April 2018

Animal Shelter Outcome Prediction

Data Science

Final Project Report

Sabrina Riebe

Introduction

The purpose of this project is to provide Austin Animal Center (and other shelters) with an understanding of trends in animal outcomes and with the ability to identify which animals are least likely to be adopted so that they can dedicate extra attention in helping these pets find a happy home.

The dataset used comes from the Austin Animal Center over the course of October 1st, 2013 to March, 2016.

Table of Contents

Table of Figures

Problem and Goal of the Project

Every day there are many animals given up as unwanted to U.S. shelters, while other animals are rescued from cruelty situations, or are found wandering the streets. This results in approximately 7.6 million pets and companions ending up in U.S. shelters each year. Many of the animals in the shelters are adopted and find loving families to take them home, however, many are not so lucky. Those that are not adopted cannot stay in the shelter forever and are euthanized. 2.7 Million cats and dogs are euthanized every year in the U.S.

The goal of this project is to create a model to understand trends in animal outcomes. These insights could benefit Austin Animal Center (and other shelters) by helping them focus their attention on those animals that are predicted to need a little extra help finding a new home.

Original Dataset

The data for this project comes from Austin Animal Center, and includes information about their shelter animals from October 1st, 2013 to March 2016.

- 10,000 animal profiles
    - Approximately 450 different breeds of animals
    - Approximately 5800 dogs
    - Approximately 4100 cats
    - Approximately 4000 female animals
    - Approximately 4500 males
- The dataset is already split into a train and test dataset
- All data reflects the status of the animal as they **leave** the shelter

The dataset includes the following information about the animals:

- **AnimalID**: a unique identifier given to each animal as they enter the shelter.
- **Name**: the animal's name if it has been supplied by the owner. Approximately 30% of the animals' names are missing, meaning that they were picked up off the streets without a collar or identification information.
- **DateTime**: the date and time that the animal left the shelter
- **OutcomeType**: the outcome of each animal
    - This is what I will be predicting with my model
    - There are 5 possible values
        - Adopted
        - Returned to owner
        - Transferred (to another shelter)
        - Euthanized

    - Died[1]
- **OutcomeSubtype**: includes additional information about the animal's outcome. I.e. where an animal was transferred to (either to the SPCA or to a partner shelter), the reason that the animal was euthanized, where the animal went after being adopted (off-site or to a foster family)
- **AnimalType**: whether the animal is a dog or cat.
- **SexuponOutcome**: includes the animal's gender, as well as whether the animal has been spayed/neutered, or not.
- **AgeuponOutcome**: the age of the animal at the time of adoption.
- **Breed**: the breed of the animal.
- **Color**: the color of the animal.

Data Preparation

The following columns were **deleted** because I felt it did not provide valuable information to my prediction:

- Name
- OutcomeSubtype

Left **unchanged**:

- AnimalID
- OutcomeType
- Color

Prepared columns:

- DateTime

  The original column contained an unparsed date and time. I deleted the time the animal was adopted because I felt it was not helpful to the prediction, and only left the date (year, month, and day) that the animal left the shelter.

- SexuponOutcome

  This column originally contained the animal's gender and status on whether it was sterilized at the time of leaving the shelter. I split the column into two columns, one containing the gender of the animal and the other containing its sterilized status. I removed the rows where the status was unknown or had no value (this only accounted for 4% of the records).

[1] Please see "What the Model will not Predict"

- AgeuponOutcome

  This column originally contained a string value containing the age of the animal which I thought was unhelpful due to the large amount of unique values. My aim was to create a column with numeric values (5), not string values (e.g. 5 years). I firstly merged all rows that has the same meanings, e.g. year and years, month and months, etc. and thereafter split the integer and the string. Then I pivoted the string around the animal's ID and used the integer values to populate these columns. I removed the columns days and weeks and stored their ages as 0 months. Finally, I joined the years and months columns into a single age column, creating a integer value which represented the animal's age.

- Breed (see what the model will not predict)

  From this column, I created an additional column which stored Boolean values stating whether the animal was a purebred or a mixed breed.

I also used the **ensemble method** to create new features which I used in the prediction model. This is explained in detail in the methodology.

What the Model will not Predict

There are variables that the model will not focus on predicting and using in the prediction. These can be considered for future action by future data scientists who will work on this data.

Outcome: Died

Of the 10,000 records, only 65 records are for animals that have died. There are too few records on this for me to make an accurate prediction for this outcome and thus I will not focus on it in the model. Also, referring back to the main outcome of this project – I aim to create a model to predict whether animals will be adopted and if not give shelters an indication of which animals they should dedicate extra time to. Thus, the outcome 'died' is not of critical importance to the goal of this project. Lastly, after examining the original data, I can see that the majority of the animals that died were either under the age of 1 or over the age of 10 (figure 1) i.e. very young or very old and hence we can assume they died of natural causes. Therefore, this outcome is not likely for healthy, average aged animals – the majority of the animals in the shelter.

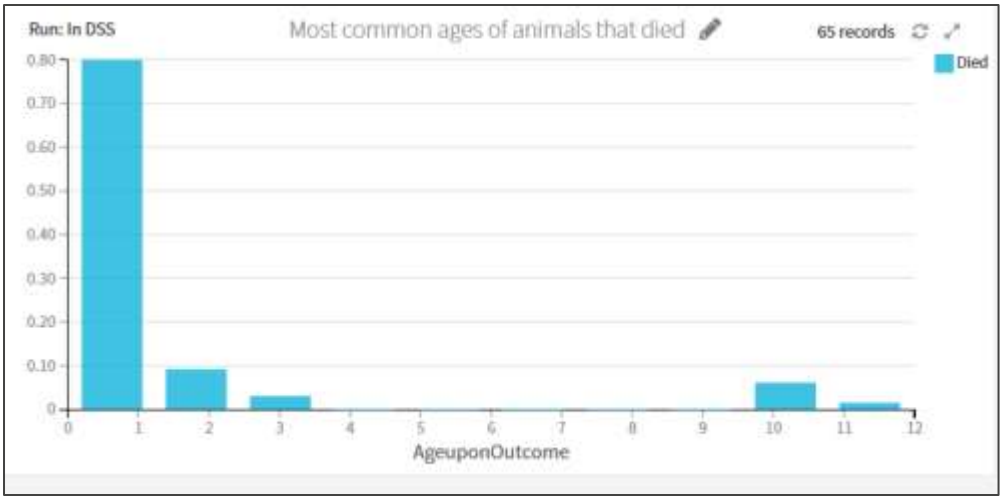

Figure 1: Outcome died analysis

Outcome based on ‘Breed’

I will be focusing on the age, gender, sterilized status, animal type, and whether the animal is a purebred or not when creating my model. The ‘breed’ column contains over 464 unique identifiers. Although I will use this column in creating the prediction model, I think the model could be improved if the breed column was given more time and consideration in preparation. External data containing information about various dog and cat breeds, including types of hair, temperaments, etc. could be used to categorize all the unique vales in the column currently to create a more accurate prediction model and give a better understanding of what breed of animals people are more likely to adopt.

Methodology

I began my project by preparing the train dataset with the above mentioned features. Once the data preparation phase was complete, I created a balanced prediction model with a Python backend to predict multiclass classifications of animal outcomes. In the design of the model, I used the ensemble method to generate A+B, A-B, and A*B pairs of numerical features to create new features on which the model would be trained. I used 5 algorithms to train the model, Logistic Regression, Random Forest, Artificial Neural Network, XGBoost, and Decision Tree. Once these models were finished training, I examined each to determine which would be the best to deploy to the flow to score against the test dataset. XGBoost had the highest score, with a score of 0.786, but as I had anticipated, the ‘Died’ outcome was not accurately predicted by the model due to the too few entries in the dataset and was hence omitted from the final prediction. The next best scoring model, Random Forest, with a score of 0.778, included the ‘Died’ prediction in its outcome, but incorrectly predicted the rest of the features generating a model where ‘Return-to-owner’ was the most frequent outcome, which is incorrect according to the original dataset. For this reason, I decided to use the XGBoost prediction and deploy this model to the flow and score it against the test dataset to get the final outcome prediction.

Dataiku Flow

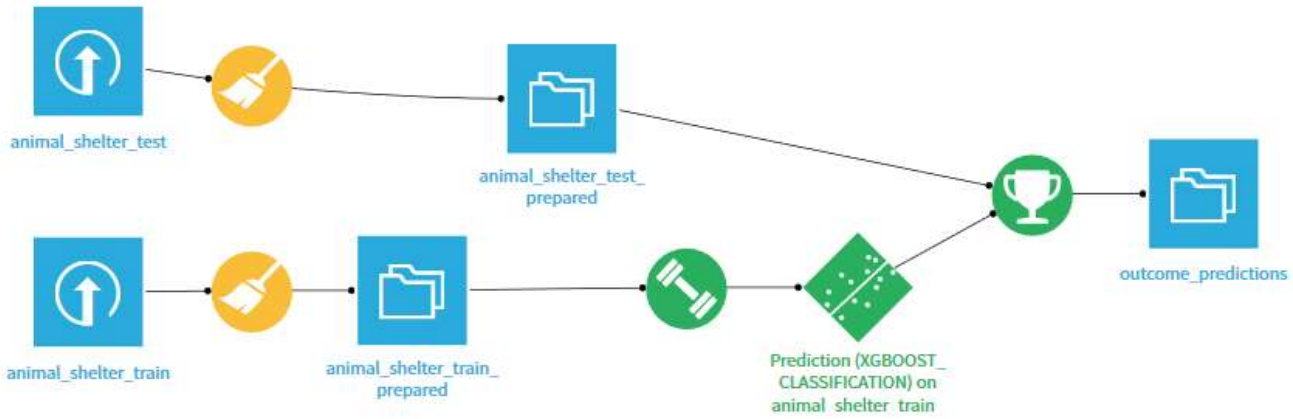

Figure 2: Dataiku Flow

Results

Final Dataset Prediction Comments

My final prediction included 4 possible outcomes: Adopted, Transferred, Returned to owner, and Euthanized. As I had thought, the model did not predict any animals to have died, but as previously explained this was due to the insufficient information on the ''died' outcome to accurately train my model.

The dashboard shown below, gives insights into which animals will be adopted and which not. The dashboard is explained in great detail in the sections below. Austin Animal Center (and other shelters) can use these graphical representations, to see which animals are predicted to need a little extra help in finding a new home.

Dashboard

Figure 3: Dashboard

What can Austin Animal Shelter do with this data?

The first graph (figure 3) shows the most likely outcome per animal type (i.e. dog or cat). From this graph we can see that both dogs and cats have reasonably good rates of adoption, but what is interesting to see is that cats are much more likely to be transferred than dogs, and dogs are much more likely to be returned to their owners than cats. Both animal types have more or less the same rates of euthanasia.

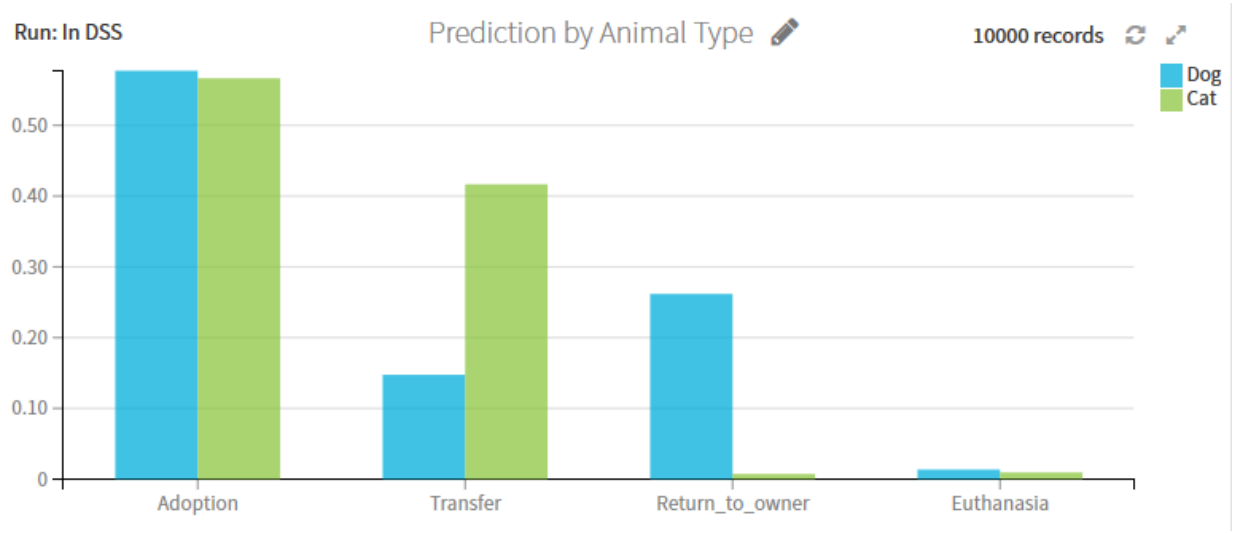

Figure 4: Outcome by animal type

The second graph (figure 4) shows the percentage of adoptions per month of the year. We can see that in December (month 12), the amount of adoptions is the highest where on the contrary, adoptions In April (month 4), is the lowest. With this information, the shelters can focus their attention on those months with the lowest predicted adoptions.

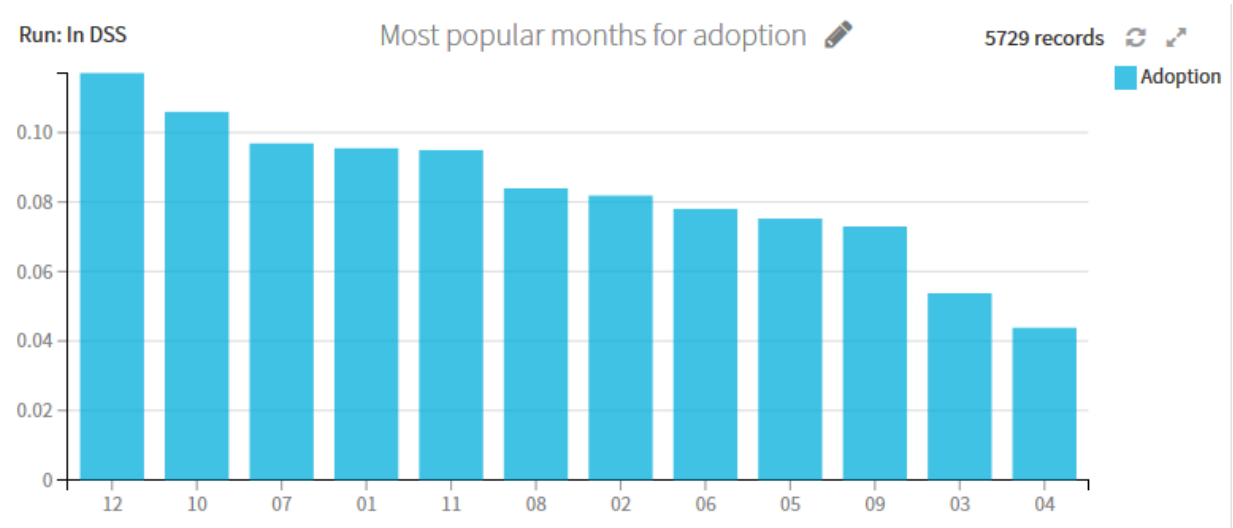

Figure 5: Most popular months for adoption

The following two graphs (figures 5 & 6) show the most likely outcomes per age group of animals. Once again we can see some interesting differences in the outcomes of each age group for dogs and cats. What is surprising to note is the amount of cats between the ages of 9 and 13 that get adopted whereas for dogs, the majority of the adoptions are between the ages 0 and 4. Furthermore, the number of dogs that are returned to their owners are much more than the number of cats who are returned to their owners, and cats are retuned mostly due to their age it would appear. Lastly, the rates of euthanasia among cats and dogs according to their ages are clearly shown – the older animals are more likely to be euthanized.

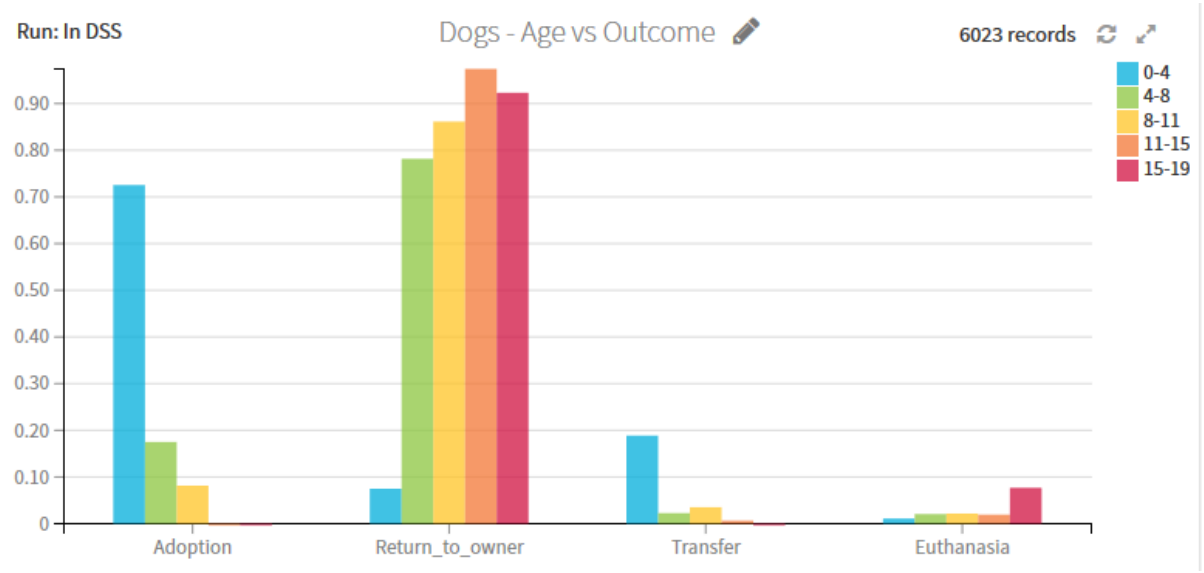

Figure 6: Dogs - age vs outcome

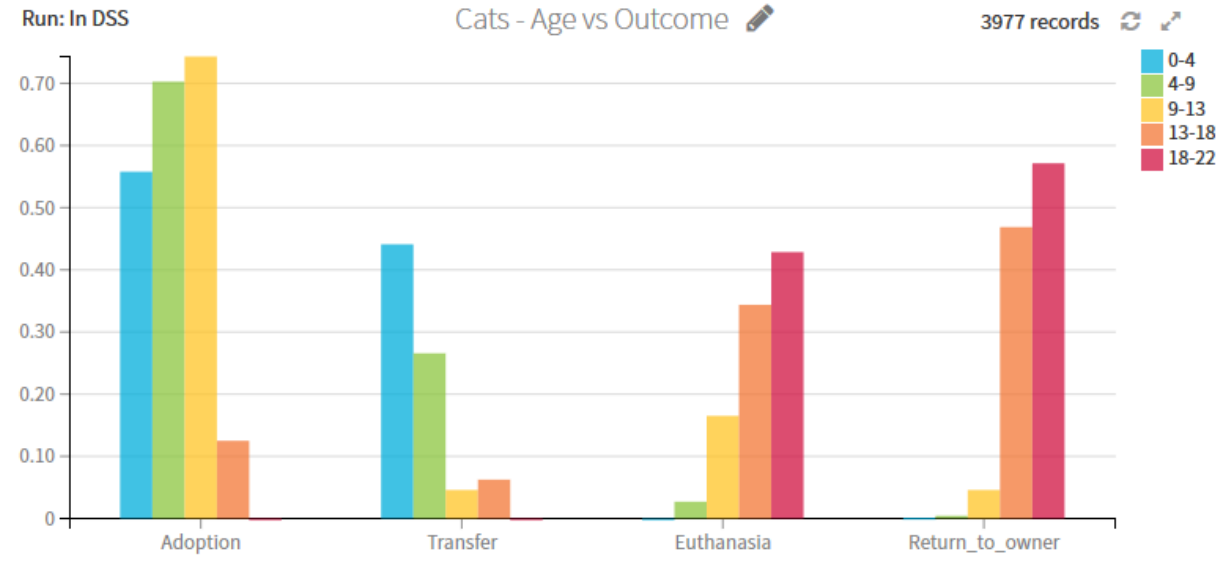

Figure 7: Cats - age vs outcome

Figure 7 shows the most likely outcome of animals by their gender and type. For example male dogs are more likely to be returned to their owners than what female dogs are, and female cats are less likely to be euthanized. Adoption and transfer rates are similar between the two genders, but return to owner and euthanasia has noticeable differences.

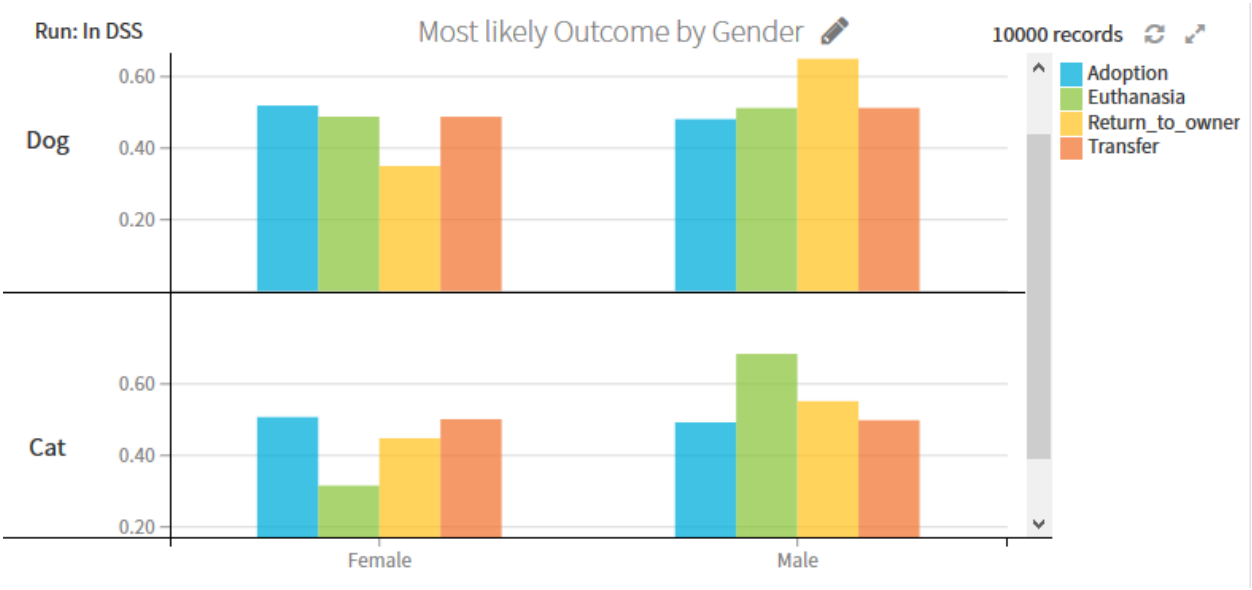

Figure 8: Most likely outcome by gender

Figure 8 shows how an animal's sterilization status affects their outcome. We can see that animals that have been spayed/neutered are much more likely to be adopted than those who have not been. Intact animals have the highest rates of transfer and euthanasia.

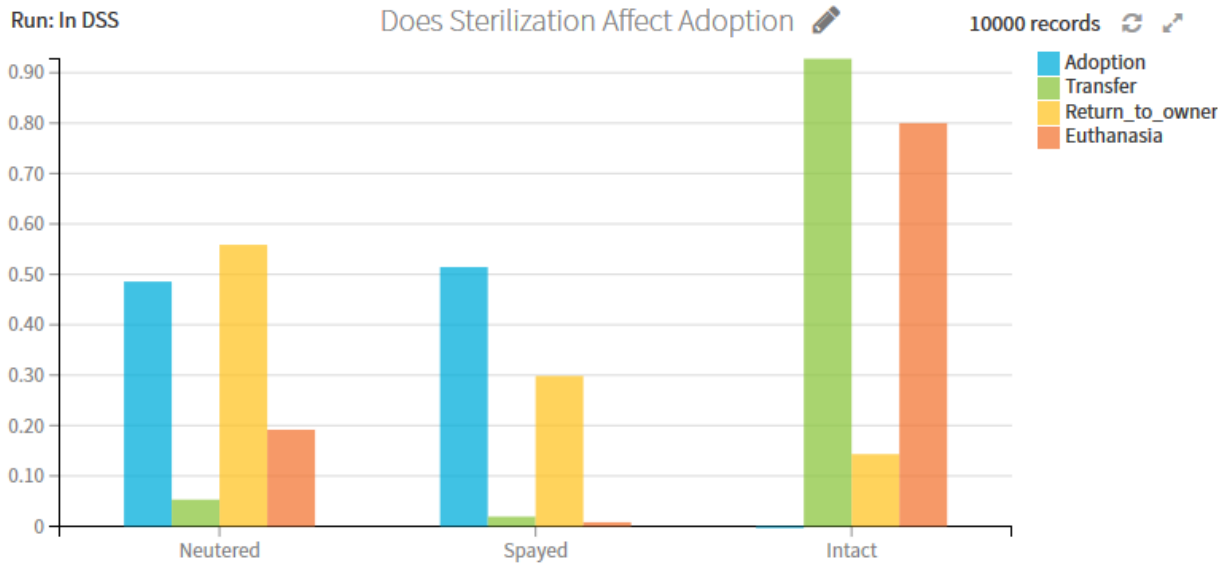

Figure 9: Sterilized status vs outcome

Ethical Implications

Firstly, the dataset only contains data about animals from one shelter in the U.S. – the Austin Animal Center. Animal outcomes and adoption rates could vary greatly from a) country to country, and b) even state to state in the U.S. Furthermore, there is no information about the animals' health conditions. This could be an important contributing factor in whether or not an animal is adopted and what outcome it is subjected to. And lastly, there are only dogs and cats in this dataset; what about all the other animals that end up in shelters – rabbits, guinea pigs, hamsters, etc., etc.

In terms of ethical implications affecting human beings, there are none because no information about the animal's original owners or new adopted family's is stored what so ever. However, storing information about the new families of the animals could possibly be helpful in helping the other animals find homes and loving families similar to those who fit the profile of animal adopters.

Suggestions for Future Action

Refine the many breeds

The breed column currently contains 464 unique identifiers. Classifying these unique breeds into categories for dog or cat types such as for example hunting, companion, working, etc, or even hair types or animal temperaments, could provide some useful information which could be fed into the predictive model and used when determining which dogs and cats are more likely to be adopted.

Social Media trends

Add external data from social media like Facebook, Twitter, and/or Instagram to see which dogs and cats people are posting/tweeting/speaking about. This could give an indication into which dogs and cats are very popular among people and the shelter could use this knowledge during popular adoption times to both help boost the popularity of the other, less popular, breeds, as well as try and get the popular breeds all adopted.

You could also identify what type of people are interested in which breeds of dogs and cats and target specific groups of people when promoting adoptions.