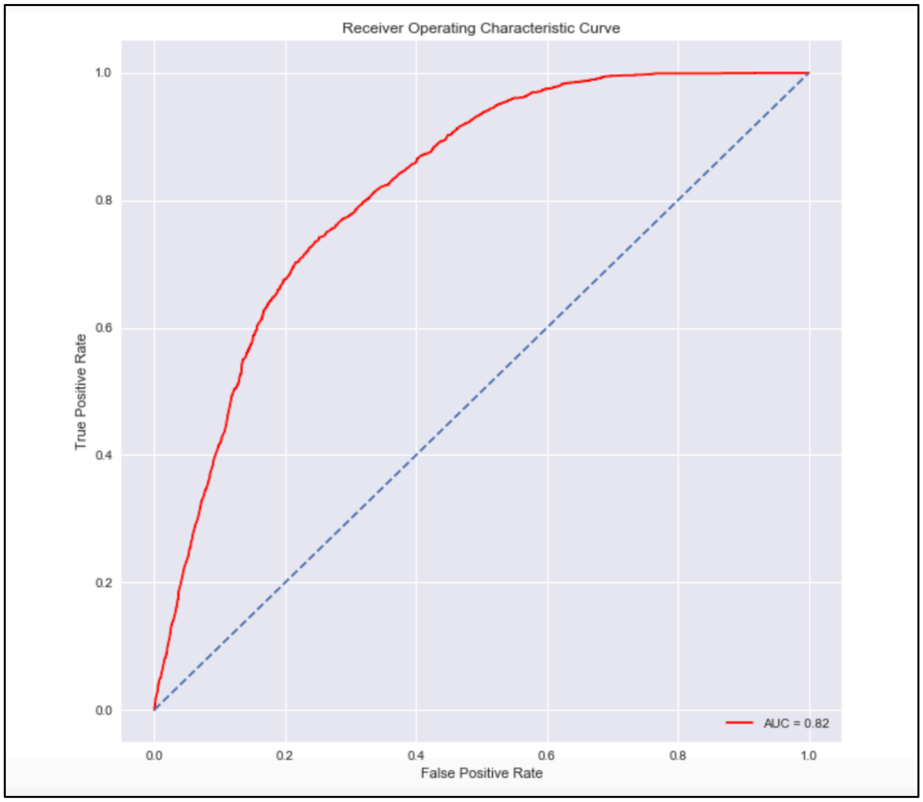
Figure 27. Confusion matrix (cat xgboost)



The classification report/confusion matrix looks almost identical to the random forest. A minor change was a decrease in 0.01 in precision and an increase in 0.01 in recall for the adoption/return class. Again, we see that the model is much better at predicting euthanasia than it is at adoptions/returns. The F1 score for the euthanasia class (class 0) is 0.81, and the F1 score for the adoptions/returns class (class 1) is 0.66. Consequently, the macro F1 score is 0.74 and the accuracy is 0.76. This is about identical to the random forest, so XGBoost did not significantly improve the model.

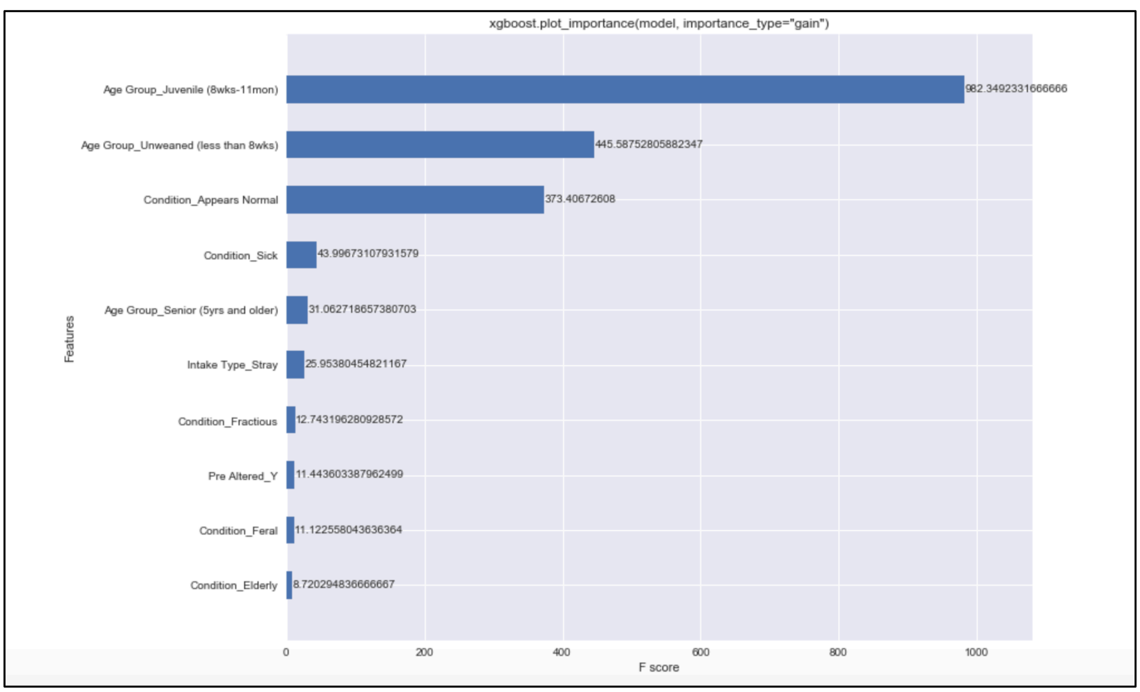
Below is the ROC curve, showing an AUC of 0.82. This is again identical to the AUC of the random forest.

Figure 28. ROC curve (cat xgboost)



Lastly, the feature importance plot is shown below. Note that this plot is slightly different than the plot shown in the random forest section due to the different Python libraries, but the idea is the same. It appears that the top five features remain the same as the top five features in the random forest model. Overall, we still conclude that the age of the cat, the condition the cat is in, the intake type, and if it was pre-altered are top predictors.

Figure 29. Feature importance plot (cat xgboost)



4.3.5.2 Predicting dog outcomes

XGBoost was also used to predict dog outcomes. The same hyper-parameters that were used in the cat model were also used for the dog model.

The dog data was split into three different folds. The optimal parameters for the learning rate, max depth, and number of estimators were 0.01, 10, and 50 respectively. The average macro F1 score during cross validation was 0.76, which is the same as the random forest.

Once the optimal parameters were found, the model was tested on the validation set that was held out in the beginning of the process. The figures below show the classification report and confusion matrix. **(Note that class 0 = Euthanasia and class 1 = Adoption/Return.)**

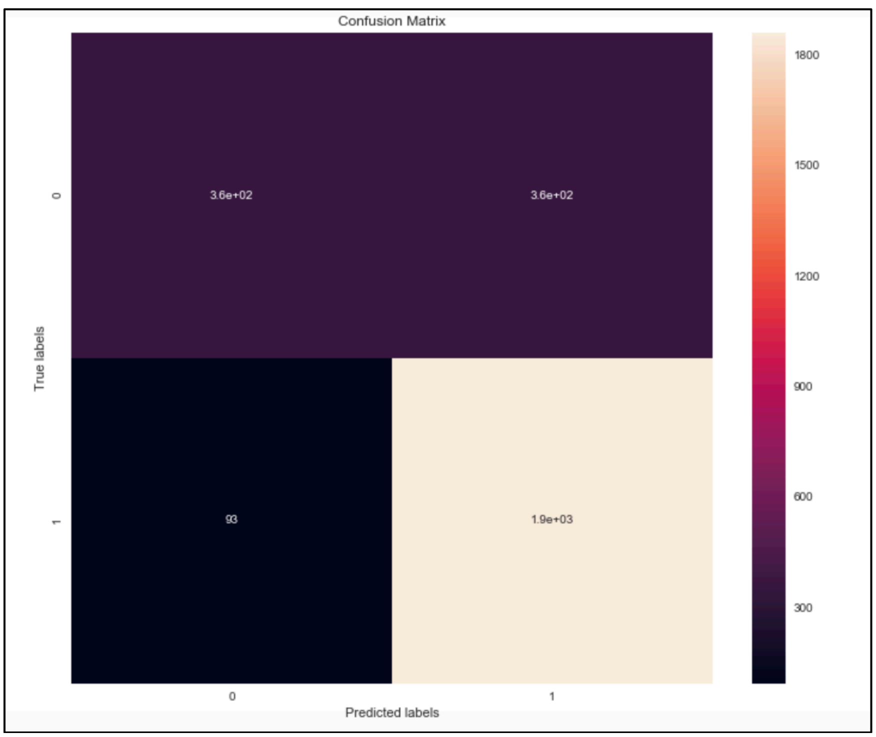
Figure 30. Classification report (dog xgboost)

```
=== Classification Report ===
              precision    recall  f1-score   support

           0       0.80      0.50      0.61       726
           1       0.84      0.95      0.89      1953

    accuracy                           0.83      2679
   macro avg       0.82      0.72      0.75      2679
weighted avg       0.82      0.83      0.81      2679
```

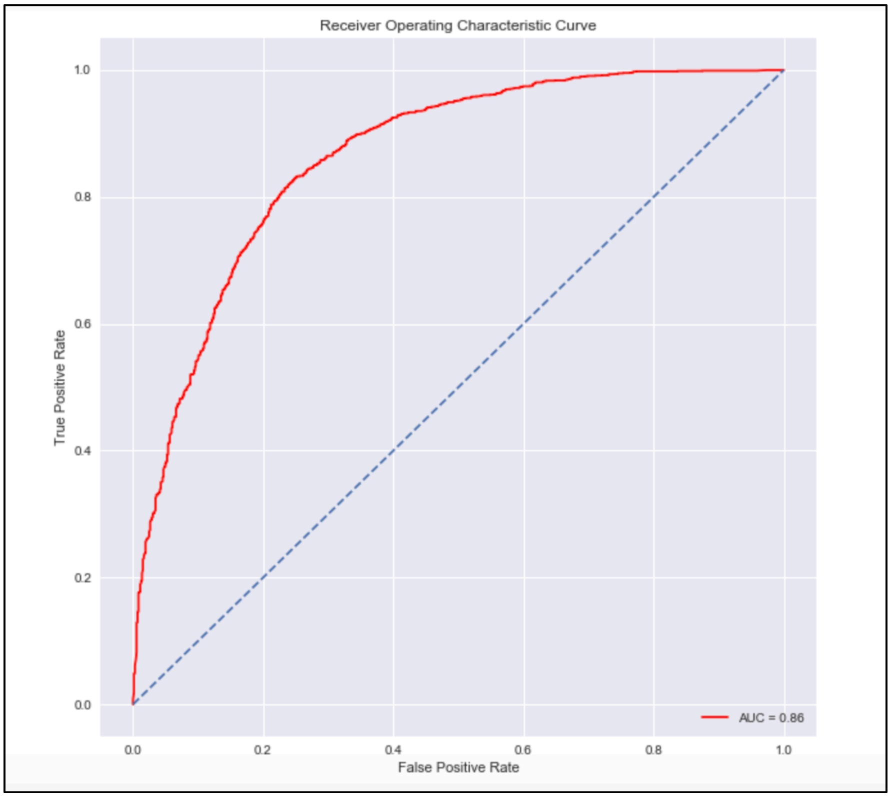
Figure 31. Confusion matrix (dog xgboost)



The results are almost identical to the random forest. Again, the dog model is better at predicting adoptions/returns than euthanasia. The F1 score for the euthanasia class (class 0) is 0.61, and the F1 score for the adoptions/returns class (class 1) is 0.89. Consequently, the macro F1 score is 0.75 and the accuracy is 0.83.

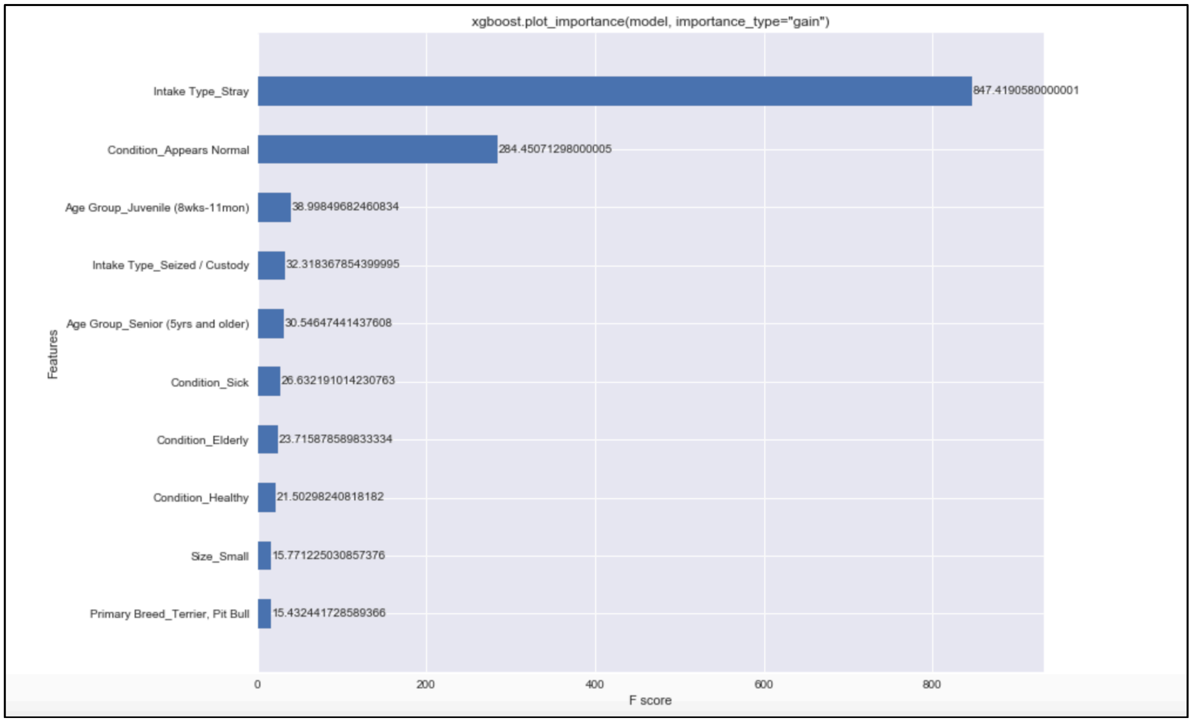
Below is the ROC curve, which shows an AUC of 0.86. This is the same as the random forest.

Figure 32. ROC curve (dog xgboost)



Lastly, the feature importance plot is shown below. It appears that the top two features remain the same as the top two features in the random forest model. However, the remaining important features are slightly different. The most interesting difference is the 10th feature, which is whether the primary breed is a pit bull or not. The random forest top 10 features did not include anything about breed, whereas the XGBoost model does. From the XGBoost model we can conclude that the intake type, condition, and age group are still key predictors.

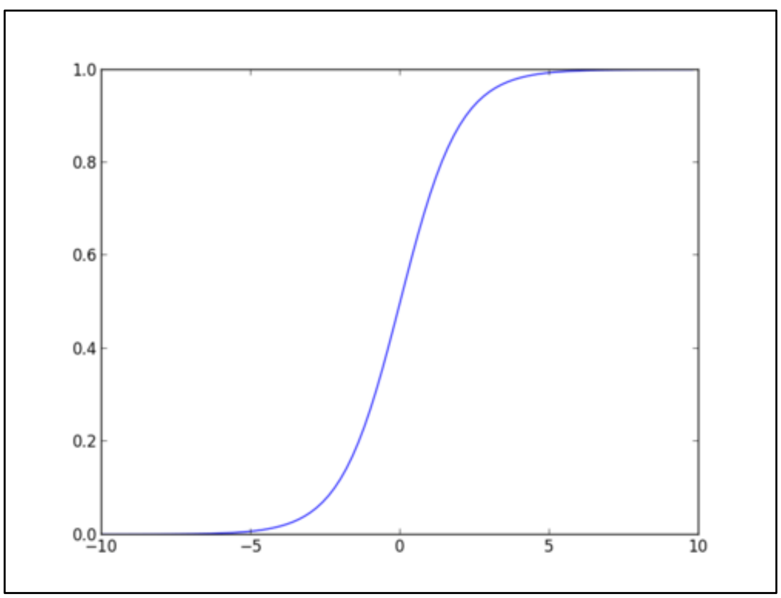
Figure 33. Feature importance plot (dog xgboost)



4.3.6 Logistic Regression (Lasso)

The last algorithm used is Logistic Regression using Lasso Regularization. Logistic Regression is a statistical technique used to model a binary outcome. It uses the logit function to predict the probability of a certain outcome. This function is in the shape of an S and falls between 0 and 1, and an example of this function is shown in Figure 34 below. If the output of this function is closer to 1, the prediction is in the positive class, if it is closer to 0, the prediction is the negative class.

Figure 34. Example of logit function curve



(Navlani, 2019)

Regularization is a method designed to reduce error and avoid overfitting. Lasso regularization adds a penalty for the absolute value of the magnitude of the non-zero coefficients in the model. This means that some coefficients become zero and are eliminated. Thus, lasso works as its own feature selector that picks out the most powerful predictive features.

Some advantages of this algorithm are that it is easy to implement, it is effective, it provides a way to select important features, and no parameter tuning is needed. The biggest disadvantage is that if features are correlated, it might choose the ones to include randomly.

4.3.6.1 Predicting cat outcomes

Since there were no parameters that needed to be tuned, grid search was not necessary. However, K-fold cross validation was still used. The training data was split into three folds. The average F1 score from cross validation was 0.72, slightly worse than the random forest and XGBoost algorithms.

The classification report and confusion matrix for the held-out validation set are shown below. **(Note that class 0 = Euthanasia and class 1 = Adoption/Return.)**

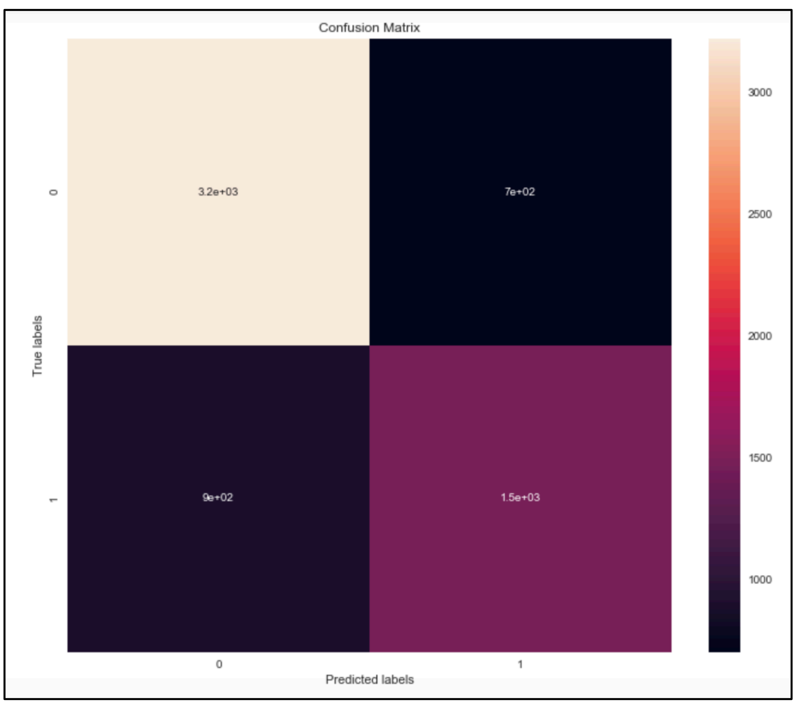
Figure 35. Classification report (cat logistic regression)

```
=== Classification Report ===
              precision    recall  f1-score   support

           0       0.78      0.82      0.80      3923
           1       0.68      0.62      0.65      2380

    accuracy                           0.75      6303
   macro avg       0.73      0.72      0.72      6303
weighted avg       0.74      0.75      0.74      6303
```

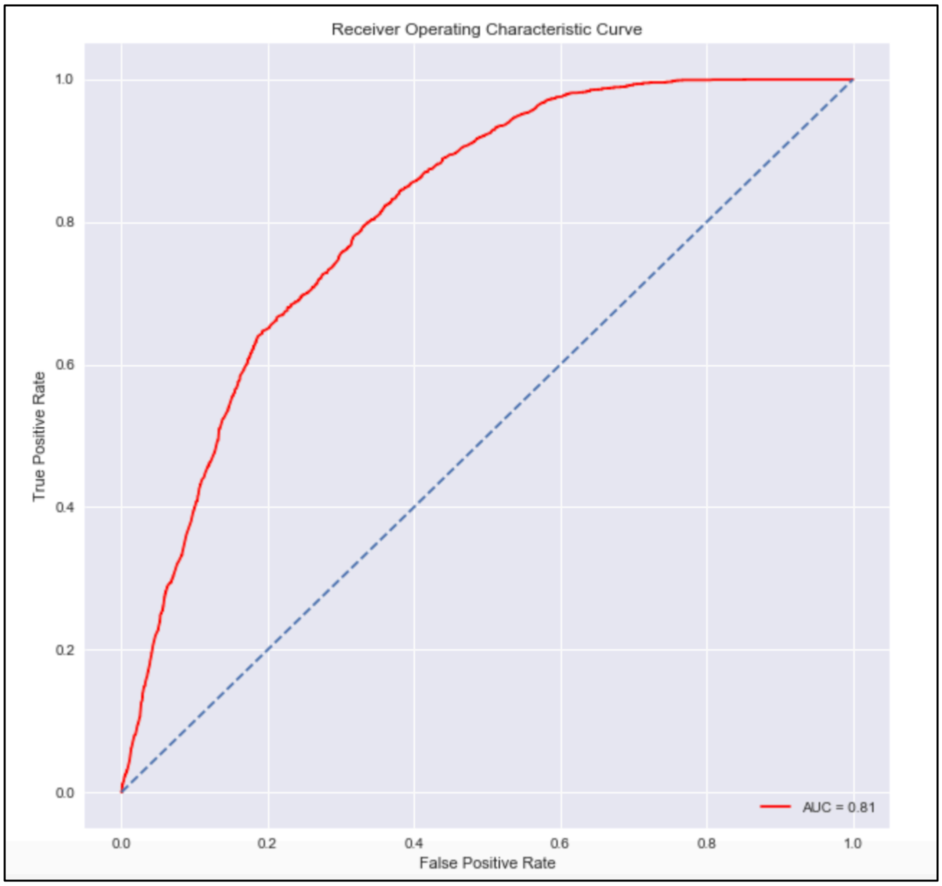
Figure 36. Confusion matrix (cat logistic regression)



The classification report/confusion matrix looks similar to those from the random forest and XGBoost algorithms. The F1 score for the euthanasia class (class 0) is 0.80, and the F1 score for the adoptions/returns class (class 1) is 0.65. Consequently, the macro F1 score is 0.72 and the accuracy is 0.75. These results are very slightly worse than the random forest and XGBoost, but the metrics are very close.

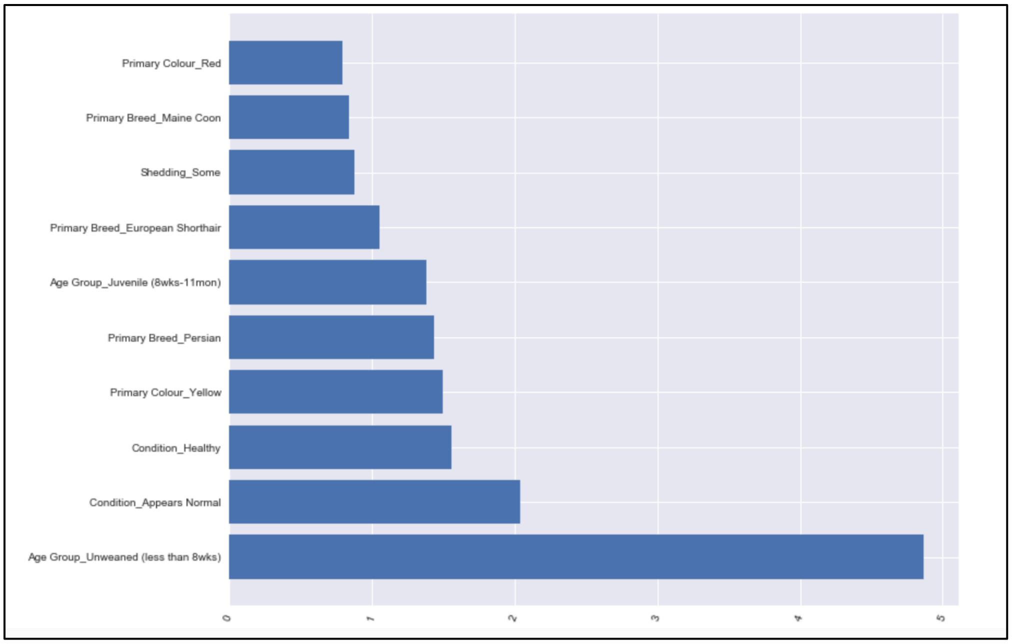
Below is the ROC curve, showing an AUC of 0.81. This is slightly worse than the AUC of the random forest and XGBoost.

Figure 37. ROC curve (cat logistic regression)



The plot below shows each feature's "importance" for the logistic regression. This is a different plot than what was shown in the random forest and XGBoost models. The values directly correspond to the absolute value of the coefficients in the logistic regression model. Because of this, the features cannot be interpreted the same way. A coefficient with a larger magnitude creates a bigger "change" in the predicted outcome. While the random forest/XGBoost features encapsulate how to best partition the data to get to a prediction, the important features in logistic regression show the features that have the biggest change in the predicted outcome.

Figure 38. Feature importance plot (cat logistic regression)



This feature importance plot looks almost entirely different than what was seen in the random forest and XGBoost. These features include features about different colors and breeds. Looking at the data, all but one "yellow" cat was euthanized out of 24 cats. This could suggest that the "top features" in logistic regression are found through very extreme cases.

4.3.6.2 Predicting dog outcomes

Next, the logistic regression was used to predict dog outcomes. The results from the validation set are shown below.

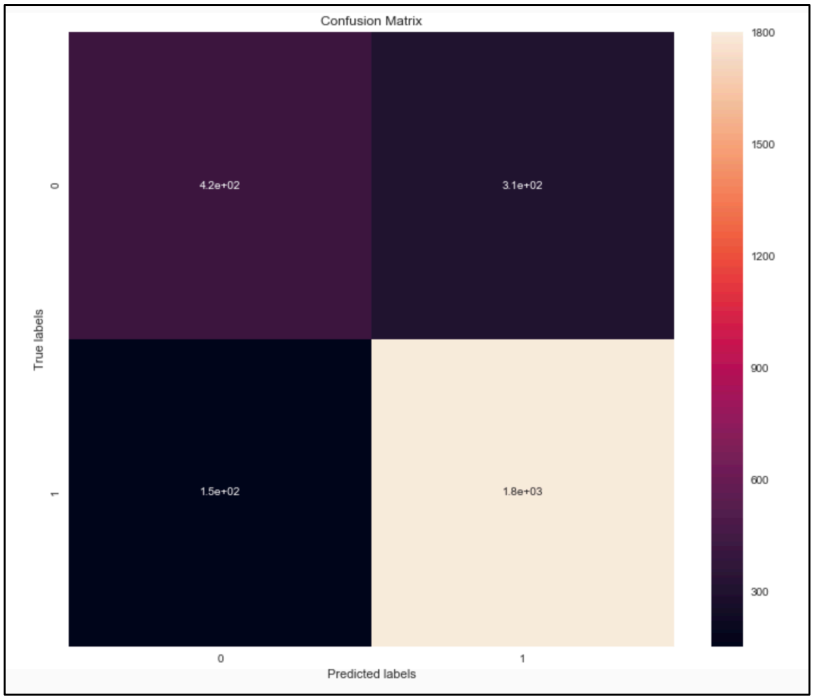
Figure 39. Classification report (dog logistic regression)

```
=== Classification Report ===
              precision    recall  f1-score   support

           0       0.73      0.58      0.65       726
           1       0.85      0.92      0.89      1953

    accuracy                           0.83      2679
   macro avg       0.79      0.75      0.77      2679
weighted avg       0.82      0.83      0.82      2679
```

Figure 40. Confusion matrix (dog logistic regression)



The results are similar to the results from the random forest and XGBoost. The F1 score for the euthanasia class (class 0) is 0.65, and the F1 score for the adoptions/returns class (class 1) is 0.89. Consequently, the macro F1 score is 0.77 and the accuracy is 0.83.

Below is the ROC curve, which shows an AUC of 0.86. This is the same as the random forest and XGBoost.

Figure 41. ROC curve (dog logistic regression)

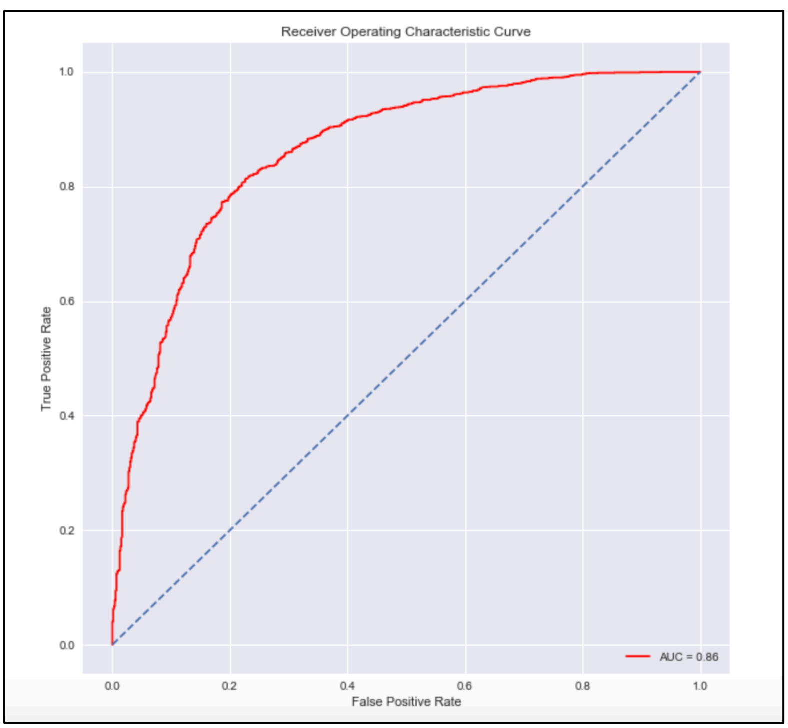
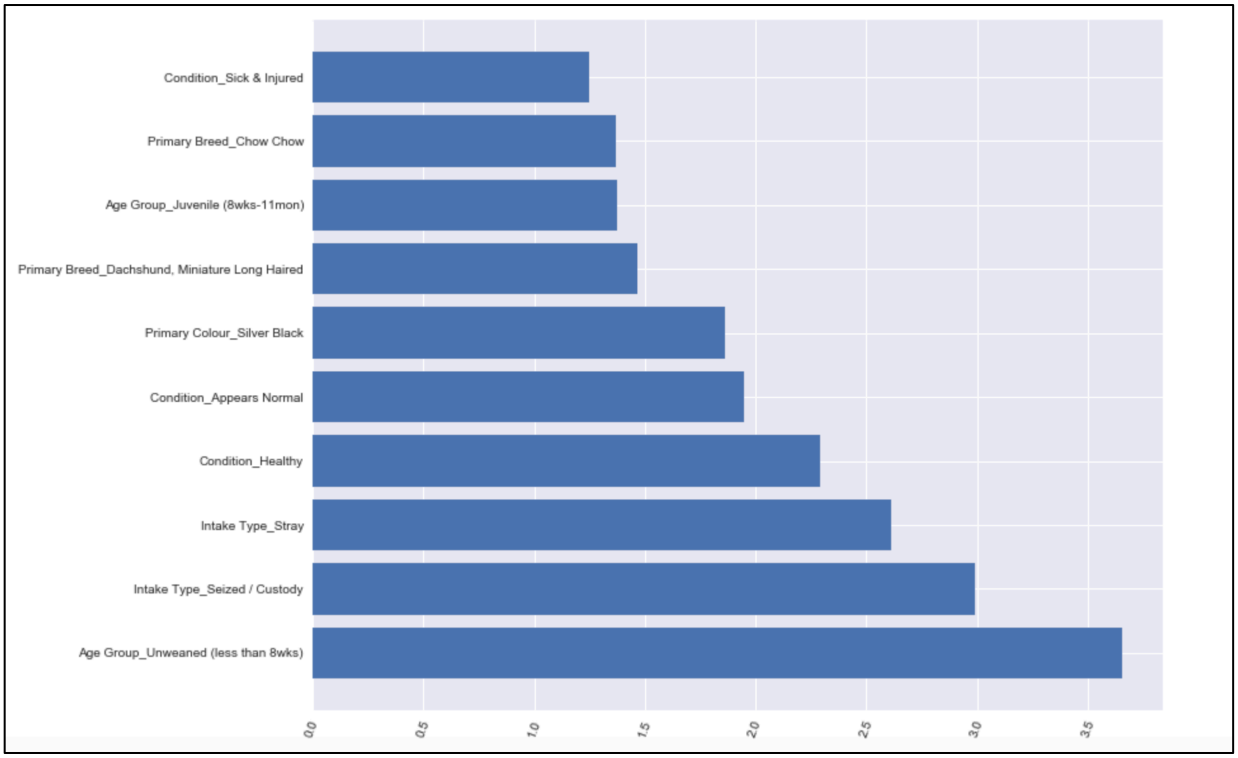


Figure 42. Feature importance plot (dog logistic regression)



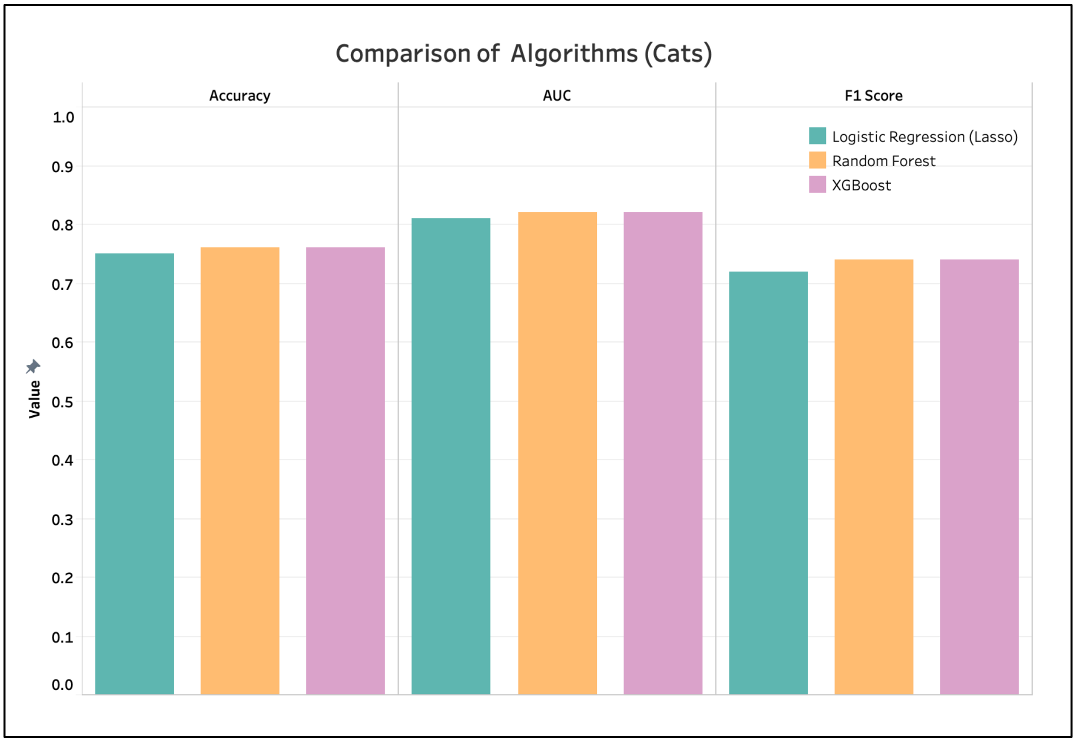
Like in the cat model, the top features from the logistic regression appear to be different than the top features from random forest and XGBoost. In the dataset, there were less than 10 dogs less than eight-weeks old, and all of them were euthanized. Again, this suggests that logistic regression with lasso is very sensitive to extreme cases.

4.3.7 Assessment

4.3.7.1 Model comparison

The plot below shows a comparison between accuracy, F1 score, and AUC for the different cat models.

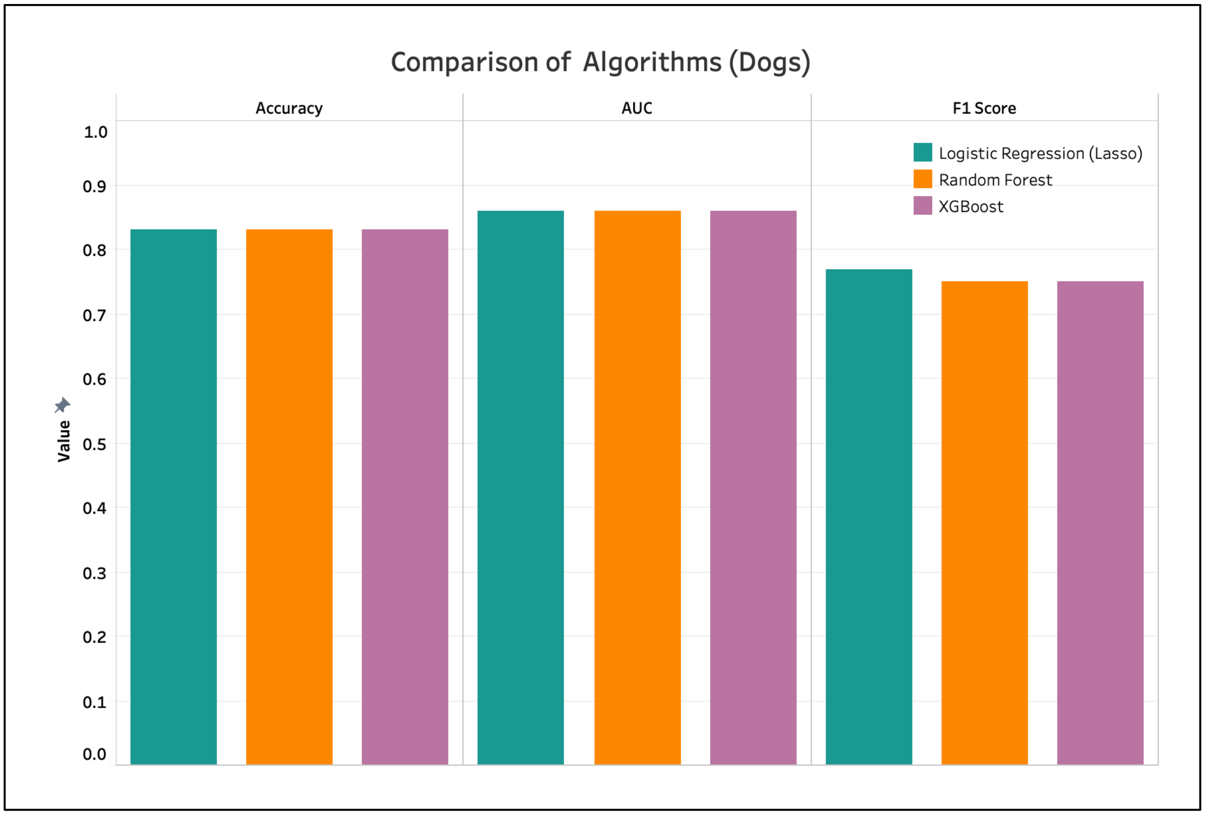
Figure 43. Cat model metrics



When predicting the different shelter outcomes for cats, the random forest and XGBoost algorithms performed exactly the same. XGBoost did not enhance performance as was initially expected. The logistic regression model using lasso regularization performed similarly to the random forest and XGBoost, but it was slightly worse. Out of the three metrics, AUC was always the highest. Recall that AUC tells us how much the model is capable of distinguishing between classes. An AUC of 0.82 means that the model is good at distinguishing between euthanasia and adoption/return. However, the F1 score tells us something different. It is used when we need to seek a balance between precision and recall. Precision means that out of those that were predicted a certain class, how many of them were actually that class. Recall means that out of the actual number of a certain class, how many correct observations did it classify into that class. More about precision and recall will be discussed below. The F1 score here is around ~0.74, which is fairly good, but not perfect. Accuracy is shown but should not be used as a way to assess the models since the classes were unbalanced.

The same plot for the dog models is below.

Figure 44. Dog model metrics



For the dog models, Logistic regression very slightly outperformed the random forest and XGBoost in the F1 score. Overall, the models performed mostly the same. The AUC of ~0.86 tells us that the model is good at distinguishing between the two classes. The F1 score of ~0.75 is also fairly good.

4.3.7.2 Precision and recall

There is always a tradeoff between precision and recall, as possibilities to increase either one could be at the expense of reducing the other. In the plots below, the precision, recall, and F1 scores are shown for each model for both dogs and cats. For the cat models, adoption/returns had fairly low recall of ~0.64. This means that the model has a lot of false negatives when predicting adoptions/returns; in other words, it classifies cats that are more likely to be adopted/returned as more likely to be euthanized around 36% of the time. For the dog models, recall was very low for the euthanasia class (~0.56). This means that false negatives are high when predicting euthanasia; in other words, it classifies dogs that are more likely to be euthanized as more likely to be adopted/returned around 44% of the time. Missing a dog that is predicted with a high chance of euthanasia is more detrimental than missing a dog that has a high chance of adoption. If the model is used to identify animals that are in danger of not being adopted/returned, correct euthanasia class predictions are critical. Therefore, the dog model needs some improvements before it can be used, if desired.

Figure 45. Cat model precision and recall metrics

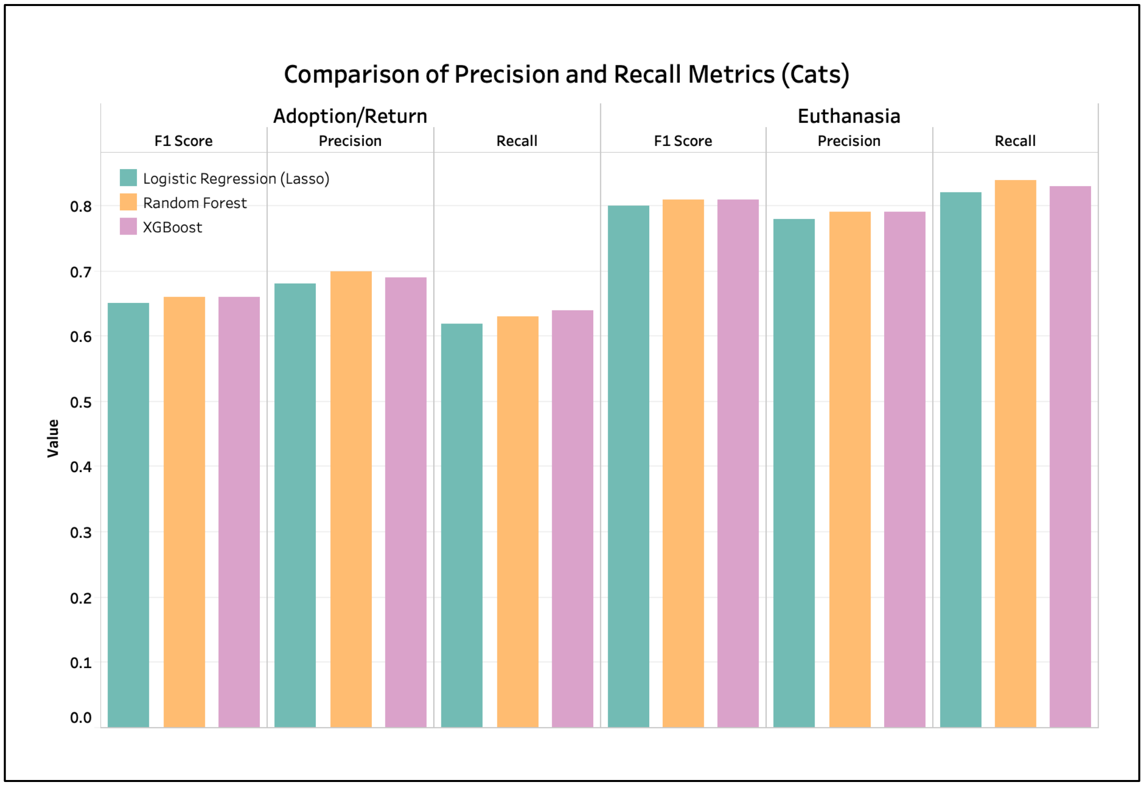
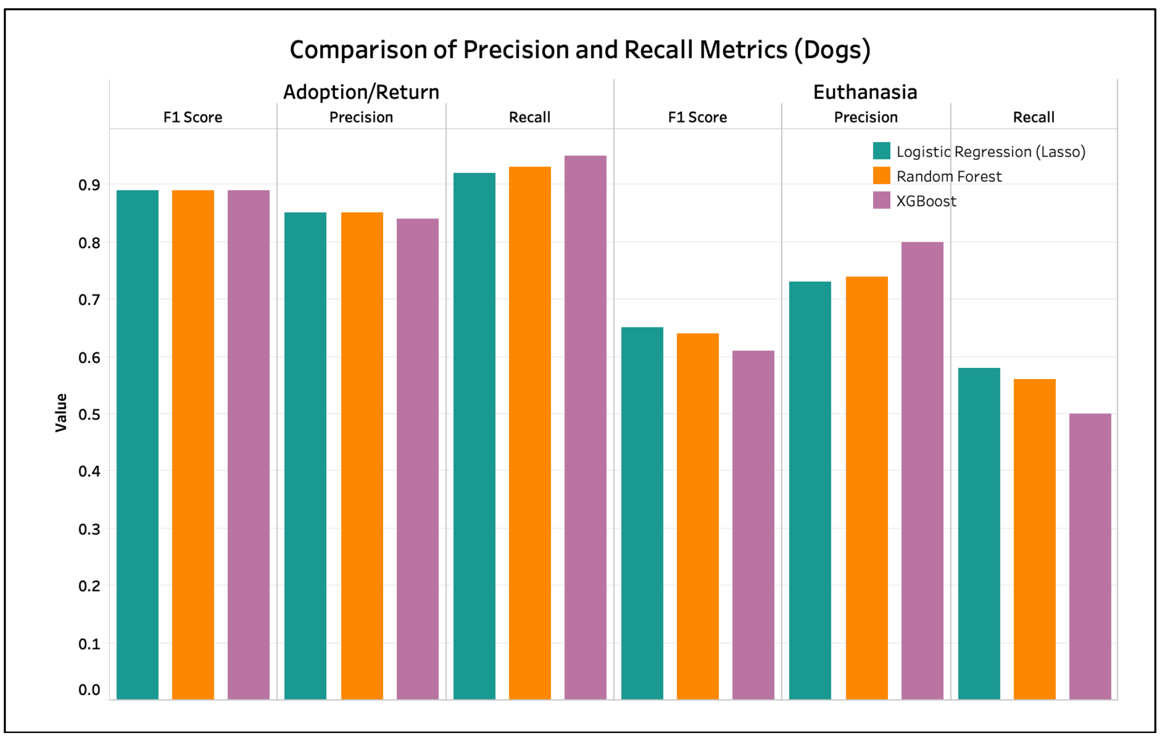


Figure 46. Dog model precision and recall metrics



4.3.7.3 Comparison of top features

Using machine learning to predict an animal's outcome can help us identify the features that have the most impact on that outcome. As stated previously, the random forest/XGBoost models provide a better view of the distinguishing features for the entire dataset. Based on the cat random forest, the top features were the age of the cat, the condition the cat is in, the intake type, and whether or not it was pre-altered. Based on the dog random forest model, the top features were the intake type, the condition the dog is in, the age of the dog, and the gender of the dog. For both cats and dogs, it is clear that both age and condition is a key predictor for their outcome. Additionally, the intake type, particularly strays, were distinguishing features for both cats and dogs. For dogs, 87% of strays were either returned to their owner or adopted. However, for cats, about 30% of strays were ether returned to their owner or adopted. Thus, a stray dog is more likely to be adopted/returned than a stray cat.

The logistic regression features were very sensitive to extreme cases (of small sample sizes), such as yellow cats and very young dogs. Thus, the random forest/XGBoost models are better at highlighting the most predictive features for the dataset as a whole.

5. Discussion

5.1 Areas of improvement

The biggest limitation of this project was the fact that very small datasets were used in the machine learning. In order to make true generalizations, hundreds of thousands of samples are needed to train a model. In this case, less than 20,000 samples were used for both the dog and cat models. Furthermore, the model would likely perform better if there were many observations in both classes (there was an issue with imbalanced data for both dogs and cats.) Therefore, the results of these models should be maintained with some degree of skepticism. This could be alleviated by combining the dataset with similar data from other shelters across the country. There were not only limitations with the amount of data, but the characteristics within the data were limited as well. For example, there was no detailed data referring to the animal's behavior, such as shy, aggressive towards people, aggressive towards other animals, etc. Including data like this for every animal would likely offer additional insights into why an animal was or was not adopted/returned to their owner.

Another limitation with this study involves the fact that a binary outcome was used. The model is assuming there are only two outcomes for an animal, which are adoption/return or euthanasia. In reality, there could be other outcomes for an animal, such as being transferred out of the shelter. In order to predict this, a large increase in data would be needed.

As discussed in the previous section, the dog model needs improvements. This is because the dog model often misclassifies dogs at risk of euthanasia as the opposite. If this model were ever to be used to help identify animals at risk of not being adopted or returned to their owner, this model would need to be improved.

5.2 Future work

Understanding animal shelter outcomes does not stop here. As stated previously, the predictive models could be improved by incorporating more detailed data about the animals, such as behavior. However, there are other interesting problems that can be looked at as well. For example, a very interesting study would be to analyze how advertising/marketing, community outreach, and/or social events impact animal adoption rates. As mentioned in Jill Dyché's study, one could analyze how putting an animal on social media (images, videos, etc.)

increase the chance of adoption or reduce the amount a time an animal would wait to be adopted. Another interesting project would be to use recommender systems to match humans to pets that are best suited for their lifestyle. Lastly, one could try to predict the length of time a particular animal would stay at the shelter, which could highlight the variables that have the most impact on their time waiting to be adopted etc.

5.3 Comparison to related work

As mentioned previously, Joanne Lin's study was the motivation for this project. Ms. Lin also decided to predict a binary outcome, but in her case, she chose to predict adoptions vs "other", where other could be euthanasia, transfer, died, etc. In Ms. Lin's analysis, she was able to reach an accuracy of around 82%. Although her study does not include other metrics like F1 score and AUC, the accuracy scores were very similar. However, Ms. Lin had a much larger amount of data to train her models. Additionally, Ms. Lin found that for dogs, the most predictive features were the breed of the dog, whether it is spayed/neutered, and the dog's age. For cats, the top features were whether the cats were spayed/neutered, whether that cat had a name, and the cat's age. In both projects, the animal's age and the fact that the animal was spayed/neutered (pre-altered) are key factors when it comes to an animal's outcome.

5.4 Reflections

Data science projects can be very challenging but rewarding. One of the key take-aways from working on this project is that the model can only be as good as the data that is used to train it. A common saying is "garbage in, garbage out". Making sure that there is adequate data to actually build a decent model is one of the most important things to remember when working on a data science project. An increase in the volume and quality of data in this project could significantly improve the results. Additionally, another take-away is that exploratory data analysis is extremely important. Being curious about the data and the underlying relationships will only improve modeling. For example, digging into the relationship between animal age group and the shelter outcomes turned out to be a key finding. One of the most rewarding aspects of this project is potentially providing actionable insights for an important cause, which in this case is helping animals find homes and helping Frederick County Animal Control improve their daily operations. The most challenging part of this project was making sure that the models weren't biased (or trying to decrease bias as much as possible). Including the spay/neuter and "has name" features would have significantly boosted the model's performance. However, these features were essentially cheating the model, so they had to be removed. Another challenging aspect is communicating the results effectively to audiences with different backgrounds. It is important to make sure the methods, key findings, and limitations are clearly presented. A suggestion for someone working on a future data science project would be to try and find a problem they are passionate about. This will make working on the project seem less like "work" and make it more enjoyable. Additionally, it is critical to understand the underlying workings of different machine learning algorithms. Without this knowledge, being able to choose the right models and interpret the results correctly would be nearly impossible.

6. Conclusion

The goals of this project were to provide an analysis highlighting different trends in FCAC's data as well as to build machine learning models to predict animal shelter outcomes.

During exploratory data analysis, the 21702, 21701, and 21703 zip-codes were identified as top regions for abandoned animals. Thurmont (21788) was also identified as a top region for stray and seized cats. FCAC can use this information to perhaps extend their adoption hours on those days or add additional staff/volunteers to assist with the adoption process. Additionally, an analysis of adoption times showed that Wednesday evenings were the most popular day and time for adoptions, followed by Saturday afternoons. FCAC can use this information to determine when to schedule the most adoption times and when to hold social events to promote adoptions. While analyzing the length of time dogs and cats spend at the shelter, an important finding was that cats wait a much longer amount of time to get adopted than dogs. Cats waited a median of 42 days, while dogs waited a median of 15 days. In particular, juvenile dogs are the quickest to get adopted, and adult cats wait the longest to get adopted. Dynamic visualizations showing additional plots and charts were compiled into an interactive dashboard via Tableau.

Machine learning algorithms were used to predict if a cat or a dog has a higher chance of being adopted/returned or euthanized. The random forest, XGBoost, and logistic regression with lasso algorithms were used to predict the outcome. Separate models were created for both dogs and cats in order to compare the features that hold the most predictive power. For both cats and dogs, the three algorithms performed nearly the same. For cats, the accuracy, F1 score, and AUC were roughly 0.76, 0.74, and 0.82 respectively. For dogs, the accuracy, F1 score, and AUC were roughly 0.83, 0.76, and 0.86. These numbers are close to 1 (meaning a perfect model) but show that there is room for improvement. Based on the random forest models, the top predictive features for cats were the age of the cat, the condition the cat is in, the intake type, and whether or not it was pre-altered. For dogs, the top features were the intake type, the condition the dog is in, the age of the dog, and the gender of the dog. It is evident that both the age and the condition of both dogs and cats are key factors when it comes to their shelter outcome.

The biggest limitation of this project was the limited amount of data and the fact that the classes were imbalanced. An increase in data (from other shelters) could provide more generalizable results and a model that can actually be used for decision making. In addition, the dog predictive model had fairly low recall scores when predicting euthanasia, which is more detrimental than low recall scores for adoption/euthanasia. The goal of the model is not to decide if an animal should be euthanized, but to determine if an animal has a higher predicted risk of being euthanized. Because of this, recall for the euthanasia class needs to be increased and improvements need to be made to the model before it could be used as a decision-making tool.

References

Ganegedara, T. (2019, June 12). *Intuitive guide to understanding decision trees*. Retrieved from https://towardsdatascience.com/light-on-math-machine-learning-intuitive-guide-to-understanding-decision-trees-adb2165ccab7

Levenson, A. (2019, July 1). *Can data science help us save animals?* Retrieved from https://www.thinkful.com/blog/how-data-science-can-help-us-save-animals/

MacPete, R. (2015). *What the heck is kitten season anyway?* Retrieved from http://www.pethealthnetwork.com/cat-health/cat-checkups-preventive-care/what-heck-kitten-season-anyway

Navlani, A. (2019). *Understanding logistic regression in python*. Retrieved from https://www.datacamp.com/community/tutorials/understanding-logistic-regression-python

Underwood, J. (2016, July 26). *Making a difference with data - animal shelters*. Retrieved from https://www.jenunderwood.com/2016/07/26/making-a-difference-with-data-animal-shelters/

GitHub Repository: https://github.com/klongo23/FCAC-Data-Analysis
Tableau Dashboard: https://public.tableau.com/profile/kaitlyn.longo#!/vizhome/FCACDataAnalysis/FinalDashboard