DS-UA 202, Responsible Data Science, Spring 2024 Course Project: Technical Audit of an Automated Decision System

rk4289, sl9854

April 2025

1 Background

Every year, over 7.6 million household pets are given up by their owners or picked up from unfortunate circumstances to end up in US animal shelters. The purpose of this Automated Decision System is to analyze the shelter outcomes and identify any trends that may be present in animal adoption and euthanasia events. These insights could be used in shelters to refocus efforts on helping specific animals that need extra assistance finding a new home.

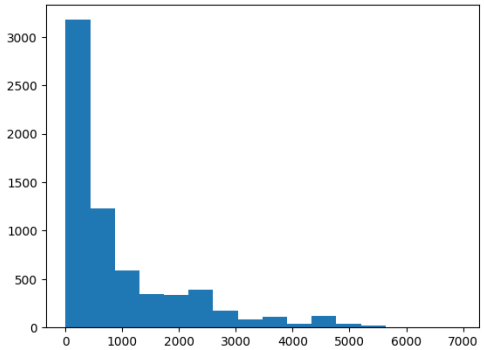
This ADS has one main goal of predicting outcomes to inform rehoming efforts.

2 Input and Output

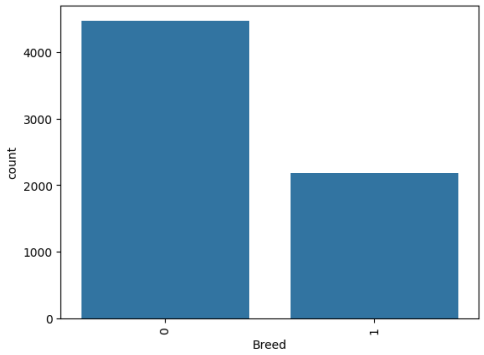
The dataset we are using is collected by the Austin Animal Center. Every row in the data set is one encounter with an animal, and individual animals are assigned an ID on entrance to the shelter. Their outcome is recorded upon their departure. The dataset is updated daily, and the ADS we are using right now is using a version of the dataset from 2016. The models and pre-processing are done separately for cats and dogs.

Below are numerical visualizations of important or interesting features, both provided and engineered.

- **Dog Features**
  - Age Upon Outcome

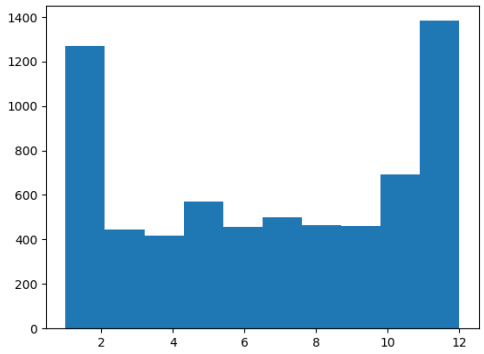


    In the chart, we are analyzing a shelter dog's age (in days) when they arrive at the shelter. The distribution of the data closely resembles an exponential distribution with the most arrivals being 1 year or younger in age. The average dog lifespan is around 11 years (4000 days) and is likely shorter among dogs in shelters (especially dogs with multiple shelter visits). Litters are likely surrendered to shelters as their mother's owners cannot care for them.
  - Breed



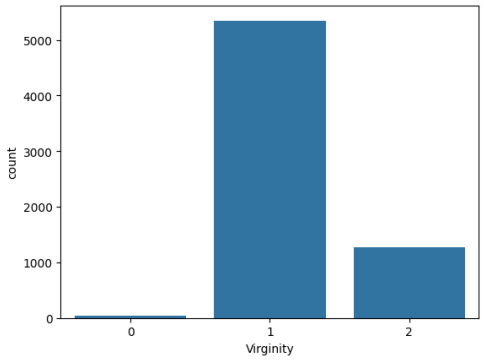
This histogram shows the number of dogs at shelters which are Mixed Breed (0) or Pure Breed (1). We can see shelters take in more mixed breeds than purebreds, likely due to the fact that purebred dogs are almost always bred intentionally and not surrendered. It is worth noting that the creator of this ADS used imperfect machine logic to sort dogs into these categories based on the breed names, and that shelters visually classify breeds which can often be inaccurate.

– Month



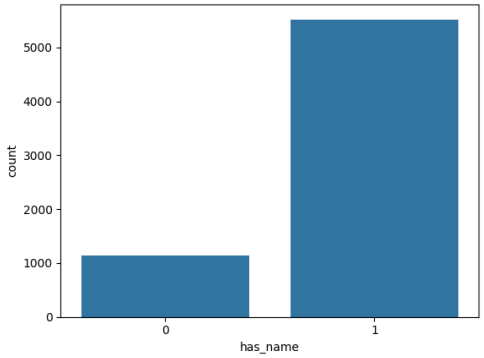
This graph measures the amount of dog arrivals based on months of the year. The distribution shows that there is a significant increase in arrivals in the coldest months of the year, which makes sense for a shelter.

– Fixed Status



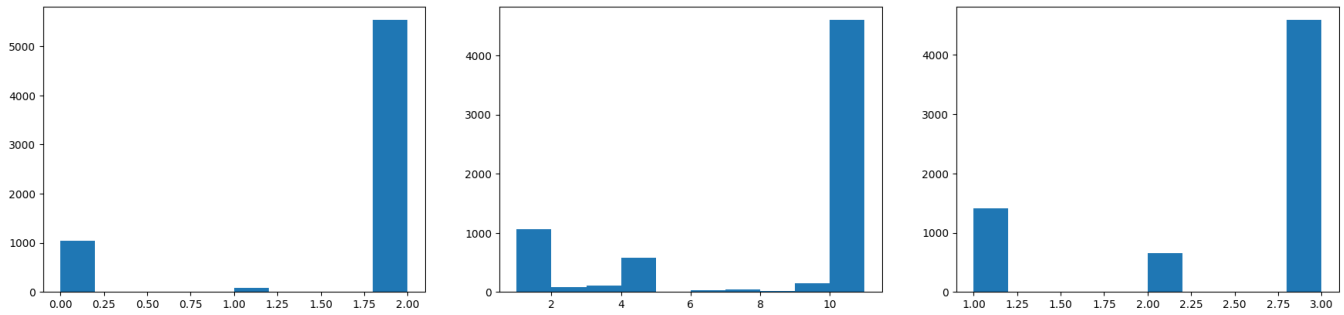
This graph shows dog arrivals split by whether or not they are neutered/spayed. In the X axis, 1 represents neutered and 2 represents not neutered. 0 represents unknown status. As seen in the distribution, there are significantly more neutered dogs that arrive than un-neutered. Neutered dogs are not born as strays, and getting a dog neutered is not cheap, meaning many of these dogs were probably surrendered by someone who intended to keep them.

– Has Name



We can see that there are significantly more dogs with names than without. Dogs that come in without names are likely strays or very young, so this feature can contain useful information that influences their outcome.

– Hairgroup, Aggressiveness, Weight

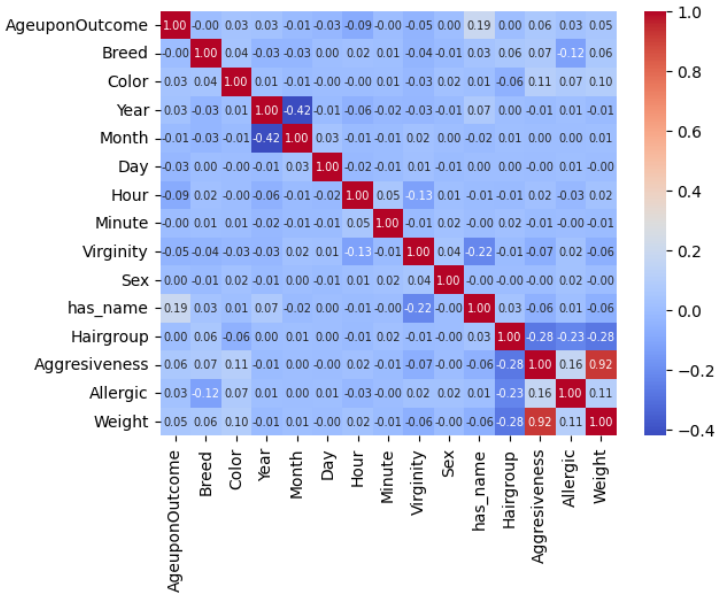


(a) Hair Group (b) Aggressiveness (c) Weight

Figure 1: Feature Visualizations: Hair Group, Aggressiveness, and Weight

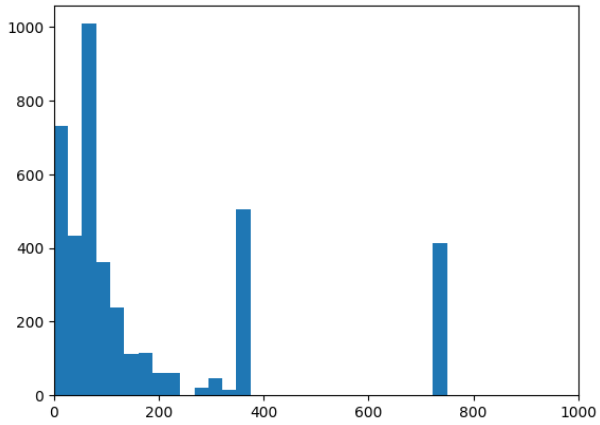
These features were engineered by flawed machine logic which determined these by the reported name of the breed only. Only 10 Breeds were specified, and the rest fall into an 11th category, as we can clearly see in these meaningless distributions.

Dog Correlation Matrix



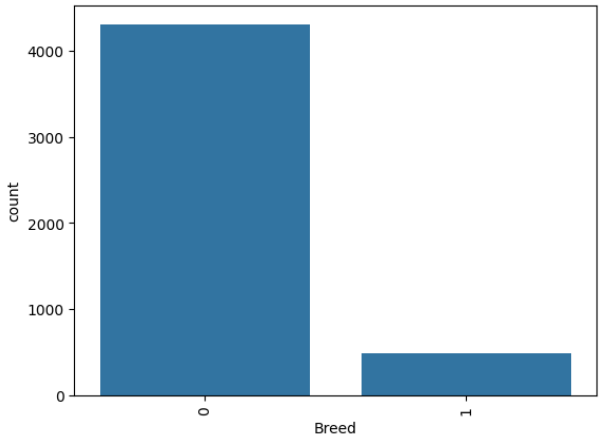
The major correlations we see are a correlation between virginity (fixed) and name, as well as correlations with the engineered features at the end. Pets that arrive with a name are more likely to have been neutered, and the features at the end are only accurate for about 10 breeds.

- Cat Features
  - Age Upon Outcome



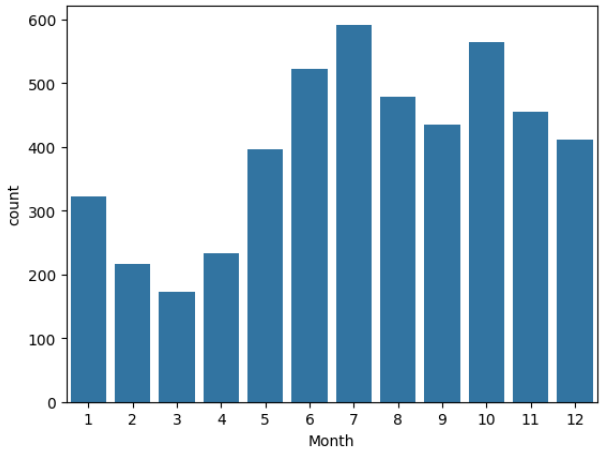
This chart shows the distribution of shelter cat ages (in days) when they arrive at the shelter. The distribution shows that there is a large spike in arrivals around day 0, or right after birth. There is another larger spike around day 80-90, which is right after weaning age, and the appropriate time for kittens to be rehomed. The gap in between days 400 and 700 is likely due to the age being recorded in years instead of days, and not included in the dataset.

  - Breed



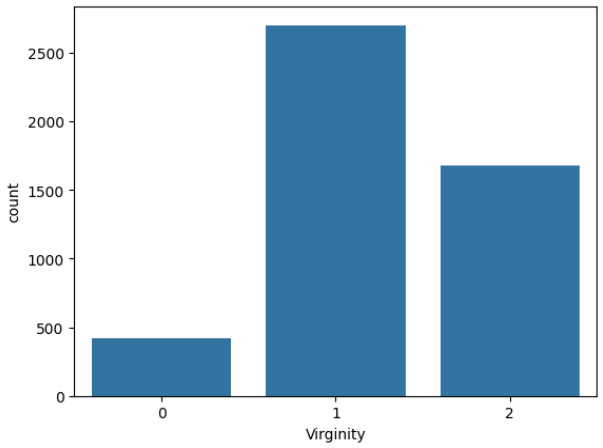
This chart shows the distribution of shelter cats based on whether they are a mixed breed(0) or purebred(1). The distribution shows a much larger quantity of mixed breed cats compared to purebred, however, the logic used by the author to identify mixed vs pure bred cats is flawed, as mentioned above, making these results less impactful. This difference could be due to the fact that purebred cat breeds are worth more in price, and are therefore less likely to end up in shelters.

  - Month



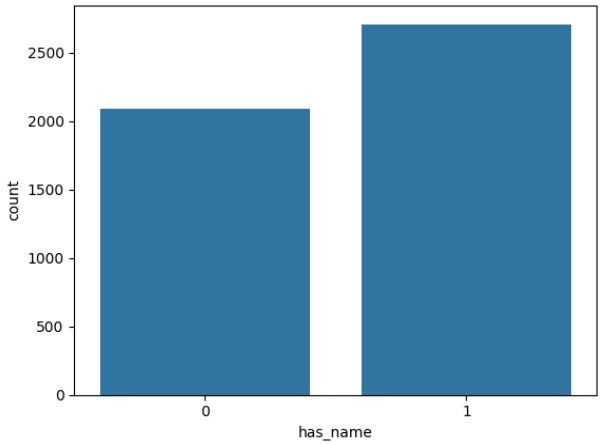
This graph measures the amount of cat arrivals in the shelter based on months of the year. The distribution shows a slight increase in cat arrivals in months 7 and 10, which seems random.

– Fixed Status



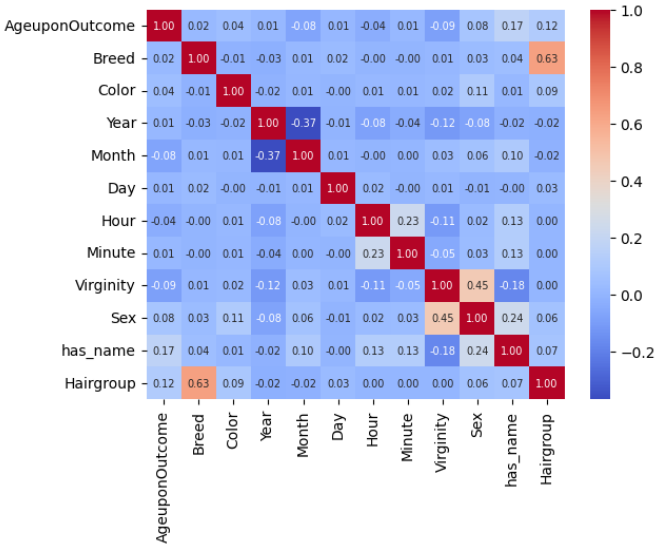
This graph shows cat arrivals by whether or not they are neutered/spayed. On the X axis, 1 represents neutered and 2 represents not neutered, with 0 being unknown status. The distribution shows a larger amount of arrivals that are fixed compared to not fixed, possibly signaling that most of the arrivals are people returning pet cats, and not strays, as those would not be neutered.

– Has Name



We can see in this graph that there are more cats in the shelter with a name (1) than without (0). This indicates that more cats come in that are previously homed compared to those without a name who are probably strays or too young.

Cat Correlation Matrix



The major correlations here are between breed and hairgroup, which makes sense as those 2 things are directly related, and between fixed status (virginity), sex. It seems female cats are for some reason more likely to be named and fixed than male cats. Other correlations are negligible or nonsensical.

3 Implementation and Validation

a). For the data cleaning step, any columns with unknown values were dropped and the set was split between cats and dogs to train individual models. From here, 3 features were engineered. Hairgroup, Aggressiveness, and Weight. These features were engineered by taking the top 10 breeds in these categories off the internet and checking if they matched the breed of the animal, leading to the vast majority of the animals remaining unclassified.

b). The non numerical data was encoded and the data was plotted for exploratory data analysis. The data was then split into training and validation sets and fed directly into the models. Gradient Boost performed best on the validation sets, so it was chosen as the final model.. The author mentioned he did not do feature reduction as it reduced accuracy.

c). The ADS was validated using the validation set mentioned above, and again when the author submitted his prediction to the dataset's competition, where it achieved a score very similar to the validation set on unseen data. The stated goal was to gain insight into resource allocation for the shelters, and with the high overall accuracy and recall demonstrated by the model, we can be sure that the model does provide insight.

4 Outcomes

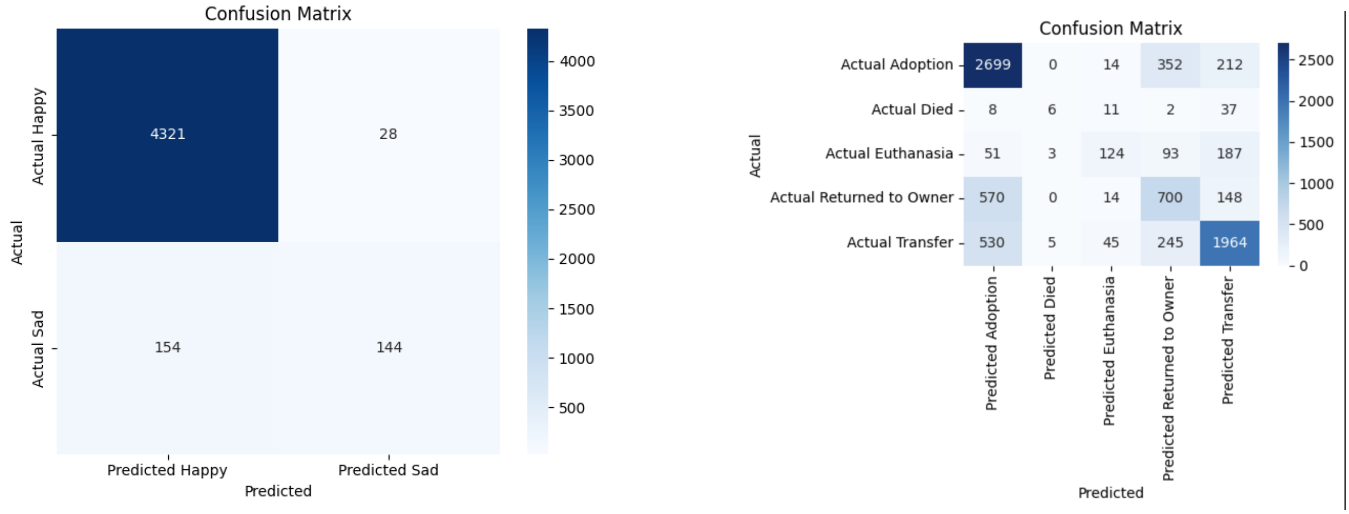
Here we will test the accuracy of the model overall as well as the accuracy for certain classes such as gender and spayed/neutered. We will check for equalized odds between these groups to ensure fair representation. (Note: We will not check equalized odds between cats and dogs as these two have separate models trained with separately processed data). Because the purpose of this ADS is to help animal shelters allocate resources, we choose to use equalized odds to make sure no group has unfair representation which could take away resources from animals that need them more.

The first step was to assess the confusion matrix for all animals as a test for overall bias of the algorithm. To do this we generate the confusion matrices for cats and dogs alike and add their values. It is worth noting that in this case a positive outcome is "Happy", meaning the animal was either returned to owner or adopted. The negative outcome is "Sad", meaning the animal either died in the shelter or was euthanized. For the purpose of this project, we will ignore those with the outcome "Transferred" as it is a neutral outcome, neither positive nor negative.

The official accuracy score of the model is 0.61 for dogs and 0.80 for cats.

(Note: While these scores are relatively good, they improve significantly if we reduce the number of outcomes from 5 to 2. When we treat "Returned to owner" and "Adopted" as the same outcome "Happy" we can see the model is not penalized for mixing between them.)

After making the above change, the newly calculated accuracy overall is 0.96. As we can see in the confusion matrices below, the positive outcomes in the training data significantly outweigh the negative outcomes. This is a notable quality of this dataset which could lead to bias.



(a) Reduced Confusion Matrix

(b) Full Confusion Matrix

Figure 2: Comparison of Reduced and Full Confusion Matrices

$$\text{TPR} \approx 0.9927$$
$$\text{TNR} \approx 0.4447$$
$$\text{FPR} \approx 0.5553$$
$$\text{FNR} \approx 0.0073$$

While our model is very accurate, we can see a significant proportional difference between our false positive and false negative errors. This is indicative of a significant positive bias in our model which is likely driven by the skewed information in the data explained earlier.

We can move on to checking for equalized odds between male and female animals.

Male Classification Metrics

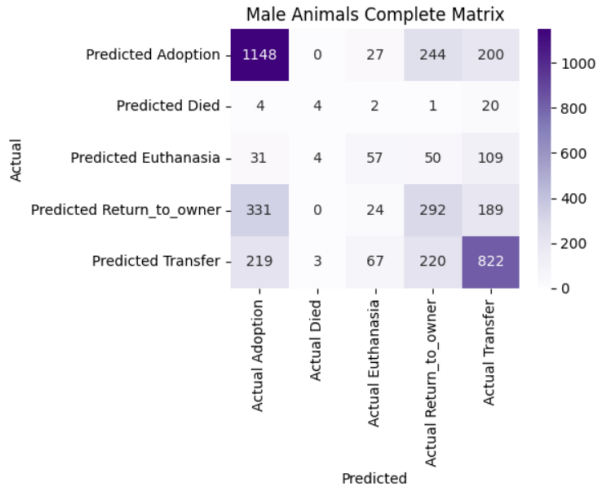
$$\text{TPR} \approx 0.9934$$
$$\text{TNR} \approx 0.4318$$
$$\text{FPR} \approx 0.5682$$
$$\text{FNR} \approx 0.0066$$
$$\text{Accuracy} \approx 0.9491$$

Female Classification Metrics

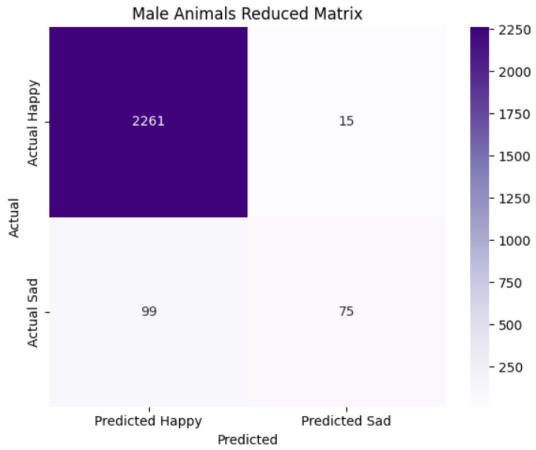
$$\text{TPR} \approx 0.9920$$
$$\text{TNR} \approx 0.3964$$
$$\text{FPR} \approx 0.6036$$
$$\text{FNR} \approx 0.0080$$
$$\text{Accuracy} \approx 0.9336$$

From these Metrics, we can clearly see that between male and female animals we have very similar false positive and false negative error rates. This is to be expected as there is no prevailing sentiment in the US about which gender animals to adopt, and these animals are represented nearly equally in the training data. This sensitive feature does

satisfy equalized odds as both groups have similar True Positive and False Positive rates.

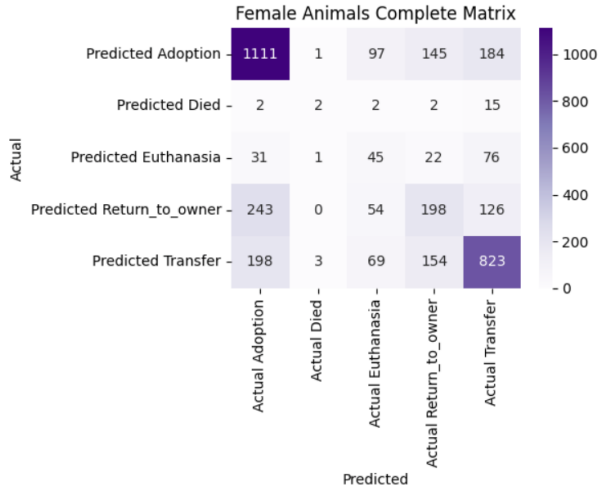


(a) Full Confusion Matrix

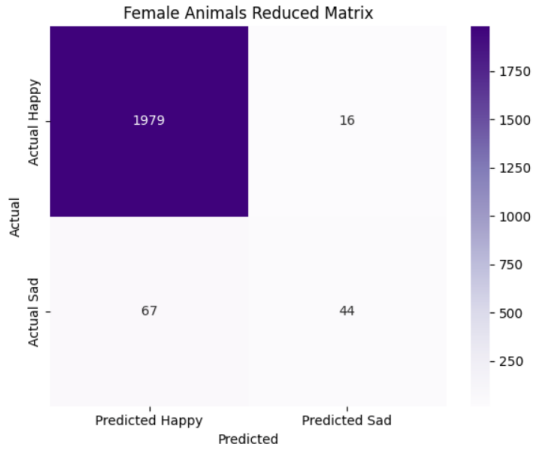


(b) Reduced Confusion Matrix

Figure 3: Male Animal Confusion Matrices



(a) Full Confusion Matrix



(b) Reduced Confusion Matrix

Figure 4: Female Animal Confusion Matrices

Now we move on to checking equalized odds between fixed animals and intact animals

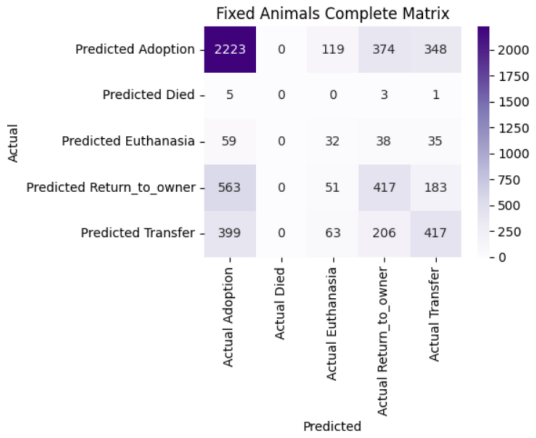
Fixed Classification Metrics

$$\text{TPR} \approx 0.9971$$
$$\text{TNR} \approx 0.1862$$
$$\text{FPR} \approx 0.8138$$
$$\text{FNR} \approx 0.0029$$
$$\text{Accuracy} \approx 0.9574$$

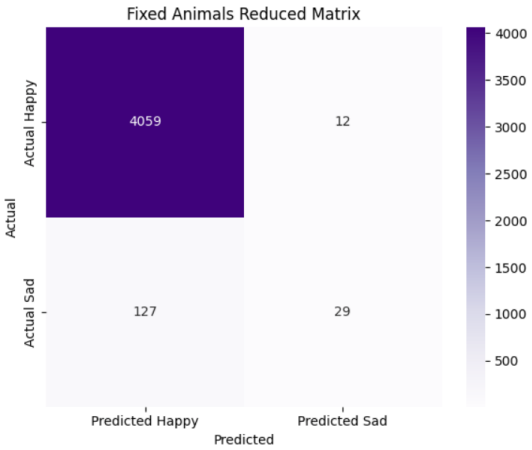
Intact Classification Metrics

$$\text{TPR} \approx 0.9050$$
$$\text{TNR} \approx 0.6977$$
$$\text{FPR} \approx 0.3023$$
$$\text{FNR} \approx 0.0950$$
$$\text{Accuracy} \approx 0.7888$$

Here we see a significant difference in the False positive rates and a noticeable difference in the true positive rates. This is because we have significant class imbalance. We also notice a significant drop in the accuracy. In the context of how this ADS is meant to be deployed, this discrepancy means that fixed animals are more likely to be falsely labeled with a positive outcome, which could lead to the group receiving less resources than they need.

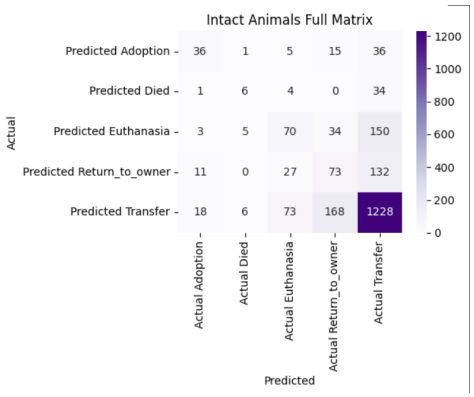


(a) Full Confusion Matrix

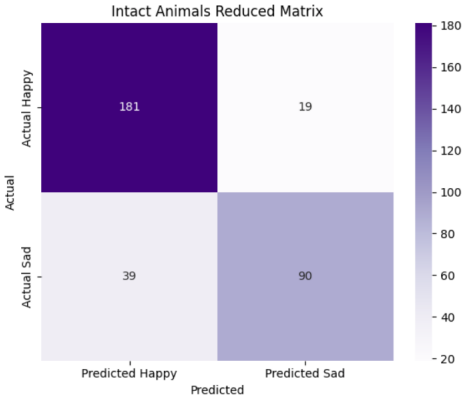


(b) Reduced Confusion Matrix

Figure 5: Fixed Animal Confusion Matrices



(a) Full Confusion Matrix



(b) Reduced Confusion Matrix

Figure 6: Intact Animal Confusion Matrices

5 Summary

a). The dataset used for this ADS was mostly appropriate for predicting shelter animal outcomes based on different variables that are expected to influence adoption or euthanasia rates. However, there are several problems affecting the quality and fairness of the results.
There is a very clear imbalance of positive over negative outcomes, which significantly skews the shelter animal predictions. Additionally, several features including hair group, weight and aggressiveness were generated using flawed machine logic, thus making them not generalizable to the specific outcome predicted.
Multiple categories with missing values, like a status of "unknown", or gaps in the age outcomes, lead to inconsistent results and may affect the predictions for the true outcome.

b). The accuracy of the implementation is affected by the class imbalance of positive vs negative outcomes. The initial model had an accuracy of 0.61 for dogs and 0.80 for cats, which is a moderate level. After reducing the outcomes to "Happy" vs "Sad", the accuracy increased to 0.96. However, we know that the negative outcomes are in this dataset are under-predicted, as seen in the low true negative rate of ≈ 0.44.
As for fairness, equalized odds appear to be satisfied between male and female animals, as both the TPR and FPR are approximately equal across groups. However, equalized odds are not satisfied between fixed and intact animals, as intact animals have a significantly lower TPR (0.905) than fixed animals (0.9971). Intact animals also have a higher FNR (0.095) than those that are fixed (0.0029). This disparity suggests that the model is less likely to correctly predict positive outcomes for intact animals, and also overestimating positive outcomes for fixed animals. This could lead to intact animals being looked over in shelter resource allocation and decision making.
Fairness across different traits would be particularly relevant to shelter staff, as it affects the allocation of resources and care for different animals, depending on their needs. The general public would expect fair treatment across all animals regardless of condition, so it is important to preserve transparency regarding how the input features influence predictions and outcomes.

c). I would not be comfortable deploying the ADS in the public sector in its current state due to its inconsistent and incomplete nature. Specifically, the disparities in FPR's and use of flawed engineering in features could have negative impacts on resource allocation and introduce discriminatory outcomes for certain breeds. Additionally, the lack of transparency in data processing and feature generation could raise questions about ethics in both the public and industry settings. However, it could still be useful in shelters where the demographics match the training demographics.

d). This system requires more thorough validation, fairness audits and significant improvements to its feature design in order to be considered for real-world use. This would include improving flawed logic used as proxy for important features, fixing variable inconsistencies and increasing transparency by using interpretable models for stakeholders to understand predictions.