Methodology

Data Cleaning

Original animal shelter data comes from Dallas OpenData. There was one CSV per year from 2015-2020. This means there were several data points from 2014 accounting for the Intake Dates prior to January 1st, 2015 all the way up until when we started this project on August 30th, 2020. The names and headings of each CSV were edited so they could be merged into a single continuous data frame in a 'for loop.' This was then filtered to include only meaningful columns. This includes:

- Animal ID
- Animal Type (Cat, Dog, etc)
- Animal Breed (Corgi, Pug, etc)
- Animal Origin (Field, Over the Counter)
- Census Tract (GeoData: ###.##, Zip, Mapsco)
- Chip Status (Chip, No Chip, Unable to Scan)
- Intake Date
- Outcome Date
- Intake Condition (Health status)
- Outcome Condition
- Outcome Type (Adopted, Euthanized, etc)

Each animal's length of stay was calculated by subtracting the Intake Date from the Outcome Date. Because some animals had stays approaching 2 years, outliers were filtered out. Outliers were defined as any stay greater than 60 days. While this range was well above the 75th percentile + 1.5*IQR (20 Days), we felt that 2 months was an unusual but reasonable amount of time to spend in a shelter.

Starting with 211,216 data points, we excluded 2,113 animals who were in the shelter longer than 60 days.

The responses for each column were then limited to reduce the number of categories by using the .str.contains() function in python.

- Animal Origin
  - Field
  - Over the Counter
- Chip Status
  - Chip
  - No Chip
  - Unable to Scan
- Intake Condition
  - Healthy
  - Contagious
  - Rehabilitable Non-Contagious
  - Manageable Non-Contagious
  - Untreatable Non-Contagious
- Outcome Type

    - Adopted
    - Euthanized
    - Returned to Owner
    - Transferred

Intake types of untreatable & contagious, and dead on arrival were removed, as they would not be considered release. This accounted for the removal of 9,571 more animals bringing us to 199,532.

The data was further split into two more data frames, one for cats containing 50,593 data points, and one for dogs containing 149,971. Other types of animals that were excluded for study are 'birds, wildlife, livestock.' As much as we would like to know more about the scorpion that was dropped off at a shelter, it is outside the scope of this study.

Of the cat and dog breeds from each data frame, they were combined into broader categories. Cats were split into categories based off of their hair length. Dogs were organized to include broader breeds that include ~5,000 animals. Corgi's and Pug's were included because they are the class mascots. All other dog breeds were grouped as 'Other.'

- Cat Breeds
    - Short Hair
    - Medium Hair
    - Long Hair
- Dog Breeds
    - Chihuahua
    - Corgi
    - Hound
    - Pitbull
    - Pug
    - Retriever
    - Shepherd
    - Terrier
    - Other

With Tableau the goal was to use the latitude and longitude in the data provided to create a map. We ran into an issue that persisted with our geographic data. The "Census Tract" column contained three varieties of data to include Mapsco (##A), Zip Codes (#####) and Census Tracts (###.##). Our team had divided our responsibilities into segments and one of the two Tableau individuals was able to capitalize on leveraging the Zip Code data successfully. Unfortunately this accounted for just 7,044 of the 180,162 data points.

In an effort to get more of the full data set involved on our maps, further cleaning was needed. Any data in the "Census Tract" column containing a letter was removed, bringing us from 180,162 to 152,790. In an effort to get in the Census Tract format of (###.##), the column was then divided by 100. We then performed an "inner" merge the catdog.csv and censustract.csv. This removed the likely 7,044 remaining Zip Codes from the data set, leaving us with 145,746 data points. This was unable to be utilized at time of production.

Modeling

To create predictive models for both the cat and dog data, we first established what we would like to predict: outcomes. Dallas Shelters took in 211,216 animals between 2015 and July 2020. To help determine how resources should be allocated, early knowledge of whether an animal is likely to be adopted or euthanized could be helpful. This is why we chose to predict outcome type.

To predict outcome type, this column was encoded with the following labels:

1. Adopted
2. Euthanized
3. Returned to Owner
4. Transferred or Fostered

The features for our model were largely categorical. To prevent limit bias, we encoded these categories using 'One-Hot Encoding.' This encoding was used for these columns:

- Animal Breed
- Origin
- Chip Status
- Intake Condition

The resulting new columns were joined with our only quantitative variable, length of stay. These values of both the categorical and quantitative variables were then scaled using sklearn's StandardScaler.

The cat and dog data were then split using sklearns train_test_split, and passed through the following machine learning models:

- Logistic Regression
- K-Neighbors Classifier
- Decision Tree Classifier
- Bagging Classifier
- Random Forest Classifier
- Ada Boost Classifier
- Gradient Boost Classifier
- eXtreme Gradient Boost Classifier

While each model has its strengths and weaknesses, we believe that, for both cats and dogs, the Gradient Boost Classifier works the best. It had similar accuracy scores (0.58 for cats, 0.59 for dogs) with the other tree ensemble models but had better F1 scores than the rest of the models. It also falsely predicted euthanization at lower rates than the other models. While it did predict more euthanized animals would be adopted than other models those models had lower F1 scores for predicting euthanization.

Cat Gradient Boost Classification Analysis Report

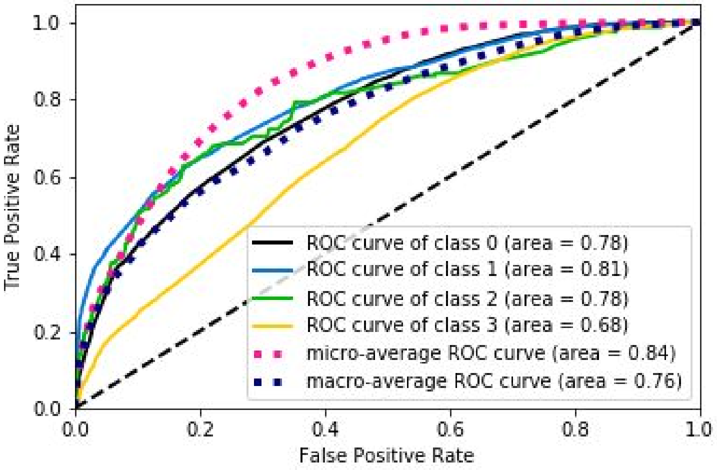
Outcome	Precision	Recall	F1-Score	Support
Adopted	0.60	0.50	0.55	3511

Euthanized	0.74	0.40	0.52	2585
Returned to Owner	0.31	0.03	0.06	159
Transferred	0.53	0.75	0.62	4710

Cat Confusion Matrix

Actual Outcome	Predict Adopted	Predict Euthanized	Predict Return	Predict Transfer
Adopted	1770	45	2	1694
Euthanized	210	1024	6	1345
Returned to Owner	68	22	5	64
Transferred	883	300	3	3524

Cat ROC Curves



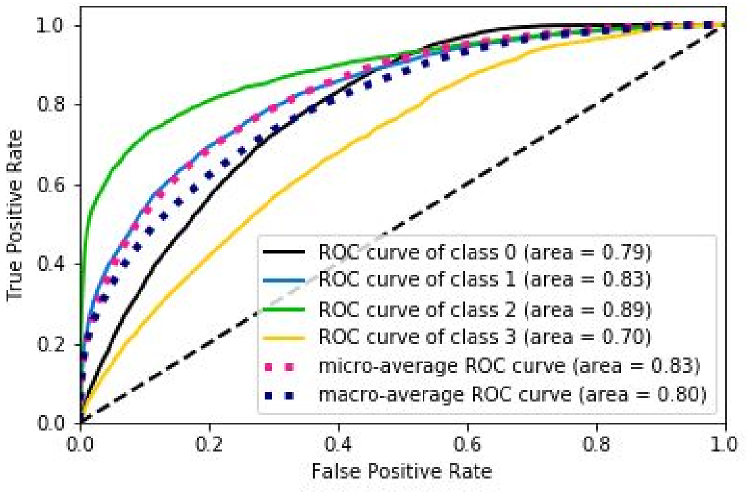
Dog Gradient Boost Classification Analysis Report

Outcome	Precision	Recall	F1-Score	Support
Adopted	0.54	0.85	0.66	12416
Euthanized	0.60	0.51	0.55	7366
Returned to Owner	0.76	0.66	0.71	7920
Transferred	0.43	0.09	0.15	6374

Dog Confusion Matrix

Actual Outcome	Predict Adopted	Predict Euthanized	Predict Return	Predict Transfer
Adopted	10570	967	566	313
Euthanized	2690	3756	642	278
Returned to Owner	1909	638	5229	144
Transferred	4496	862	459	557

Dog ROC Curves



Model features varied in importance between cats and dogs. While length of stay was the most important feature for both (0.26 for cats, 0.40 for dogs), the origin of a dog seems to matter more than its health, and the opposite for cats. Breed matters more for dogs, while breed was the less important feature for cats.