Abstract

The focus of this project was to analyze the Dallas OpenData spreadsheets from January 2015 – August 2020 and to explore the outcome of the animals that enter the Dallas Shelter and their location. The dataset included over 200,000 animals. This dataset was filtered down to eliminate non-viable animals, (dead on arrival, or both found untreatable and contagious), also the animals that are recorded in the shelter for over 60 days. The data was later analyzed and visualized using the tools provided by Tableau. To make the predictive models, the data was divided into cats and dogs and then built using SciKit-Learn's 'Gradient Boost Classifier' so they can predict live models on the chosen values by users.

Introduction

Approximately 6.5 million animals pass through U.S. animal shelters every year. Of these animals, 3.2 million end up adopted, while 1.5 million are euthanized. As city and county budgets are shrinking in the current economic downturn, being able to determine which animals are likely to be adopted vs which ones are likely to be euthanized can help ease the strain on dwindling resources. The knowledge of these statistics is important because it provides the understanding of what happens to the animals and how under-funded cities deal with the excess of animals in their shelters.

In the timespan of January 2015 to August 2020 the county of Dallas had over 200,000 animals that either are in shelters or have passed through (adopted or euthanized). Majority of the animals are dogs and cats; they also hold multiple breeds of these animals and not all the breeds have the same outcomes. The data from the shelters will be analyzed and filtered down to see the discrepancies between not only cat and dogs, but also the breeds and their outcomes (adopted, euthanized, or returned to the owner) in the shelter. The findings will be input in a machine learning algorithm and shown on an app that can be manipulated by users to see the predictions.

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.

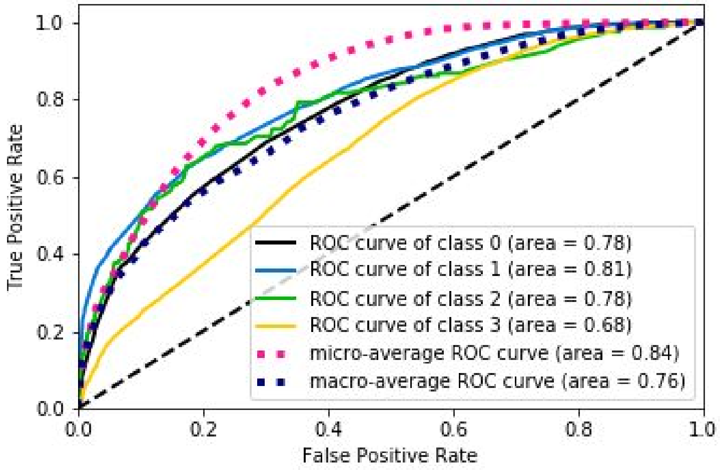
Cat Gradient Boost Classification Analysis Report

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



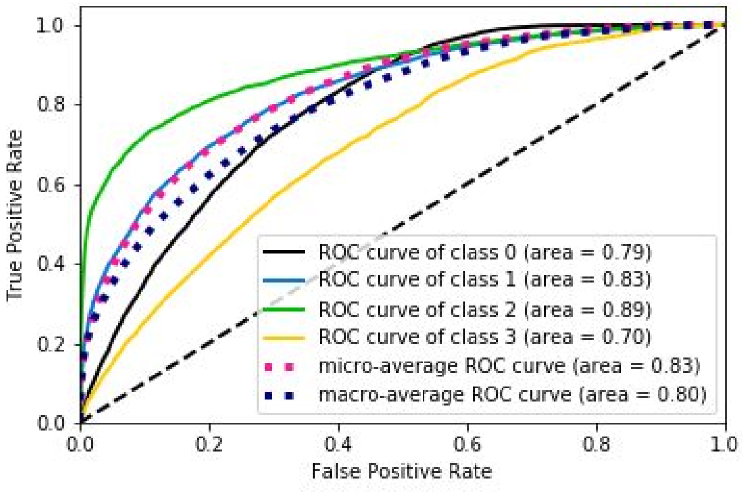
Dog Gradient Boost Classification Analysis Report

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.

Analysis

Figure 1 displays 184,338 sheltered animals from January 2015 to July 2020, the total amount for that time.

From this graph we see that there has been a steady increase of animal intake from 2016 through 2019 with a sharp decline during 2020. This decline may be due to the current Covid-19 quarantine with people having more time to keep up with the needs of a pet and because of the need for companionship.

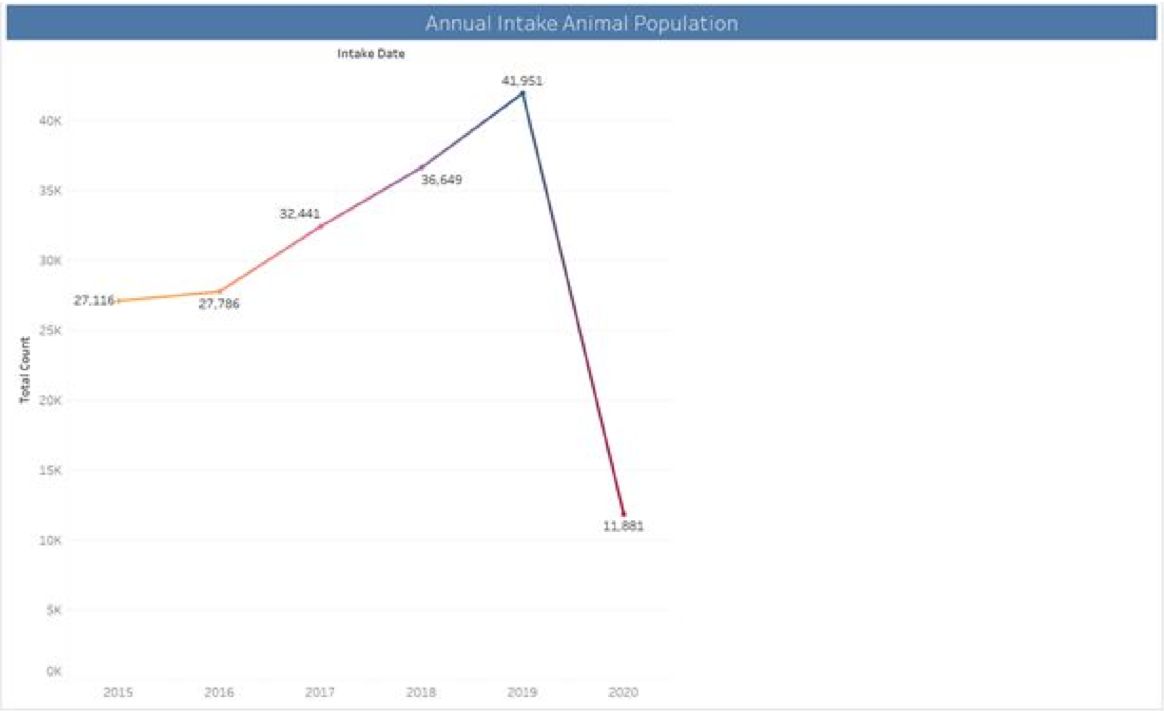


Fig. 1

Figure 2 displays the breakdown of animals in the shelter by animal type.

The bar chart shows that dogs make up three quarters of the total animals impounded while cats account for almost a quarter, with a mix of other animals making up the difference. Based on this information, the likelihood of a dog being impounded is three times higher than any other animal type.

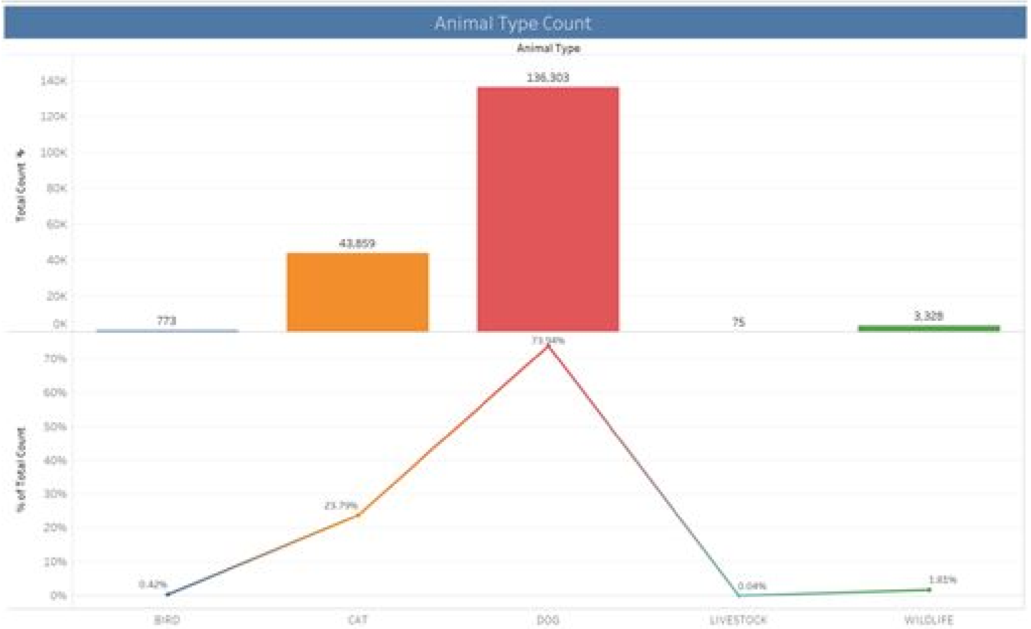


Fig. 2

Figure 3 displays the outcome of animals in the shelter by animal type.

The pie chart below shows that out of all animals impounded, dogs are more likely to be adopted and or returned to owner compared to cats. While this looks promising, this still leaves a high percentage of dogs that were euthanized.

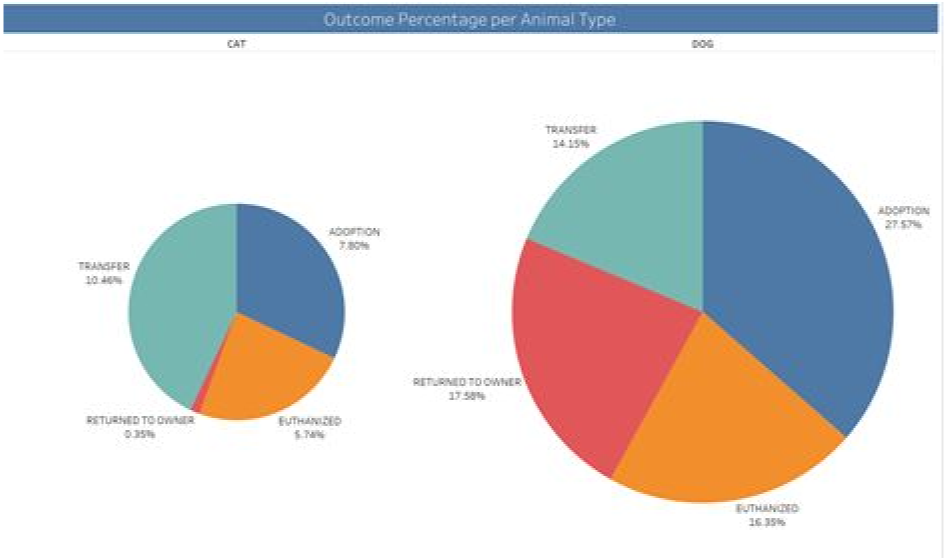


Fig.3

Figure 4 displays the shelter's dog count per breed.

The bar chart shows that Pitbull breed make up the population of dogs most impounded while the least most impounded dog population is Corgis and Pugs. The reason Pit Bulls may end up in shelters is due to fear of aggressive behavior. In general, larger dogs can be intimidating even to their owners, and the price tag to keep up with feeding and housing them may be more than imagined for many inexperienced owners.

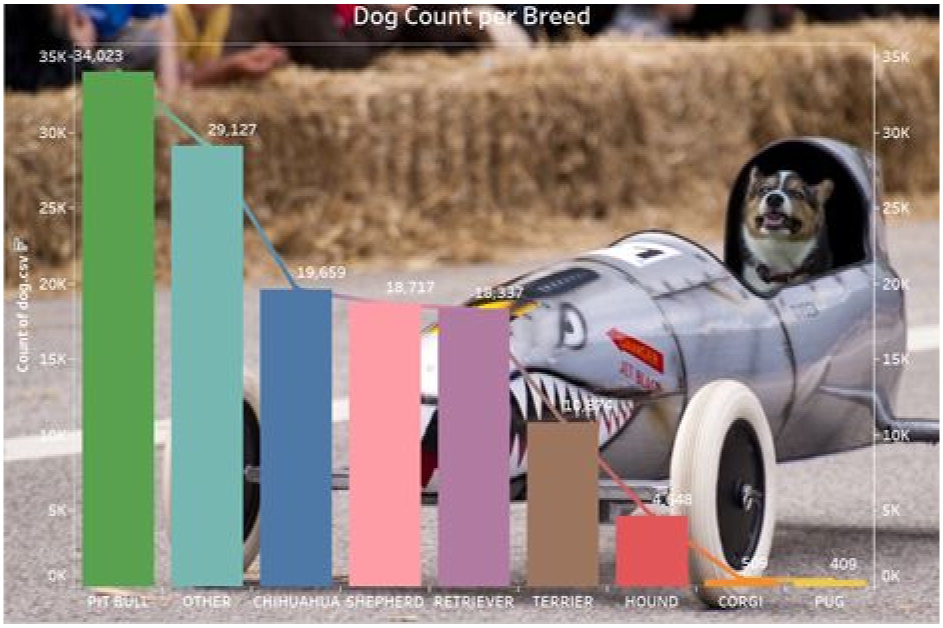


Fig. 4

Figure 5 displays the count of dogs impounded with a tracking microchip and the outcome type.

The pie chart below displays the count and percentage of dogs with or without tracking microchips. This shows that most dogs impounded were not identified as having an owner. The bar chart shows that of the 136,303 dogs, 49,663 (36.44%) were adopted, 31,681(23.24%) returned to owner, and 29,464 (21.62%) were euthanized. The number of dogs returned to owner closely correlates to the number of dogs with tracking chips, which points to a theory that more use of tracking chips may increase the likelihood of dogs finding their way back to their owners.

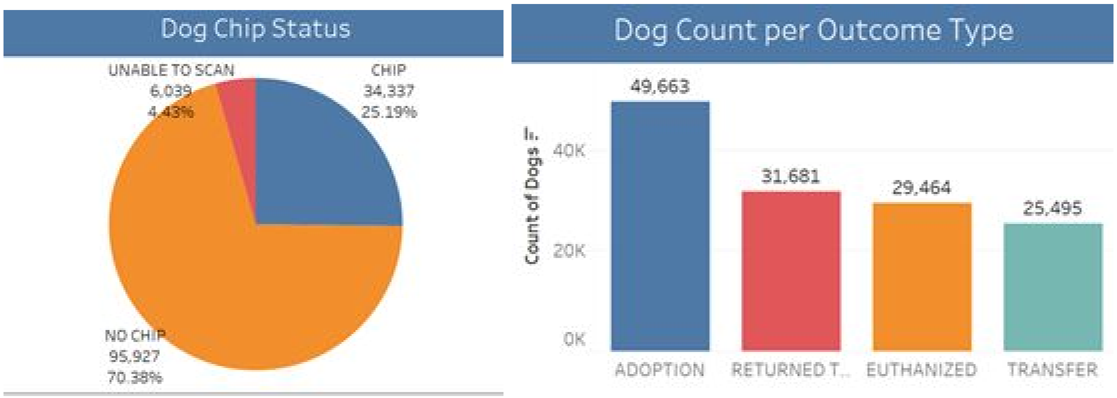


Fig. 5

Figure 6 displays the count of cats impounded with a tracking microchip and outcome per breed.

The pie chart shows that under 10% of cats impounded have a tracking microchip. The plot shows that of the cats that were impounded less than 2% were returned to owner. Based on this we theorize that a higher number of cats with microchip may increase the likelihood of an increase in returned cats to the owner.

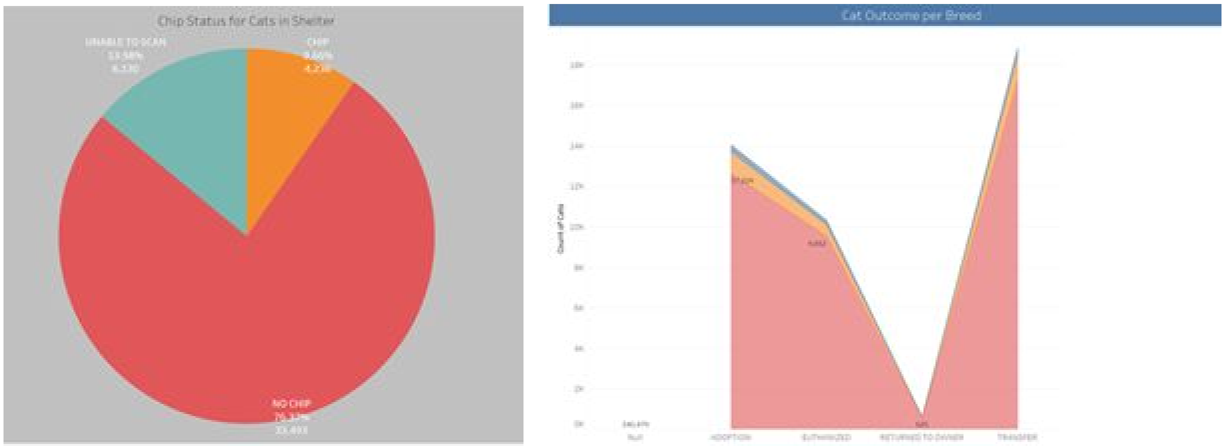


Fig. 6

Limitations

We experienced some limitations when it came to our modeling accuracy, quality of data and our geodata.

There was not a high degree of accuracy within the model as it only made a correct prediction 59% and 58% of the time for dogs and cats respectively.

We uncovered a few limitations within how our data was collected at the Animal Shelter itself. We were using data from 2015 through to August 2020. It appears that sometime in 2015, they transitioned from using Mapsco format or Zip Codes to the more accurate Census Tract format to track where the animals were found upon admission to the shelter. The difficulty we encountered was that all three types of data (Mapsco, Zip Code, and Census Tract) were all found in the Census Tract column of the dataset.

We were limited when it came to using the maps within Tableau. The maps contain different colored polygons that make up the Dallas metroplex and they are in fact Zip Codes. In a perfect world there would be nearly double that many polygons and nearly twenty five times the data to fill them.

Future Work

Although the provided modeling, methodologies and analysis are quite good and establish a competent baseline, there is always room for improvement.

As mentioned a little bit earlier, we ran out of time to get the full scope of usable data involved because of our troubles with geography. The goal moving forward will be to expand upon the work done to outline the Zip Codes on the map. Census Tracts are broken up into polygons with many segments and points. Zip Codes are outnumbered by Census Tracts nearly 2:1, making Census Tract a more geographically significant piece of data.

Other Cons About Zip Codes

- They may be created or eliminated at any time
- They exist only where U.S. Mail service is provided
- They are made up of groups of lines whose exact structural definition is not officially established
- A population of one Zip Code can exceed 100,000 (Census Tract averages 4,000)
- The total area of land (or water) is not known

So as you can see, it is important that we tap into the geographic advantage of using not just Zip Codes, but Census Tracts as well. This will allow us to map the other 90% of the data set and map it more accurately as well.

We will implement Web-Scraping on a monthly basis, as this is how often the CSV files are published on Dallas OpenData. This data will then be incorporated into our Modeling, Tableau Visualizations, and ultimately our application.

Breaking the data up into the categories of Cat and Dog was relatively simple. In the future, our group will parse the data into finer groups, like breeds of dog for example. Instead of modeling just by animal type, Cat / Dog, we will model by breed, Pitbull / Retriever / Corgi / etc.

When it came to modeling the data, we leveraged eight supervised machine learning models. We elected to use the Gradient Boosting machine learning technique because of its optimal Accuracy and F1 Scores. F1 is the weighted average of the Precision (ratio of correctly predicted positive observations to the total predicted positive observations) and the Recall (or Sensitivity, is the ratio of correctly predicted positive observations to all the observations in the affirmative). In the future, as more data is accumulated and we introduce modeling by breed, we will reevaluate which Machine Learning Algorithm will be best suited to predict the outcome of the animals that visit the Dallas Animal Shelter.

Conclusion

The ability to predict outcomes of a variety of animals is critically important in a climate of shrinking municipal, state, and federal budgets. The sharp decline in oil prices and spending in Texas will likely put extra strain on public services such as Dallas Animal Services. and their affiliated shelters. Luckily, our data shows a large decline in animal populations passing through Dallas shelters. This should alleviate some budgetary pressure.

Our models for cats and dogs are able to produce with some accuracy whether an animal will be adopted, euthanized, returned to its owner, or transferred to another facility or foster home. While these models are likely not accurate enough to determine whether or not to euthanize an animal on intake, it can be used to triage animals and more effectively allocate resources.

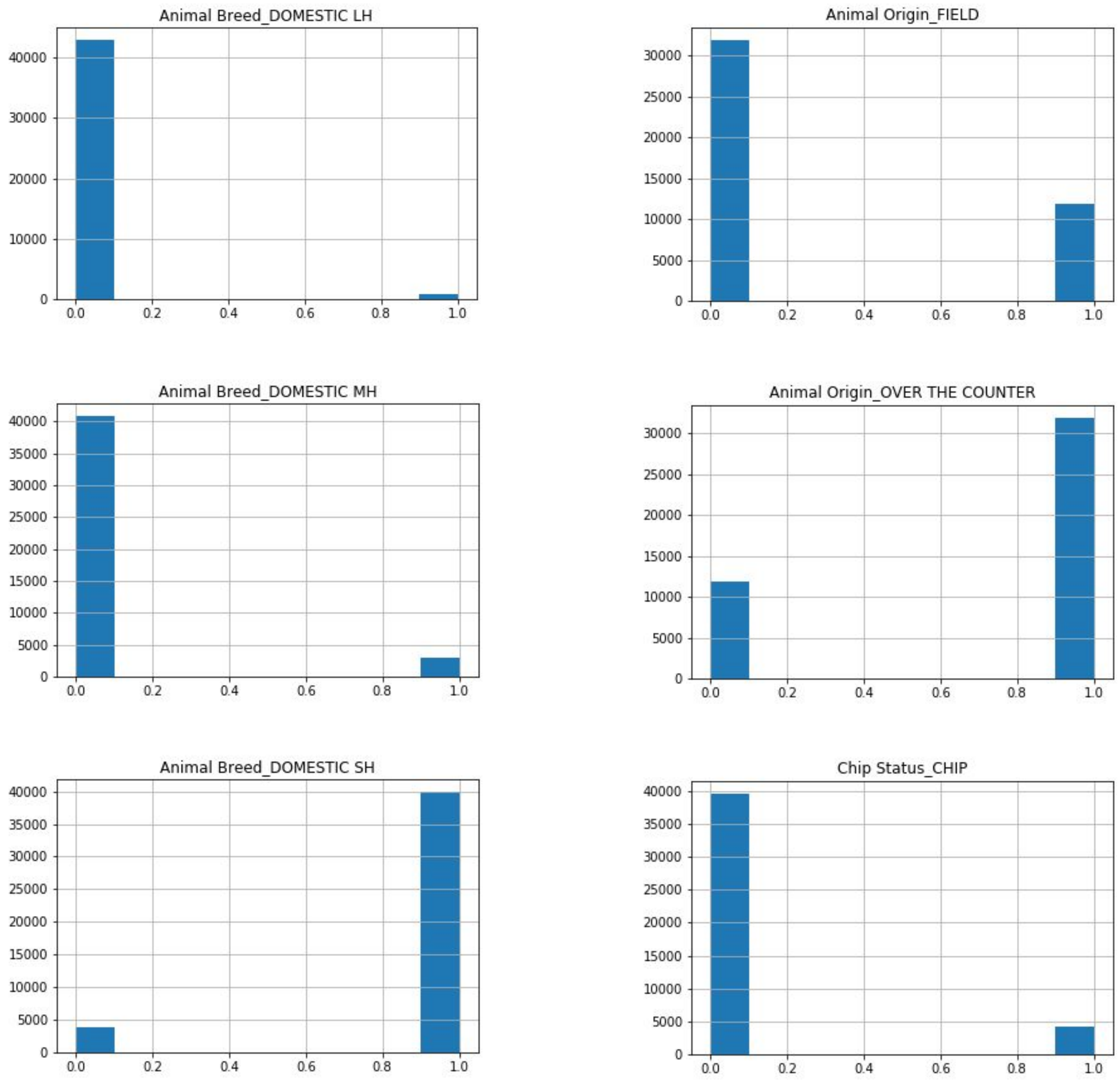
Even if the model is not used, the feature importance of each model still can help in allocating resources. The most important feature of each model is an animal's length of stay. This import in determining the return value of an animal's extended stay. Our features also suggest that a dog's origin is a better classifier than it's health, while the opposite is true for cats.

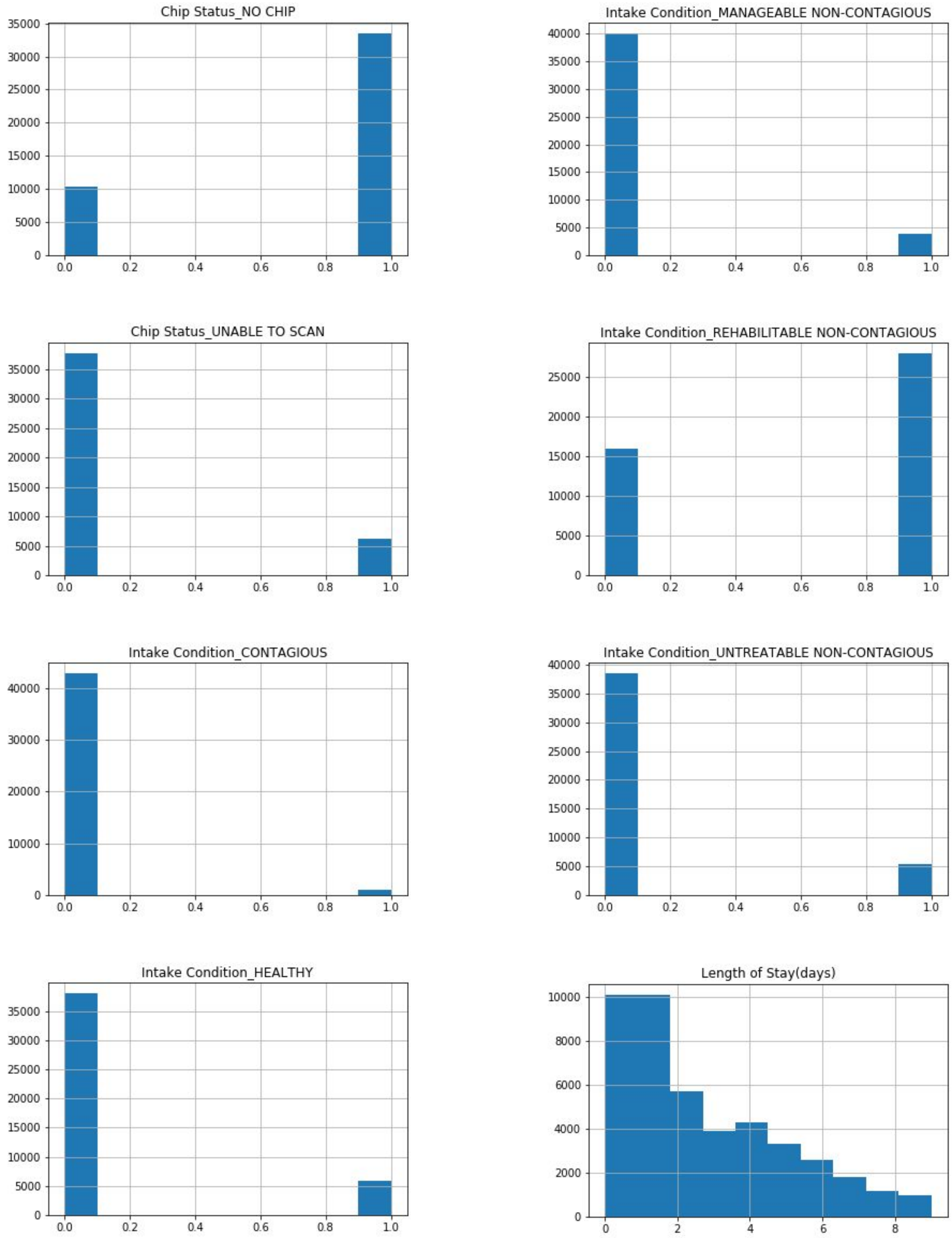
Appx

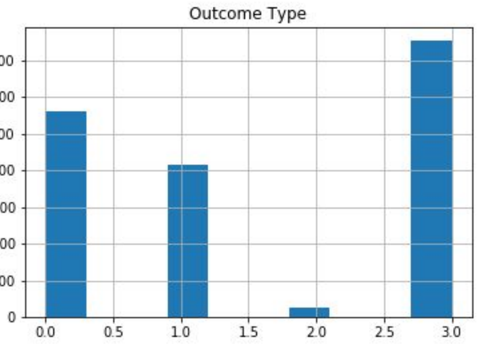
Statistical Analysis

Cat Modeling

Distribution by Classifier

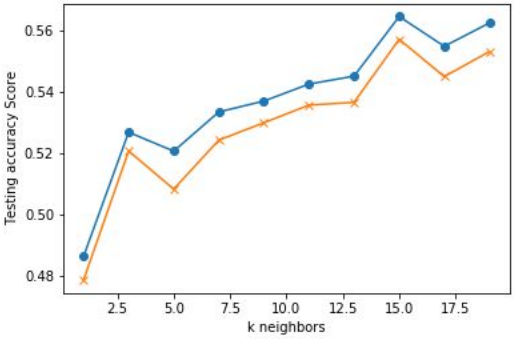



Chip Status_NO CHIP
0
5000
10000
15000
20000
25000
30000
35000
0.0
0.2
0.4
0.6
0.8
1.0
Intake Condition_MANAGEABLE NON-CONTAGIOUS
0
5000
10000
15000
20000
25000
30000
35000
40000
0.0
0.2
0.4
0.6
0.8
1.0
Chip Status_UNABLE TO SCAN
0
5000
10000
15000
20000
25000
30000
35000
0.0
0.2
0.4
0.6
0.8
1.0
Intake Condition_REHABILITABLE NON-CONTAGIOUS
0
5000
10000
15000
20000
25000
0.0
0.2
0.4
0.6
0.8
1.0
Intake Condition_CONTAGIOUS
0
10000
20000
30000
40000
0.0
0.2
0.4
0.6
0.8
1.0
Intake Condition_UNTREATABLE NON-CONTAGIOUS
0
5000
10000
15000
20000
25000
30000
35000
40000
0.0
0.2
0.4
0.6
0.8
1.0
Intake Condition_HEALTHY
0
5000
10000
15000
20000
25000
30000
35000
0.0
0.2
0.4
0.6
0.8
1.0
Length of Stay(days)
0
2000
4000
6000
8000
10000
0
2
4
6
8

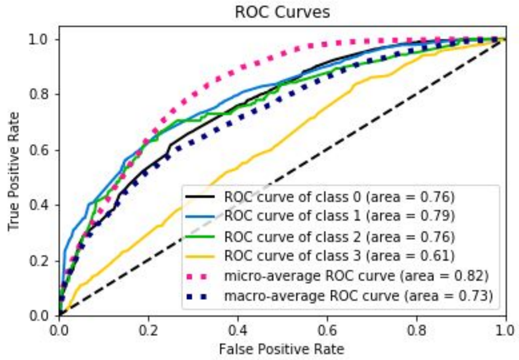


ROC Curves by model

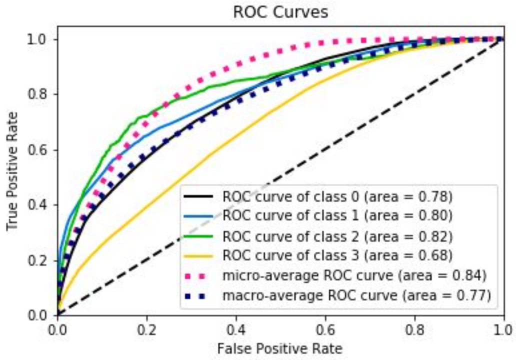
K-Neighors Elbow Curve



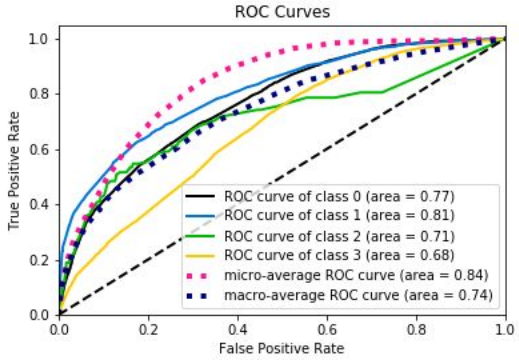
Ada Boost Model



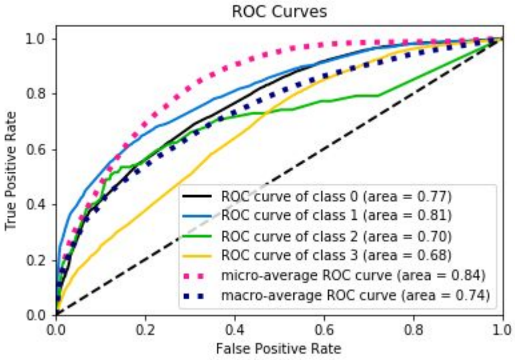
Final Production Model



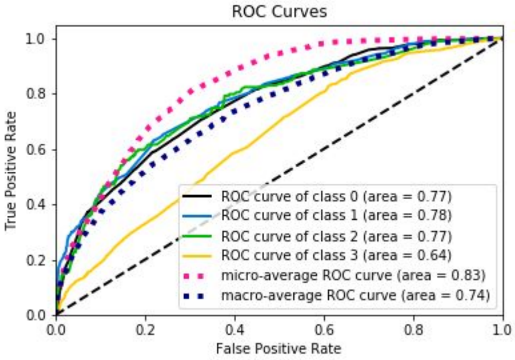
Bagging Model



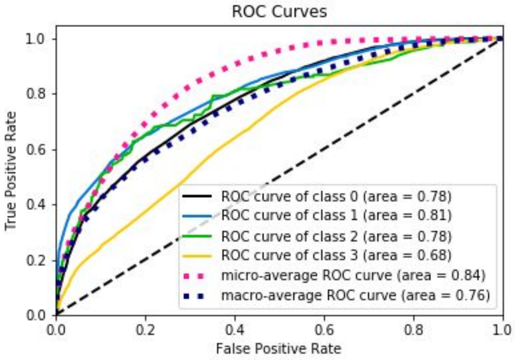
Decision Tree Model



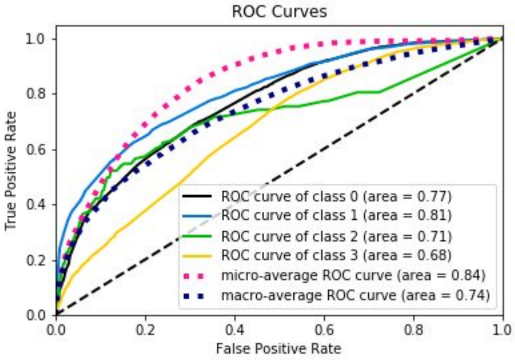
Logistic Regression Model



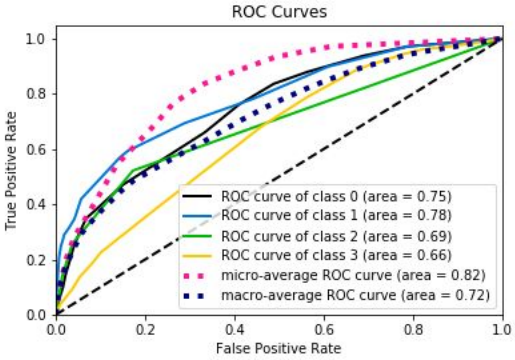
Gradient Boost Model



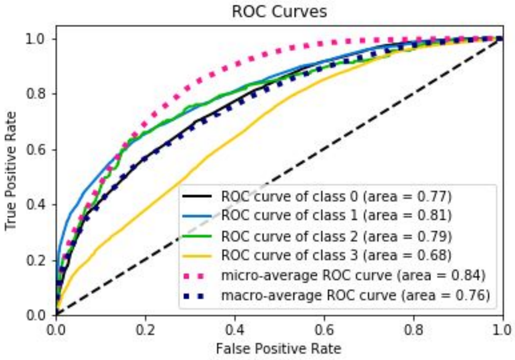
Random Forest Model



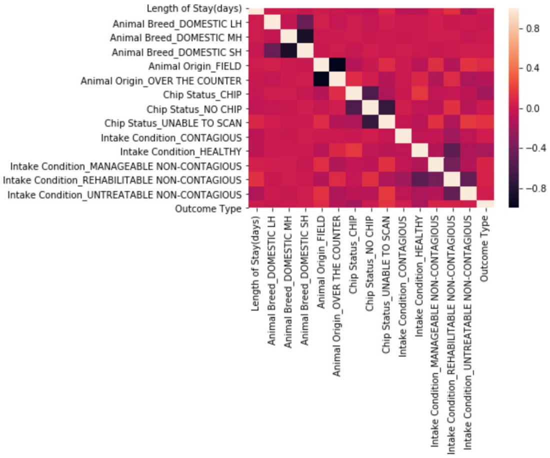
K-Neighbors Model



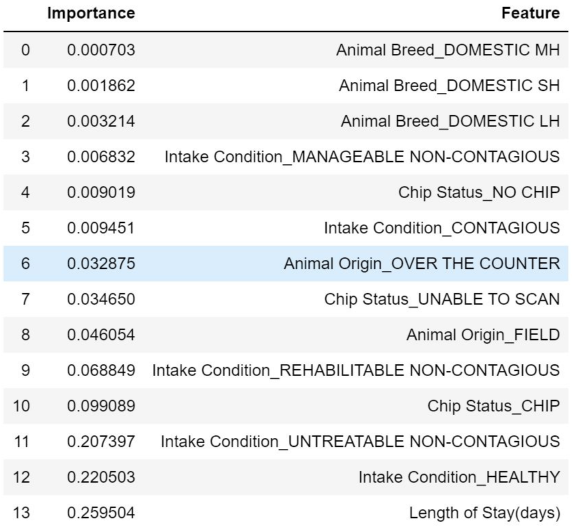
eXtreme Gradient Boost Model



Correlation Heatmap



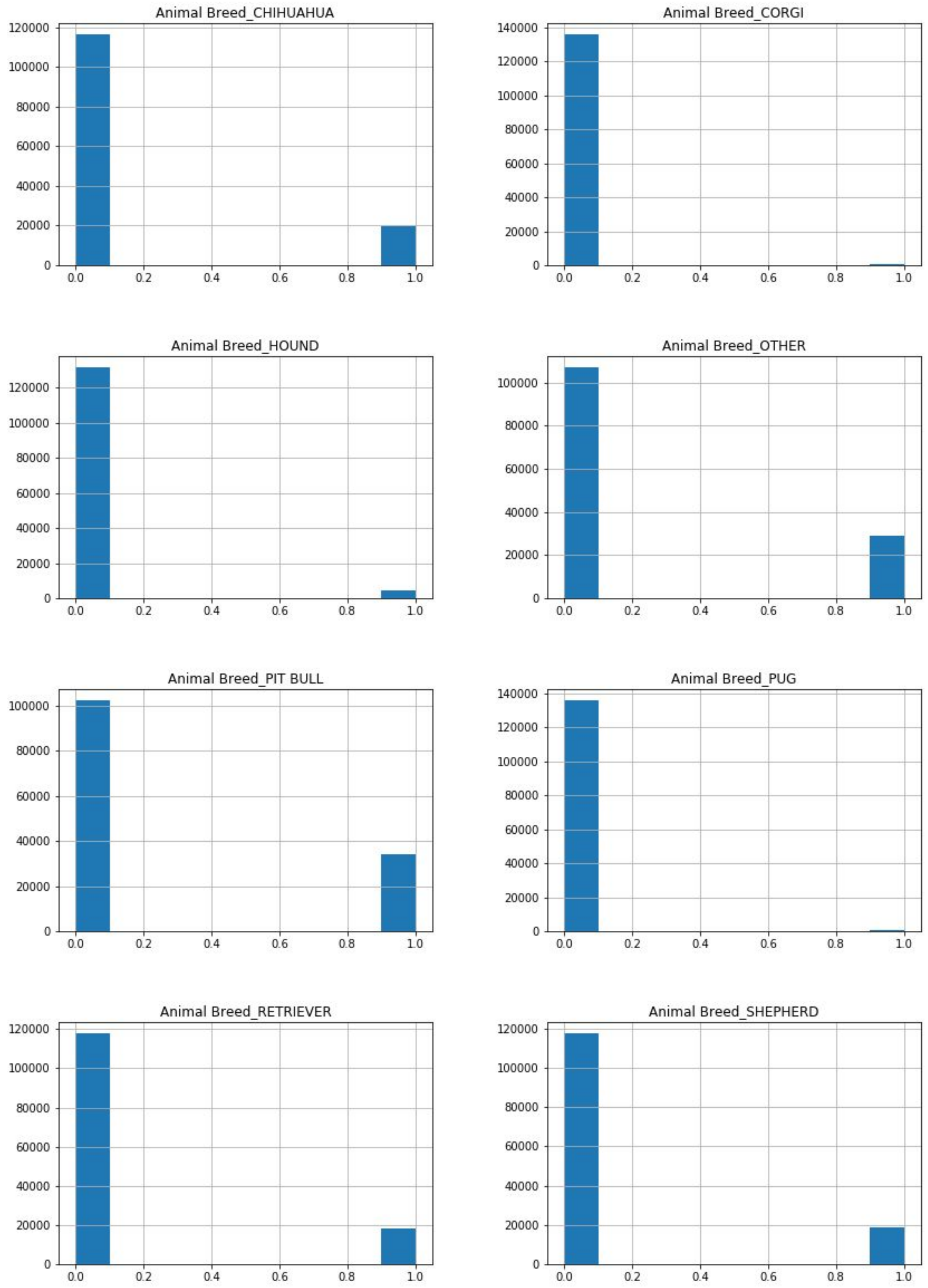
Feature Importance

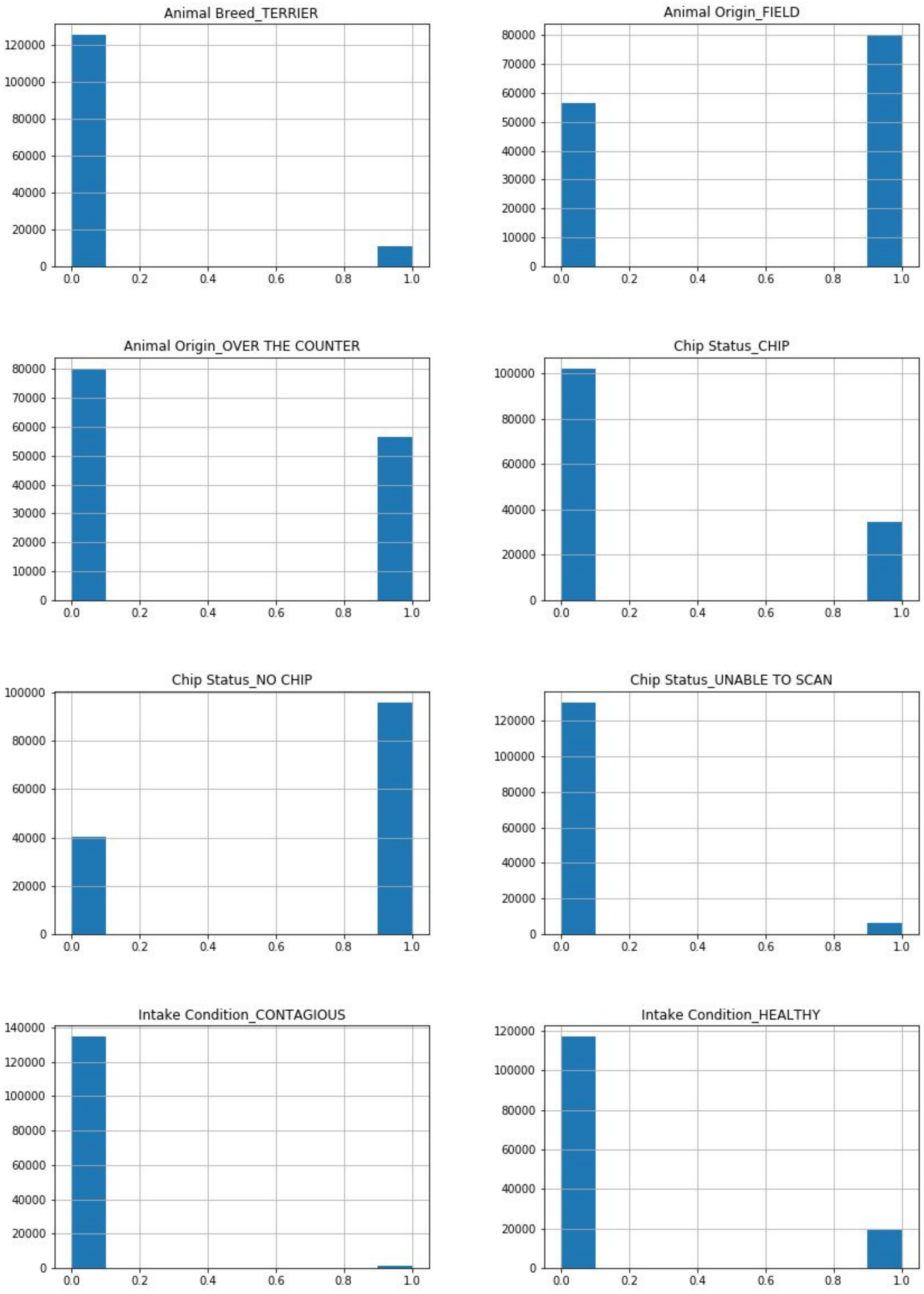


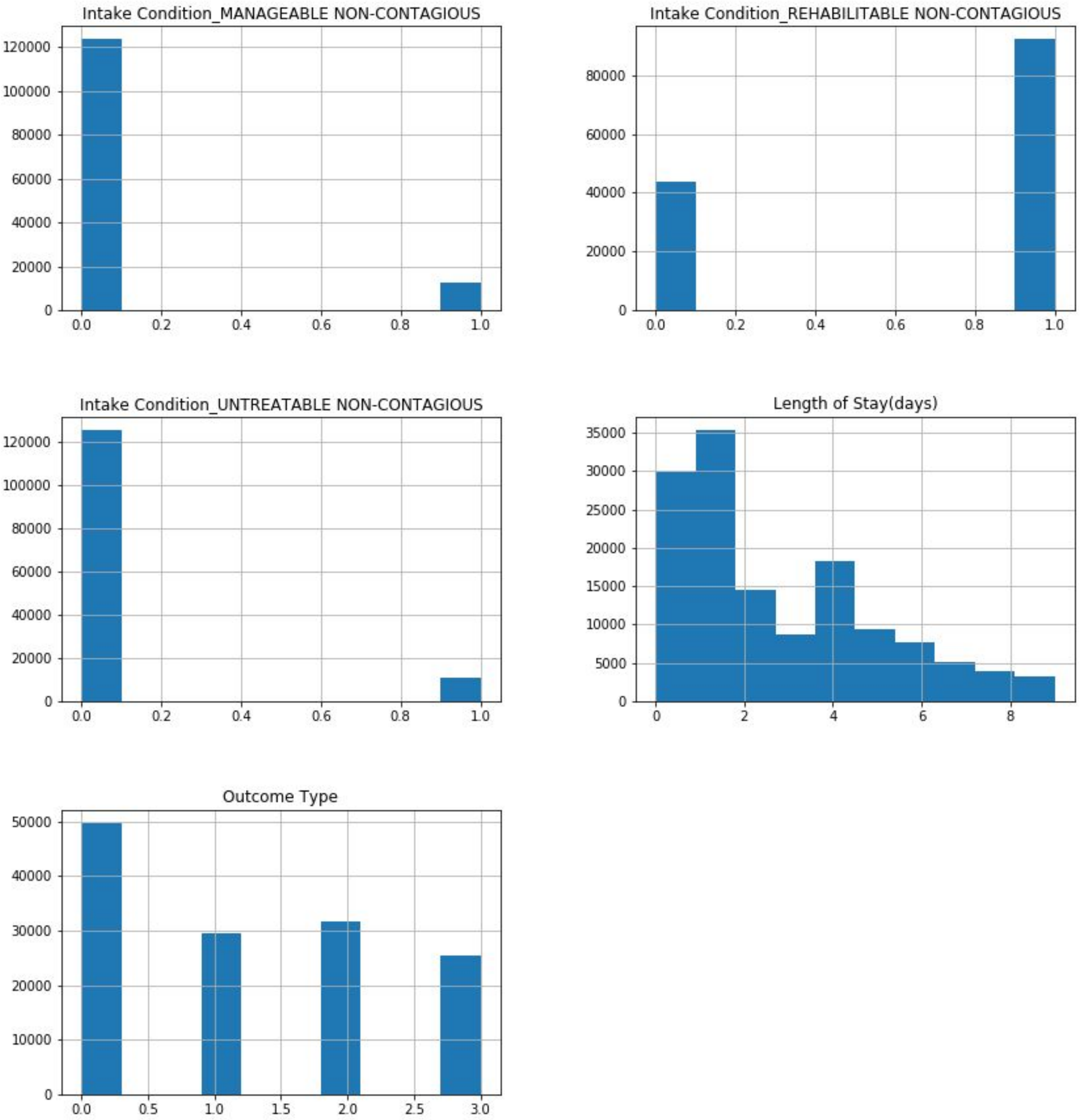
	Importance	Feature
0	0.000703	Animal Breed_DOMESTIC MH
1	0.001862	Animal Breed_DOMESTIC SH
2	0.003214	Animal Breed_DOMESTIC LH
3	0.006832	Intake Condition_MANAGEABLE NON-CONTAGIOUS
4	0.009019	Chip Status_NO CHIP
5	0.009451	Intake Condition_CONTAGIOUS
6	0.032875	Animal Origin_OVER THE COUNTER
7	0.034650	Chip Status_UNABLE TO SCAN
8	0.046054	Animal Origin_FIELD
9	0.068849	Intake Condition_REHABILITABLE NON-CONTAGIOUS
10	0.099089	Chip Status_CHIP
11	0.207397	Intake Condition_UNTREATABLE NON-CONTAGIOUS
12	0.220503	Intake Condition_HEALTHY
13	0.259504	Length of Stay(days)

Dog Modeling

Distribution by Classifier

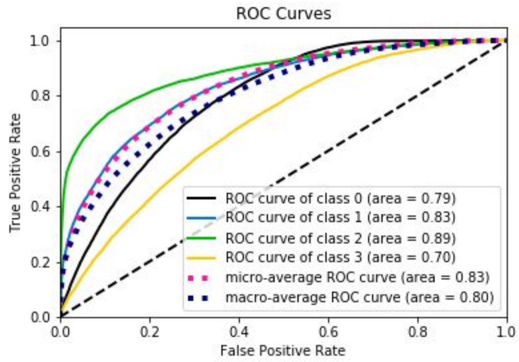



Animal Breed_TERRIER
0
20000
40000
60000
80000
100000
120000
0.0
0.2
0.4
0.6
0.8
1.0
Animal Origin_FIELD
0
10000
20000
30000
40000
50000
60000
70000
80000
0.0
0.2
0.4
0.6
0.8
1.0
Animal Origin_OVER THE COUNTER
0
10000
20000
30000
40000
50000
60000
70000
80000
0.0
0.2
0.4
0.6
0.8
1.0
Chip Status_CHIP
0
20000
40000
60000
80000
100000
0.0
0.2
0.4
0.6
0.8
1.0
Chip Status_NO CHIP
0
20000
40000
60000
80000
100000
0.0
0.2
0.4
0.6
0.8
1.0
Chip Status_UNABLE TO SCAN
0
20000
40000
60000
80000
100000
120000
0.0
0.2
0.4
0.6
0.8
1.0
Intake Condition_CONTAGIOUS
0
20000
40000
60000
80000
100000
120000
140000
0.0
0.2
0.4
0.6
0.8
1.0
Intake Condition_HEALTHY
0
20000
40000
60000
80000
100000
120000
0.0
0.2
0.4
0.6
0.8
1.0

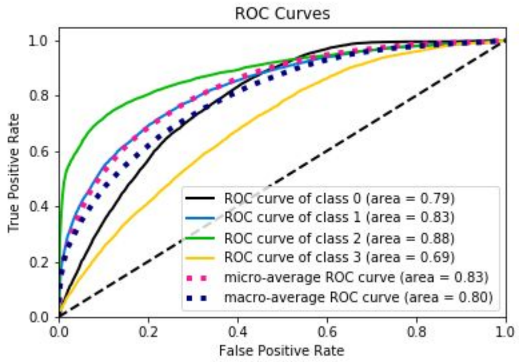

Intake Condition_MANAGEABLE NON-CONTAGIOUS
0
20000
40000
60000
80000
100000
120000
0.0
0.2
0.4
0.6
0.8
1.0
Intake Condition_REHABILITABLE NON-CONTAGIOUS
0
20000
40000
60000
80000
0.0
0.2
0.4
0.6
0.8
1.0
Intake Condition_UNTREATABLE NON-CONTAGIOUS
0
20000
40000
60000
80000
100000
120000
0.0
0.2
0.4
0.6
0.8
1.0
Length of Stay(days)
0
5000
10000
15000
20000
25000
30000
35000
0
2
4
6
8
Outcome Type
0
10000
20000
30000
40000
50000
0.0
0.5
1.0
1.5
2.0
2.5
3.0

ROC Curves by Model

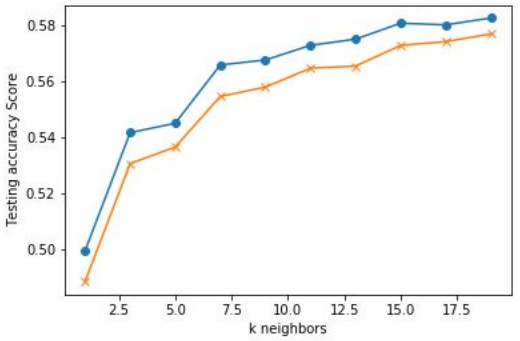
Production Model



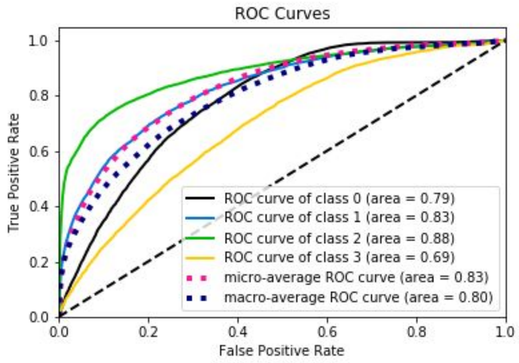
Bagging Model



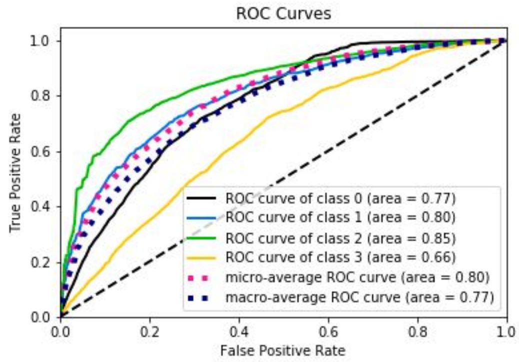
K-Neighbors Elbow Curve



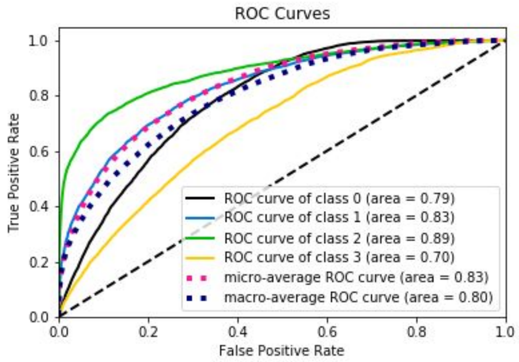
Decision Tree Model



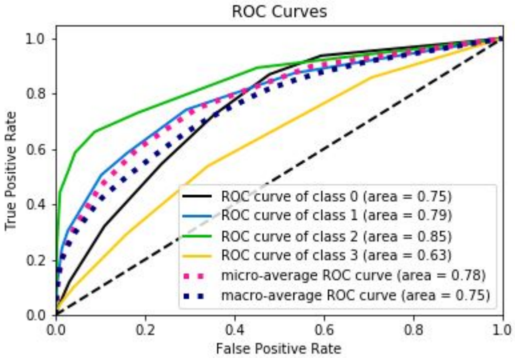
Ada Boost Model



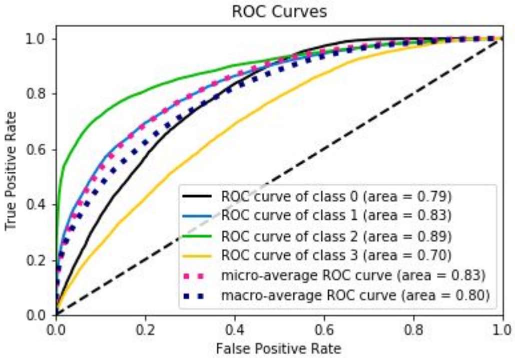
Gradient Boost Model



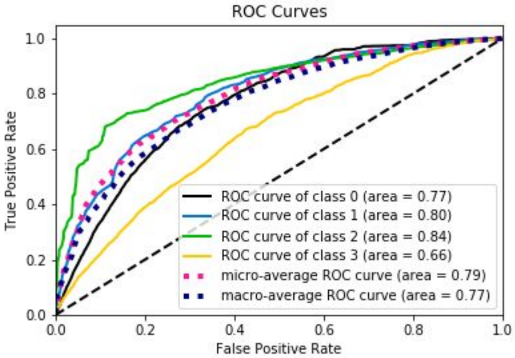
K-Neighbors Model



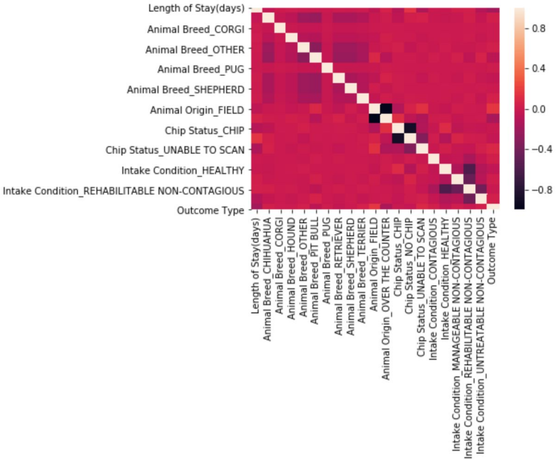
eXtreme Gradient Boost Model



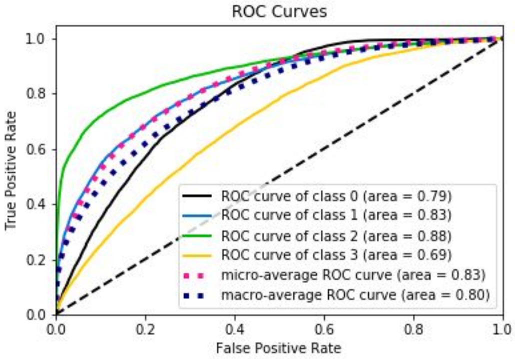
Logistic Regression Model



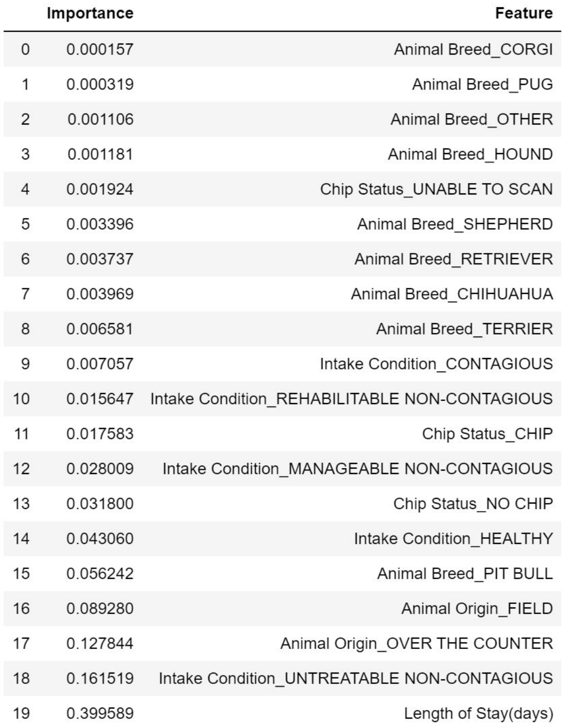
Correlation Heatmap



Random Forest Model



Feature Importance



	Importance	Feature
0	0.000157	Animal Breed_CORGI
1	0.000319	Animal Breed_PUG
2	0.001106	Animal Breed_OTHER
3	0.001181	Animal Breed_HOUND
4	0.001924	Chip Status_UNABLE TO SCAN
5	0.003396	Animal Breed_SHEPHERD
6	0.003737	Animal Breed_RETRIEVER
7	0.003969	Animal Breed_CHIHUAHUA
8	0.006581	Animal Breed_TERRIER
9	0.007057	Intake Condition_CONTAGIOUS
10	0.015647	Intake Condition_REHABILITABLE NON-CONTAGIOUS
11	0.017583	Chip Status_CHIP
12	0.028009	Intake Condition_MANAGEABLE NON-CONTAGIOUS
13	0.031800	Chip Status_NO CHIP
14	0.043060	Intake Condition_HEALTHY
15	0.056242	Animal Breed_PIT BULL
16	0.089280	Animal Origin_FIELD
17	0.127844	Animal Origin_OVER THE COUNTER
18	0.161519	Intake Condition_UNTREATABLE NON-CONTAGIOUS
19	0.399589	Length of Stay(days)