# Analyzing and Predicting Wildlife Incidents

Loyola University of Maryland
Carlin Soler
May 4th, 2021

## Abstract

Despite possessing a rich wildlife diversity and an increased number of agencies whose objective is to preserve and educate the community on the human impact on wildlife, there are limited studies that have evaluated and analyzed trends in urban wildlife in the U.S. There is also scarce information regarding the impact of the community on these animals. According to biologists at the DC The Department of Energy and Environment, birds that were thought to be extinguished are appearing because of the decrease in human activity caused by COV-19. Regardless of a decrease in human activity, wildlife rehabilitation agencies are working to their maximum capacity since the pandemic started. The purpose of this project is that, at a high level, we evaluate trends and predict wildlife incidents in the U.S.

My client, the Wildlife Center of VA, is one of the top wildlife rehabilitation centers in the DMV area. They are also one of the few rehabilitation centers to provide care to eagles in the U.S. Their work is not only limited to providing hospital care to wildlife, but also to educate the world to care for wildlife and the environment. They offer veterinary training, rehabilitation training, and education outreach training. The Wildlife Center of VA also has the Wildlife Care Academy, which helps others take an active and educated role in the conservation of wildlife. The center created a platform called WILD — ONe to assist wildlife other rehabilitation centers with managing and utilizing their records.

In the first phase of this report, we will find a thorough analysis of multiple wildlife centers in the United States using the WILD ONe dataset. It also contains a comparison in trends of the multiple centers to the Wildlife Center of VA. In the second phase of the report, we will build three classification models to predict which factors have the highest impact on predicting animal outcome.

The exploratory data analysis phase led to many relevant findings. First, the number of incidents increased almost every year except for 2019, in which the number of incidents decreased. May, June and July are the busiest months overall. Additionally, the top causes for intake were trauma, orphaning and habitat loss. While analyzing where the animals were coming from, we discovered that the busiest states were of Florida, Ohio, Wisconsin, and Virginia. We tested the gender hypothesis, and found gender did not improve the outcome.

The three classification algorithms used were random forest, XGBoost, and logistic regression. We implemented the algorithm on the 3 classes (birds, mammals and reptiles). Random forest and XGBoost had the best performance, and this can be attributed to the nature of both models. Random forest mammals dataset (accuracy score= 0.91, F1 score = 0.92 and AUC score = 0.97) had the best performance overall, followed by the random forest birds dataset (accuracy score = 0.90, F1 score = 0.90 and AUC = 0.96).

In conclusion, the models were able to successfully predict the probability of an animal being euthanized or released. We were also able to find significant trends in the animal intake and disposition. The exploratory data analysis and predictive models can still be improved. A proper data documentation and a more efficient feature engineering process can help increase the accuracy of the predictive models. Although models should never be used as a decision tool, they could be used as a resource to identify the factors that have the highest impact on predicting an animal being released.

# Table of Contents

## List of Figures

## List of Tables

# 1. Introduction

## 1.1 Project Background

According to Harvard University, approximately 30,000 species per year, about three per hour, are going extinct. The population of wildlife throughout the world decreased in size by approximately 52 percent between 1970 and 2010, and approximately 80 percent of the decline is caused by habitat destruction, injured by people and the urban environment. The fast urban development has greatly impacted wildlife habitat, and wild animals have been forced to adapt to living in proximity with people. These animals are instrumental in keeping the forests healthy, the city green, and the streets clean. Wildlife rescue and rehabilitation agencies survive out of the support they receive from government institutions and donations, and constantly face the challenges of making the most efficient use of their limited resources to protect these animals. The present report was prepared for the Wildlife Center of Virginia using their WILD — ONe database. This platform was created to facilitate wildlife rehabilitation centers with a tool that helps them collect and manage their patient data. It was also designed to assist researchers and wildlife health monitoring professionals with obtaining incident data on injured and orphaned wildlife. WILD — One is currently being used by 130 rehabilitation centers across 36 different states.

## 1.2 Objectives

Hundreds of animal incidents are entered into the system using WILD-ONe every day. The rehabilitation centers use this data to understand trends, such as the number of animals that are taken in each month, the reasons why these animals are being brought in, etc. However, the Wildlife Center of VA is interested in a detailed data analysis that covers all the centers that are using WILD — ONe compared to their center in VA. The analysis is intended to examine more carefully into their data to find relevance trends that help understand the relationships between the variables. Thus, the end goal of this project is to provide an extensive analysis of WILD- ONe data in order to help them gain a better understanding on the factors that have the highest impact on predicting animal outcome. This project will be divided into two sections or two goals.

Descriptive Goals:

- Identify trends in the categorization of injury/reasons for admissions over the years
- Identify the states in the U.S. that received the higher number of patients and the outcome
- Identify trends in the reasons for admissions by state
- Determine if there is a relationship between the time length animals spend at the center and its outcome.
- Identify trends in the hunting season and the busiest dates for animal intake

Predictive Goals:

- Determine if we can successfully predict animal outcome
- Identify the top factors that can help predict animal outcome

Due to the high number of distinct animals, for the descriptive goals, we will focus on the top three classes most centers bring for admission (birds, mammals and reptiles). For the prediction goals, we will focus on the top twenty common species for the three classes. The top twenty common species represent sixty percent of the total bird data, eighty percent of the total mammal data, and ninety percent of the total reptile data.

The analysis included in this report might help provide an insight about the animals that have the higher probability of surviving and the ones that have the higher probability of getting euthanized. It will also help determine the reasons for admission and injury types for patients that have the highest chance of receiving euthanasia and gain insight into causes of species decline. In addition, the results can help educate the community about the urban impact on these animals, raise awareness, help lower the rate of animal intakes and reduce the cost related to length of stay. Lastly, it could also motivate the centers that use WILD-ONe and other wildlife centers to collect more data and understand the importance of properly documenting the incidents.

# 2. Methods

## 2.1 Research Platform

All the analysis was conducted using the python programming language in a Jupyter notebook environment. Jupyter notebook is a flexible open-source software that allows the manipulation of code, equations, visualizations, and narrative text.
Jupyter notebook was used for data cleaning and transformation, data visualization, and machine learning. To perform the analysis, we used multiple libraries such as Scikit-learn, Pandas, and NumPy. Scikit-learn library is a machine learning open-source that supports various classification, regression, and clustering algorithms as well as data processing. Pandas was used to transform the Excel file into a pandas data frame. Since the client will be using the results for educational purposes, the visualization part of the project is very important. Most of the graphs included in this project were created using Tableau. This powerful interactive visualization software is a great tool to produce aesthetically pleasing graphs.

## 2.2 Dataset

The wildlife rehabilitation centers that use WILD — ONe gather and record the data of all of its patients upon entry and their outcomes. This includes information on birds, mammals, reptiles, amphibians, chondrichthyes, insects and malacostracans. The dataset provided contains a total of 567,549 admissions and 52 variables. Out of those, 566,739 were unique patients, so some animals came more than one time for rehabilitation. Some of these variables are common species name, class, date of admission, the reason for admission, anatomical site of the injury, disposition, and other fields. The data covers from July 18th, 2010 through December 31st, 2019 and contains roughly 1,282 unique species of animals. Between 2010-2018, the breakdown of the number of individual wildlife cases were 291,934 (51.5%) birds, 252,353 (44.4%) mammals, 22,545 (3.9%) reptiles, 796 (0.14%) amphibians, 7 chondrichthyes, and 5 insects. The dataset contains very few missing data, and the source of the data is reliable since it was provided by the WILD-ONe team at the Wildlife Center of VA.

# 3. Related Word

### 3.1 Species, causes, and outcomes of wildlife rehabilitation in New York State by Melissa Hanson, Nicholas Hollingshead, Krysten Schuler, William F. Siemer, Patrick Martin, and Elizabeth M. Bunting

Although I was unable to identify any research paper whose work focuses on using classification models to predict the probability of wildlife patients outcomes. Species, causes, and outcomes of wildlife rehabilitation in New York State paper explore factors that have the highest probability of predicting animal outcome.

They explored multiple factors, such as common trauma, distress category, final disposition, and others. The dataset included 59,370 patients comprising 31,229 (52.6%) birds, 25,490 (42.9%) mammals, 2,423 (4.1%) reptiles, and 73 (0.1%) amphibians. In the report found, it is not clear as to which tools they use in order to perform the analysis.
The data processing aspect of the report was similar to the one performed on this analysis. Simple data entry errors by the rehabilitation centers (such as minor incorrect spelling of species or place names) were corrected and the top admitted species per class were used to conduct the analysis. The reasons for admissions, category of the injury, and final disposition were also fairly similar to the dataset provided by the WCV.

# 4. Data Analysis

## 4.1 Data processing

### 4.1.1 Data Cleaning

The data cleaning process involved different steps, such as removal of irrelevant data, outliers, and duplicates, and fixing syntax errors. The first step in data processing involved the removal of unnecessary columns. These variables didn't add any value to the analysis, so they were dropped. They include: Organization Address, Organization City, Organization Jurisdiction, Organization Country, Organization Phone, Organization Email, Organization Elevation, Case Number, Subspecies, Species Code, Primary, Suspected/Observed, Admission Comments, Rescue Address, Rescue State, Rescue Jurisdiction, Rank, Suspected or Confirmed, Date Dispositioned, Disposition Address, Disposition State, Disposition Jurisdiction, Disposition Latitude, Disposition Longitude, and Disposition Elevation.

After taking a closer inspection of the variables left, we found that the field “Duration of Care (in days)” had some negative numbers and numbers that exceeded the length of the years of the dataset, for instance, some rows had -41,685 days or 365,243 days. It is important to note that outliers shouldn’t be removed unless there is a good reason for it. As previously mentioned, in this scenario they were removed because we cannot have a negative digit for days; 365,243 was removed because if divided by 365 is more than 1000 years. We can tell that these numbers are incorrect and can affect the results of the machine learning model.

The dataset was already for the most part clean, so only a few fields needed further processing in order to be consistent. For instance, the field class and common species name had some duplicates with syntax errors. In order to fix it, these variables were combined into one.

### 4.1.2 Feature Engineering

Feature engineering is the process of creating new features from original data. To gain a better understanding of the busiest days and months of the year, we added the month and day features by breaking down the Date admitted variable.

The causes of distress included a total of 124 different unique strings. For analytical purposes, these categories were aggregated into 7 groupings: 1) Orphaned (orphaned, orphaned due to human intervention, nutritional, metabolic, developmental or congenital abnormality); 2) Trauma (collision, entrapment, entanglement, injured by another animal, human o machinery, projectile);  3) Infectious (parasite, bacterial infection, fungal, viral disease); 4)Toxicity (poison or toxin ingestion, soaked or similar damage); 5) Habitat Loss (habitat loss due to human disturbance, environmental disturbance); 6) Confiscation (legal or illegal possession, born in captivity); 7) Unknown (disease or illness whose identity is unknown or lack information for proper classification).

The disposition field included a total of 11 categories: 1) Active (still under care); 2) Transferred (transferred to another rehabilitation center); 3) Self-release; 4) Released ; 5) Died (died prior to or under care); 6) Euthanized; 7) Surrogate; 8) Placed; 9) Adoption; 10) Permanent Resident; 11) Education Animal. Again, for analytical purposes, these categories were aggregated into 4 groups: 1) Active (animal still under care); 2) Died; 3) Euthanized; 4) Released (transferred, adoption, placed, surrogate, etc.). Other features were added to facilitate the visualization of the data. The features added are presented in the table below.

In addition to the date, causes of distress and disposition, the age field, which contained 12 categories: Adult, Infant, Juvenile, Hatchling, Fledgling, Undetermined, Pouch Young, Egg, Indeterminate, Neonate, Development of Back Legs, and Hatchling/Tadpole. These categories were aggregated into 3 groups: egg to infant, juvenile and adult.

## 4.2 Hypothesis and Assumptions

As previously mentioned, the final goal of this project is to provide an in-depth analysis of WILD-One data, to identify which fields have the most power in predicting outcome, and to implement a machine learning technique that can later be used to predict the animal's final disposition.

Global declines in wildlife numbers raise questions regarding specific causes and trends in the decline of wildlife. Many factors can be taken into consideration when trying to predict animal outcome, some of which are hunting season, anthropogenic activity, diseases, and exploitation.

The majority of the species in the dataset fall into 3 different classes: birds, mammals, and reptiles. Birds being the number one admitted class followed by mammals. Separate models were created for the top 3 classes (birds, mammals, and reptiles). Since the number of unique species exceeded 1,282, the top 20 species for each class were used for building the machine learning models. The intake reasons were desegregated from confiscation, orphaning, trauma, orphaned, habitat loss, infection toxicity, and unknown to their 124 individual causes of distress. Since the primary goal of this project was to understand and predict final disposition (died, euthanized, and released), the model will attend to predict which reasons for admission have the most impact on predicting the chances of an animal being released or euthanized.
In other words, the models built will assume two possible outcomes (e.g. binary problem). It is important to note that some rehabilitation centers euthanized some species upon entry, as a result, those species were removed from the analysis and models.

## 4.3 Data Exploration

The purpose of this section is to respond to the descriptive analytic goals and provide a better understanding of the relationship between the variables in the dataset.

Below you will find a table with the variables used for the exploratory data analysis portion of the project along with its description.

Table 1. Features used for exploration data analysis

| Variable | Description |
|---|---|
| **Anatomical Site of the Injury** | Description of the physical location of the injury if admitted due to an injury (e.g. Skin/Integumentary System, Thin, Eye/Ocular system). |
| **Categorization of Injury** | Category in which the injury falls into (e.g. nutritional, fleas, external). |
| **Intake Reasons** | The circumstances of Rescue category included 124 unique strings. These categories were aggregated into 7 groupings: Orphaned, trauma, Infectious, toxicity, habitat Loss, confiscation and unknown. |
| **Class** | Type of animal (e.g. birds, reptiles, mammals, etc) |
| **Common Species Name** | It is the name known to the general public (e.g, Mallard, Painted turtle) |
| **Month Admitted** | The date of admission feature was disaggregated into month, year and date. Month in which the patient was admitted for service. |
| **Year Admitted** | Year in which the patient was admitted for service. |
| **Disposition Group** | The disposition category contained 11 different outcomes. Those outcomes were aggregated into 4 different categories( e.g. active, died, euthanized, released) |
| **Gender** | The gender of the species (e.g. male, female, indeterminate) |
| **Life Stage Group** | The life stage category was aggregated into 3 groups to facilitate analysis (e.g. egg-infant, juvenile and adult). |
| **Circumstances of rescue** | This category contains information on how the animal was injured (e.g. collision with moving object car/truck/ motorcycle). |
| **Organization State** | The abbreviation for the rehabilitators state. |
| **Organization Name** | The name of the rehabilitation center. |
| **Days in Care** | The number of days the animal spent in care. |

## 4.3.1 Admissions

### 4.3.1.1 Admissions per year by species class

The figure below represents the number of animals admitted each year broken down by the top 3 species classes. It seems that there has been a slight decline in the number of intakes in 2019. In addition, in 2019, the number of mammals admitted surpassed the number of birds for the first time.

**Figure 1. Admissions per year by species class**

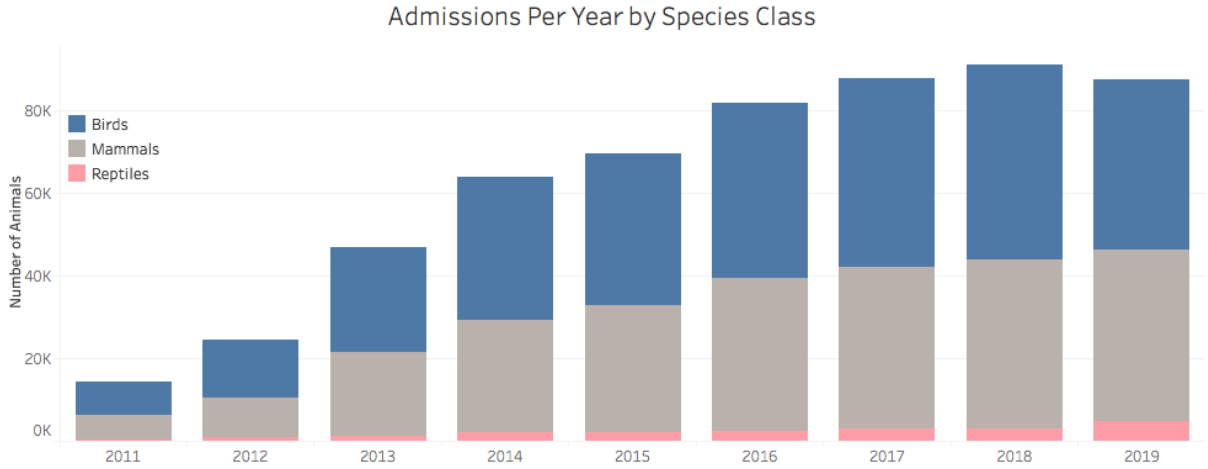


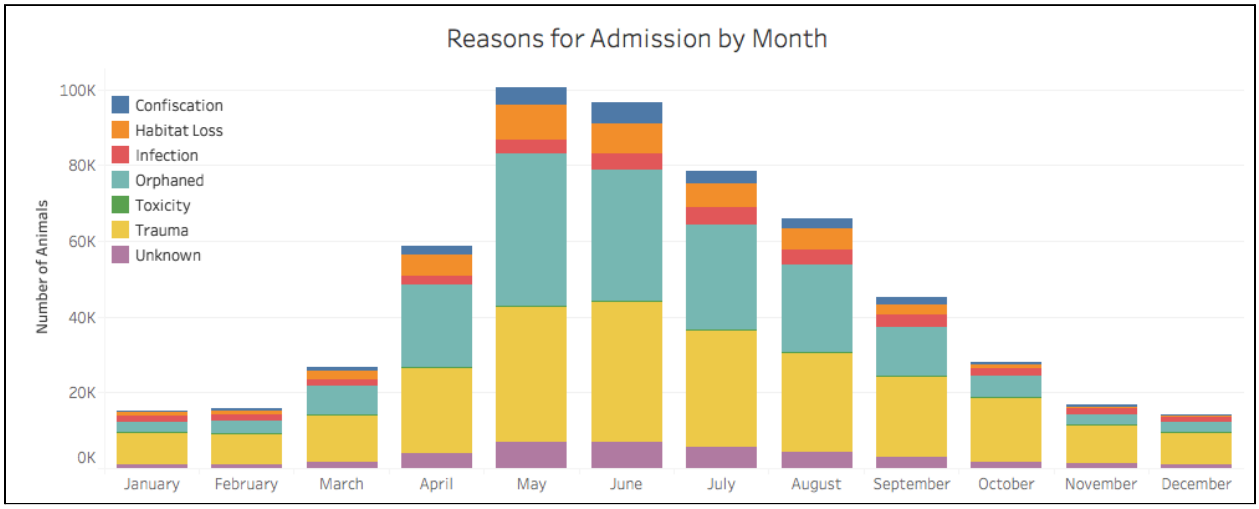

**Table 2. Admissions per year by species class**

| | Birds | | Mammals | | Reptiles | | |
|---|---|---|---|---|---|---|---|
| **Year** | **Cases** | **% Total** | **Cases** | **% Total** | **Cases** | **% Total** | **Total Admissions** |
| **2011** | 7,644 | 53% | 6,113 | 42.7% | 558 | 3.9% | 14,324 |
| **2012** | 13,704 | 55.8% | 9,759 | 39.8% | 1,580 | 6.4% | 24,545 |
| **2013** | 24,912 | 53.2% | 20,308 | 43.3% | 1,580 | 3.4% | 46,840 |
| **2014** | 34,380 | 53.8% | 27,176 | 42.5% | 2,242 | 3.5% | 63,870 |
| **2015** | 36,349 | 52% | 30,748 | 44.2% | 2,405 | 3.4% | 69,586 |
| **2016** | 41,946 | 51.2% | 36,956 | 45.2% | 2,784 | 3.4% | 81,803 |
| **2017** | 45,198 | 51.5% | 39,070 | 44.5% | 3,344 | 3.8% | 87,770 |
| **2018** | 46,817 | 48.2% | 40,710 | 44.7% | 3,424 | 3.7% | 91,082 |
| **2019** | 40,975 | 46.7% | 41,512 | 47.3% | 5,047 | 5.7% | 87,715 |

### 4.3.1.2 Reasons for admission by month

Below are the trends in admissions per month. Spring seems to be the busiest season for wildlife rehabilitators and one reason could be timing. Spring is an ideal time for babies to be born. Admissions tend to decline during the winter because fewer animals and people are outside, and it's not nursery season.

**Figure 2. Reasons for admission by month**




Trauma was the number one cause of admission overall (n=238,389, 42.43%) followed by orphaning (n= 184,668, 32.87%); habitat loss (n= 44,410, 7.90%); infectious disease (n= 31,110, 5.54%); confiscation (n= 23, 238, 4.1%): and intoxication (n= 1,528, 0.27%). The rest of the cases were categorized as unknown (38,491, 6.8%). As mentioned before, this could be due to: 1) a cause of distress was not provided, 2) the cause of distress was unknown; or 3) lack enough information to properly assign a category.

### 4.3.1.3 Analyzing where the animals are coming from

Birds are one of the largest population rehabilitation centers reported for admission followed by mammals and reptiles. Let's recall that 291,934 birds, 252,353 mammals, and 22,545 reptiles were admitted by the wildlife centers in this dataset. The figure below portrays where most of the birds, mammals, and reptiles are coming from. The states of Florida, Ohio, Wisconsin, and Virginia had the highest rate of admission for birds and mammals. On the other hand, Florida, Virginia, Wisconsin, and New Jersey had the highest intake rate for reptiles.

Figure 3. Where the birds are coming from

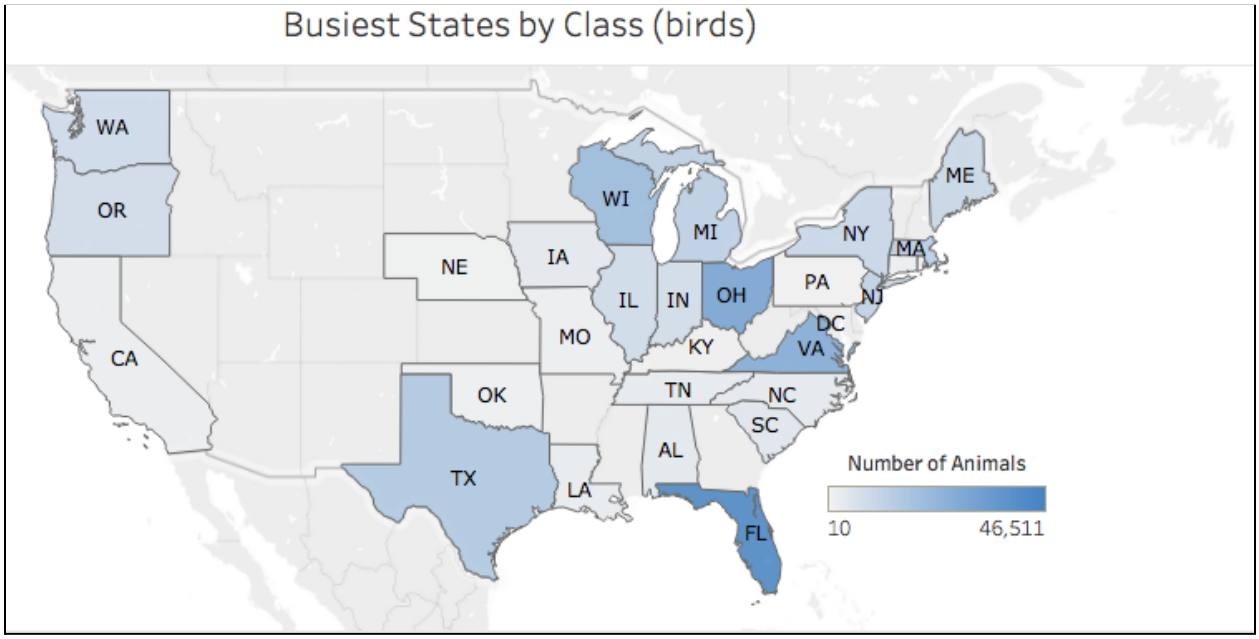



Figure 4. Where the mammals are coming from

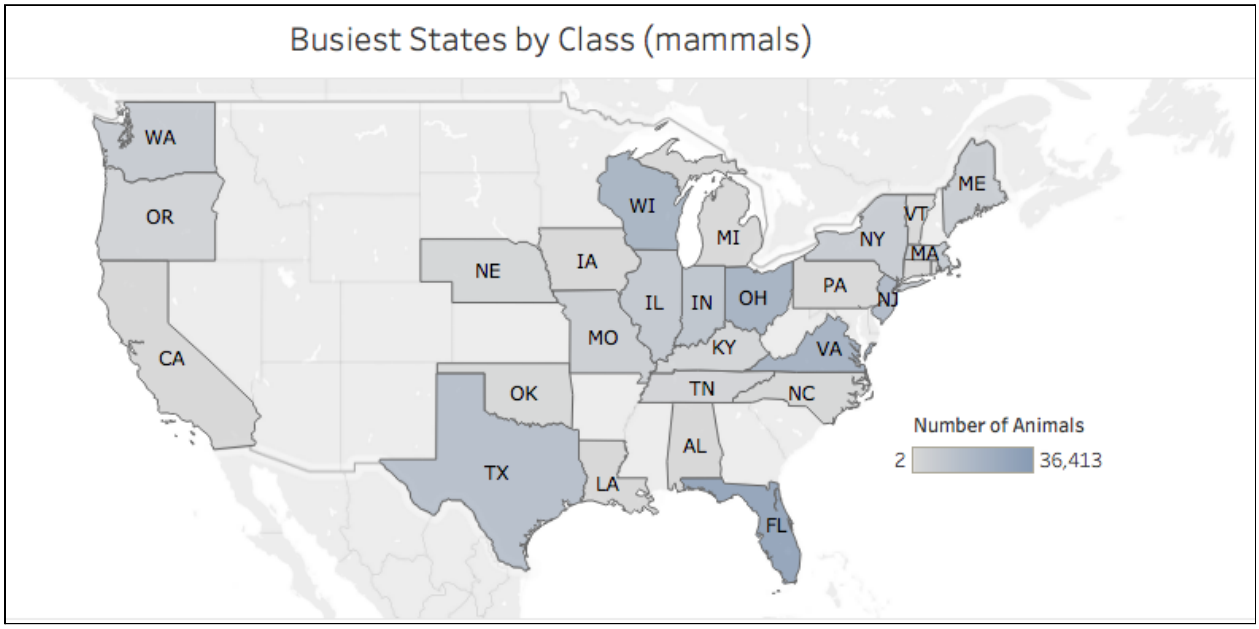
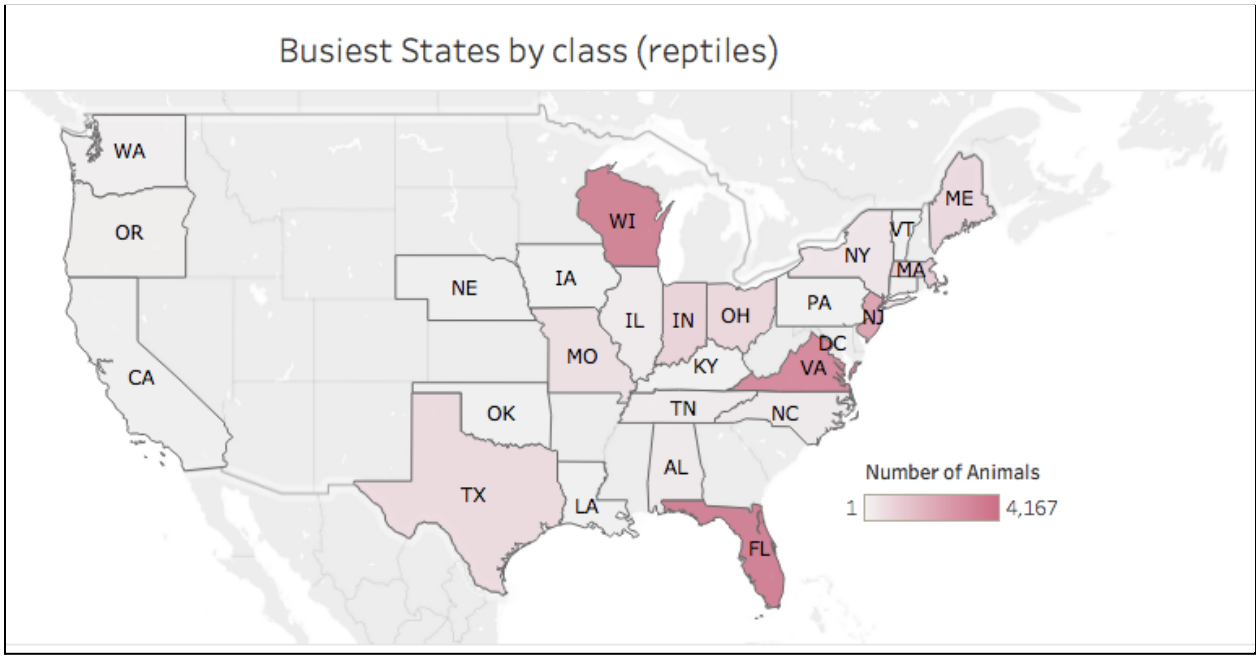


Figure 5. Where the reptiles are coming from




Florida had the highest rate of admissions overall for birds, mammals, and reptiles with a total of 46,511 (17.41%) birds, 36,413 (15.82%mammals and 4,167 (19.66%) reptiles. The state of Ohio admitted a total of 32,870 (12.30%) birds, 25,697 (11.16%) mammals, and 801 (3.78%) reptiles. A total of 29,544 (11.06%) birds, 26, 278 (11.42%) mammals, and 3,719 (17.55%) reptiles were admitted by the state of Virginia.

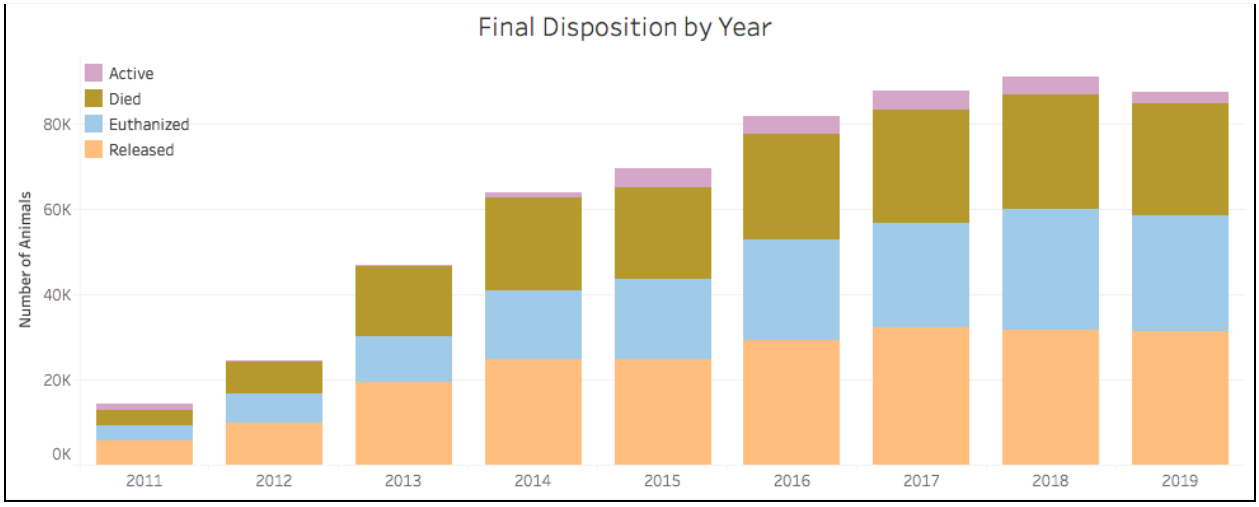
The state of Wisconsin admitted a total of 22,980 (8.60%) birds, 24,767(10.76%) mammals and 4,035 (19.05%) reptiles. The state of Wisconsin had the second largest number of admissions for reptiles.

## 4.3.2 Disposition

### 4.3.2.1 Final disposition by year

Final dispositions were reported for 99.9% (n=566,739) of the cases. Approximately 36.89% (n= 209,352) of the animals were released to the wild. Rehabilitators reported that 174,641 (30.82%) died naturally before or during care. Euthanizations accounted for 28.18% (n= 159,707) of the disposition cases. The overall released rate decreased slightly 4.08% in 2013 and 8.95% in 2015 in proportion to the overall admission increase rate.

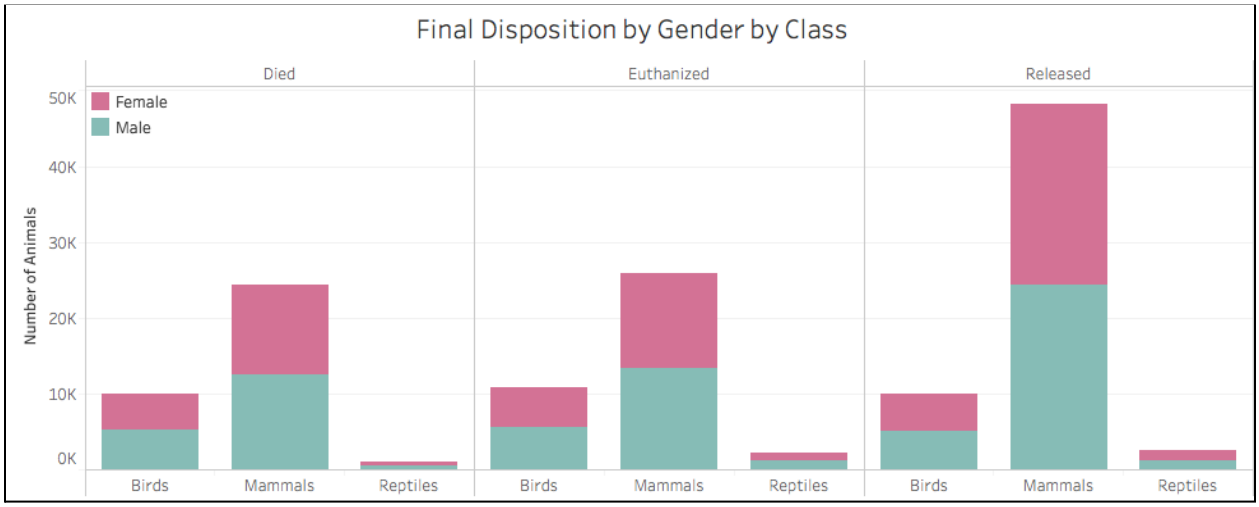
**Figure 6. Final disposition by year**




### 4.3.2.2 Final disposition by gender

Gender was identified for 25% of the patients brought for treatment. Overall, males accounted for 12.71% of the cases and females for 12.14% of the cases. There is not a significant impact of gender on the outcome. Both groups tend to die, be euthanized, and be released at a very similar rate.

**Figure 7. Final disposition by gender and class**



### 4.3.2.3 Reasons for euthanizing Birds

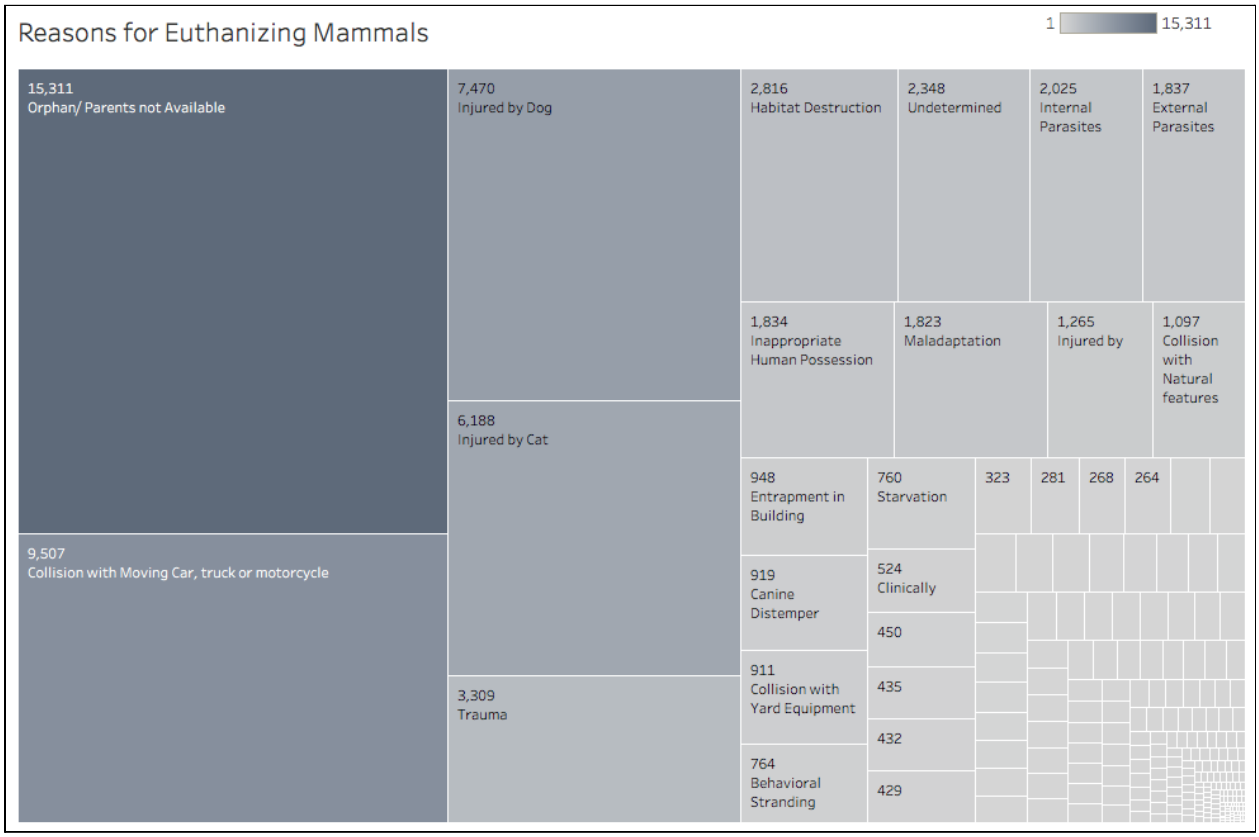
A total of 82,567 birds were euthanized due to multiple causes. Collision with a moving object was the number one reason for euthanizing birds accounting for 17.76% of the euthanized birds followed by a trauma different from collision and animal interaction (9.22%). Cat incidents accounted for 7.44% of the euthanizations. Due to the great interest from the WCV in understanding the impacts of domestic cats on declining birds, trauma from cats has been disaggregated and discussed further in the report. Most of the reasons for euthanizing birds are related to a collision, animal interaction, and infections.

**Figure 8. Reasons for euthanizing birds**


### 4.3.2.4 Reasons for euthanizing mammals

A total of 70,962 were euthanized over 9 years. Orphaning was the number one reason for euthanizing mammals accounting for 21.57% of the euthanizations. The second reason was a collision with a moving car, truck, or motorcycle (13.39%) followed by injury by a dog and cat, which accounted for 10.52% and 8.72% respectively of the mammals euthanized. Similar to the reasons for euthanizing birds, orphaning, collision, injured by another animal and infection were the top reasons for euthanizing mammals.

**Figure 9. Reasons for euthanizing mammals**




### 4.3.2.5 Reasons for euthanizing reptiles

Of the admitted reptiles, a total of 6,178 were euthanized. Similar to birds, the number one reason for euthanizing reptiles was a collision with a car, truck, or motorcycle, which accounted for 63.38% of the cases. The second reason was trauma other than collision and animal interaction, which accounted for 4.62% of the cases. Overall, collision, trauma, or animal interaction were the top reasons for euthanizing reptiles.

Figure 10. Reasons for euthanizing reptiles


### 4.3.2.6 Analyzing time spent at rehabilitation center by disposition

Most of the euthanizations occurred within a day or the same day of the admission. Reptiles spent the most time in care before being released followed by mammals and birds respectively. Reptiles spent the longest time in care before dying.

Figure 11.Distribution of time in shelter by age disposition

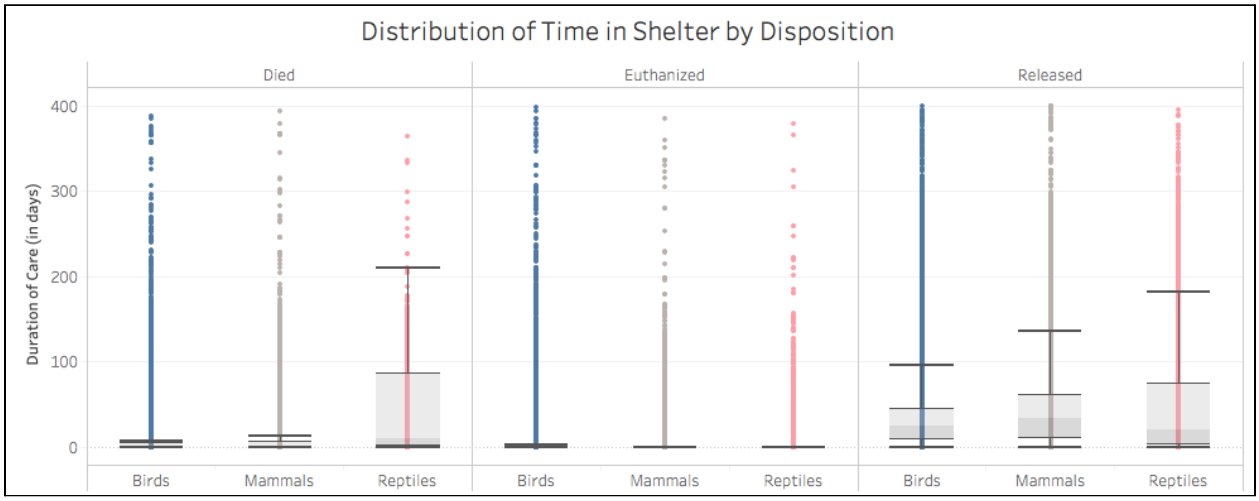


Table 3. Distribution of time in shelter by disposition (birds data)

| Birds | | | |
|---|---|---|---|
| | **Died** | **Euthanized** | **Released** |
| **Maximum** | 1,485 days | 908 | 1,111 days |
| **75 Percentile** | 3 days | 1 day | 43 days |
| **50 Percentile** | 1 day | Less than a day | 25 days |
| **25 Percentile** | Less than a day | Less than a day | 8 days |
| **Minimum** | Less than a day | Less than a day | Less than a day |

Table 4. Distribution of time in shelter by disposition (mammals data)

| Mammals | | | |
|---|---|---|---|
| | **Died** | **Euthanized** | **Released** |
| **Maximum** | 1,179 days | 1,042 days | 806 days |
| **75 Percentile** | 5 days | Less than a day | 60 days |
| **50 Percentile** | 1 days | Less than a day | 32 days |
| **25 Percentile** | Less than a day | Less than a day | 9 days |
| **Minimum** | Less than a day | Less than a day | Less than a day |

Table 5 . Distribution of time in shelter by disposition (reptiles data)

| Reptiles | | | |
|---|---|---|---|
| | **Died** | **Euthanized** | **Released** |
| **Maximum** | 1,324 days | 1,178 days | 1,039 days |
| **75 Percentile** | 84 days | Less than a day | 72 days |
| **50 Percentile** | 8 days | Less than a day | 18 days |
| **25 Percentile** | 1 days | Less than a day | 2 days |
| **Minimum** | Less than a day | Less than a day | Less than a day |

#### 4.3.2.7 Analyzing time spent at rehabilitation center by age group

Adult birds, mammals and reptiles spend the least time during rehabilitation followed by egg to infant group and juvenile respectively. The egg to infant group of reptiles spend the longest time in care.

Figure 12. Distribution of time in shelter by age group

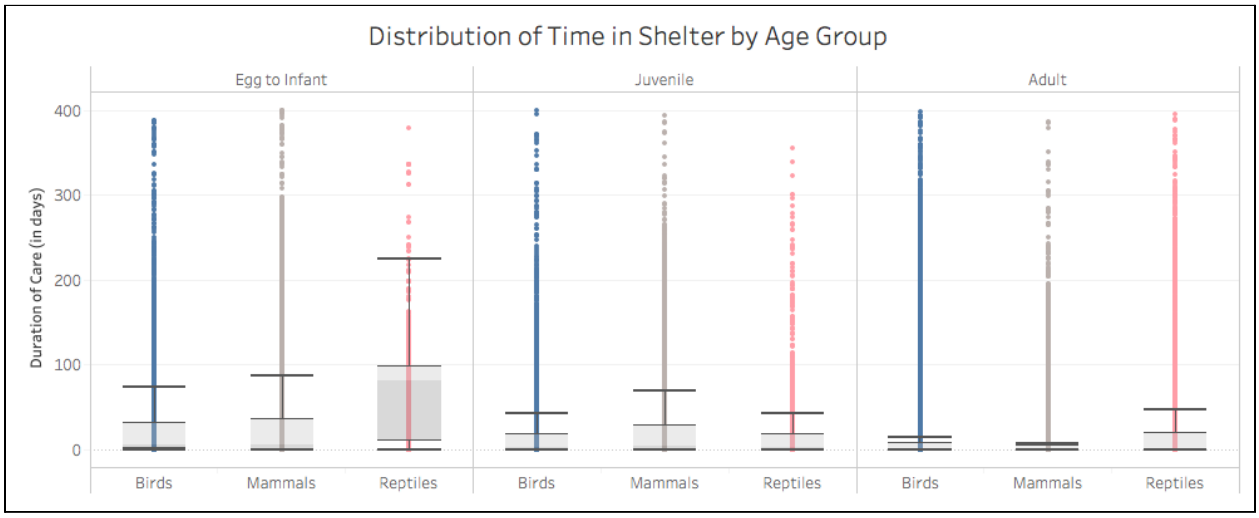



Table 6. Distribution of time in shelter by age group (birds data)

| | Birds | | |
|---|---|---|---|
| | **Egg-Infant** | **Juvenile** | **Adult** |
| **Maximum** | 714 days | 632 days | 1,485 days |
| **75 Percentile** | 30 days | 17 days | 6 days |
| **50 Percentile** | 5 days | 2 days | 1 days |
| **25 Percentile** | 1 day | Less than a day | Less than a day |
| **Minimum** | Less than a day | Less than a day | Less than a day |

Table 7. Distribution of time in shelter by age group (mammals data)

| | Mammals | | |
|---|---|---|---|
| | **Egg-Infant** | **Juvenile** | **Adult** |
| **Maximum** | 1,179 days | 655 days | 1,132 days |
| **75 Percentile** | 35 days | 28 days | 3 days |
| **50 Percentile** | 5 days | 3 days | Less than a day |
| **25 Percentile** | Less than a day | Less than a day | Less than a day |
| **Minimum** | Less than a day | Less than a day | Less than a day |

Table 8. Distribution of time in shelter by age group (reptiles data)

| | Reptiles | | |
|---|---|---|---|
| | Egg-Infant | Juvenile | Adult |
| Maximum | 435 days | 1,010 days | 1,324 days |
| 75 Percentile | 97 days | 17 days | 19 days |
| 50 Percentile | 81 days | 2 days | 1 day |
| 25 Percentile | 10 days | Less than a day | Less than a day |
| Minimum | Less than a day | Less than a day | Less than a day |

### 4.3.2.8 Causes of admission and corresponding released rate for top 20 admitted birds

Birds had the highest overall rate of release (51.9%) when compared with mammals and reptiles. Of the 138,425 birds (top 20 species) that were admitted, 71,842 were eventually released. Of these successfully released birds, 39,980 had arrived at the facilities due to orphaning, while 62,443 had arrived due to trauma. The mallard duck had the highest overall rehabilitation rate of all the birds evaluated with 77% being released (16,594 admitted; 12,777 released). For comparison purposes, 572 (16%) of the 3,581 Cooper's Hawk were successfully released. Cooper's Hawk represented 2.6% of the birds admitted for care and possessed the lowest release rate.

Table 9. Intake reasons and corresponding released rate for top 20 admitted birds

| | Confiscation | | Habitat Loss | | Infection | | Orphaned | | Toxicity | | Trauma | | Unknown | | Total | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Cases | % Rel. | Cases | % Rel. | Cases | % Rel. | Cases | % Rel. | Cases | % Rel. | Cases | % Rel. | Cases | % Rel. | Cases | % Rel. |
| American Robin | 967 | 59% | 2,459 | 49% | 1,261 | 22% | 5,885 | 41% | 28 | 3% | 9,633 | 22% | 1,325 | 47% | 21,558 | 47% |
| Mallard | 484 | 75% | 236 | 64% | 538 | 26% | 10,194 | 76% | 31 | 35% | 3,871 | 51% | 1,240 | 77% | 16,594 | 77% |
| Mourning Dove | 715 | 64% | 1,096 | 54% | 1,125 | 16% | 3,208 | 41% | 18 | 33% | 8,592 | 28% | 1,069 | 48% | 15,823 | 48% |
| Red-tailed Hawk | 67 | 43% | 95 | 57% | 1,354 | 20% | 1,622 | 31% | 76 | 36% | 4,672 | 29% | 448 | 52% | 8,334 | 52% |
| Canada Goose | 159 | 68% | 57 | 46% | 784 | 21% | 2,392 | 61% | 67 | 16% | 4,376 | 28% | 450 | 67% | 7,285 | 67% |
| Blue Ray | 470 | 47% | 469 | 44% | 292 | 13% | 1,453 | 33% | 6 | 33% | 2,851 | 22% | 689 | 37% | 6,230 | 37% |
| Northern Cardinal | 186 | 41% | 324 | 37% | 397 | 12% | 978 | 27% | 1 | 0% | 3,821 | 21% | 473 | 28% | 6,180 | 29% |
| Common Grackle | 345 | 53% | 759 | 50% | 295 | 11% | 1,793 | 29% | 12 | 17% | 2,410 | 17% | 447 | 35% | 6,061 | 35% |
| House Finch | 265 | 52% | 833 | 45% | 540 | 28% | 1,995 | 34% | 9 | 0% | 1,832 | 20% | 465 | 42% | 5,939 | 42% |
| Eastern Screech-Owl | 92 | 81% | 410 | 71% | 359 | 31% | 838 | 68% | 16 | 50% | 3,229 | 38% | 443 | 65% | 5,387 | 65% |
| American Crow | 151 | 63% | 120 | 56% | 1,054 | 7% | 1,383 | 23% | 33 | 18% | 1,896 | 16% | 524 | 39% | 5,161 | 39% |
| Great Horned Owl | 68 | 66% | 162 | 77% | 714 | 21% | 1,178 | 39% | 36 | 24% | 2,501 | 29% | 334 | 59% | 4,993 | 59% |
| Northern Mockingbird | 544 | 60% | 517 | 50% | 177 | 14% | 1,468 | 39% | 1 | 100% | 1,659 | 34% | 556 | 45% | 4,922 | 45% |
| Barred Owl | 49 | 65% | 72 | 71% | 336 | 27% | 432 | 51% | 13 | 38% | 3,158 | 33% | 303 | 54% | 4,363 | 27% |
| Carolina Wren | 221 | 48% | 926 | 41% | 32 | 12% | 1,720 | 43% | 2 | 100% | 1,135 | 28% | 298 | 46% | 4,334 | 46% |
| Brown Pelican | 8 | 100% | 125 | 37% | 242 | 45% | 1,033 | 47% | 87 | 52% | 2,177 | 56% | 207 | 63% | 3,879 | 63% |
| Cooper's Hawk | 15 | 33% | 82 | 60% | 444 | 18% | 568 | 33% | 13 | 23% | 2,190 | 31% | 269 | 49% | 3,581 | 16% |
| Red-shouldered Hawk | 38 | 72% | 59 | 72% | 284 | 22% | 563 | 55% | 35 | 49% | 1,507 | 30% | 307 | 59% | 2,793 | 59% |
| Common Pigeon | 65 | 53% | 104 | 60% | 361 | 24% | 719 | 52% | 15 | 20% | 873 | 33% | 387 | 57% | 2,524 | 57% |
| Rock Pigeon | 116 | 67% | 123 | 58% | 416 | 24% | 558 | 38% | 9 | 67% | 1,060 | 29% | 202 | 27% | 2,484 | 57% |
| Total | 5,025 | 59% | 9,028 | 50.3% | 11,005 | 18.7 | 39,980 | 49.6% | 508 | 31.7% | 62,443 | 29.2% | 10,436 | 51.8% | 138,425 | 51.9% |

### 4.3.2.9 Causes of admission and corresponding released rate for top 20 admitted mammals

Mammals had the second-highest overall rate of release (47%) when compared with reptiles. Of the 235,217 mammals (top 20 species) that were admitted, 110,551 were eventually released. Of these successfully released mammals, 96,272 had arrived at the facilities due to orphaning, while 86,379 had arrived due to trauma. The coyote had the highest overall rehabilitation rate of all the mammals evaluated with 71% being released (723 admitted; 556 released). For comparison purposes, 910 (25%) of the 3,643 White-footed Mouse were successfully released. White-footed Mouse represented 1.5% of the mammals admitted for care and possessed the lowest release rate.

**Table 10. Intake reasons and corresponding released rate for top 20 admitted mammals**

| | Confiscation | | Habitat Loss | | Infection | | Orphaned | | Toxicity | | Trauma | | Unknown | | Total | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Cases | % Rel. | Cases | % Rel. | Cases | % Rel. | Cases | % Rel. | Cases | % Rel. | Cases | % Rel. | Cases | % Rel. | Cases | % Rel. |
| **Big Brown Bat** | 114 | 56% | 950 | 59% | 416 | 20% | 1,416 | 32% | 22 | 36% | 3,660 | 36% | 865 | 51% | 7,443 | 32% |
| **Coyote** | 60 | 18% | 18 | 56% | 49 | 35% | 329 | 71% | 1 | 0% | 219 | 33% | 47 | 64% | 723 | 71% |
| **Deer Mouse** | 66 | 50% | 243 | 41% | 13 | 23% | 804 | 25% | 2 | 0% | 323 | 35% | 87 | 53% | 1,538 | 25% |
| **Eastern Chipmunk** | 46 | 48% | 111 | 56% | 279 | 12% | 858 | 38% | 38 | 5% | 2,233 | 26% | 197 | 29% | 3,762 | 38% |
| **Eastern Cottontail** | 4,350 | 15% | 6,912 | 45% | 2,287 | 6% | 16,918 | 37% | 50 | 4% | 38,316 | 23% | 2,895 | 46% | 71,728 | 37% |
| **Eastern Fox Squirrel** | 184 | 6% | 450 | 66% | 94 | 10% | 1,433 | 37% | 0 | 0% | 746 | 20% | 299 | 41% | 3,206 | 37% |
| **Eastern Gray Squirrel** | 2,664 | 16% | 8,470 | 64% | 2,040 | 16% | 25,439 | 52% | 91 | 14% | 15,012 | 27% | 2,983 | 57% | 56,699 | 52% |
| **Eastern Wood Rat** | 39 | 23% | 237 | 33% | 2 | 0% | 192 | 33% | 0 | - | 113 | 26% | 11 | 36% | 594 | 33% |
| **Fox Squirrel** | 27 | 11% | 119 | 61% | 48 | 15% | 410 | 46% | 1 | 0% | 341 | 23% | 78 | 65% | 1,024 | 46% |
| **Little Brown Bat** | 24 | 4% | 66 | 35% | 64 | 19% | 149 | 43% | 6 | 50% | 275 | 22% | 162 | 51% | 746 | 43% |
| **Norway Rat** | 50 | 62% | 83 | 6% | 16 | 0% | 524 | 28% | 4 | 0% | 159 | 26% | 104 | 42% | 940 | 28% |
| **Raccoon** | 947 | 24% | 887 | 57% | 1.418 | 3% | 12,779 | 46% | 34 | 26% | 4,100 | 33% | 1,805 | 41% | 21,970 | 46% |
| **Red Fox** | 93 | 28% | 13 | 85% | 264 | 39% | 757 | 58% | 7 | 0% | 720 | 40% | 130 | 66% | 1,984 | 58% |
| **Red Squirrel** | 73 | 14% | 341 | 64% | 22 | 41% | 1,886 | 57% | 1 | 100% | 491 | 38% | 77 | 83% | 2,891 | 57% |
| **Southern Flying Squirrel** | 70 | 13% | 256 | 55% | 36 | 19% | 441 | 43% | 1 | 0% | 818 | 46% | 86 | 57% | 1,708 | 43% |
| **Striped Skunk** | 132 | 16% | 56 | 86% | 259 | 4% | 2,683 | 63% | 9 | 22% | 877 | 34% | 293 | 47% | 4,309 | 63% |
| **Virginia Opossum** | 646 | 14% | 345 | 56% | 698 | 18% | 24,065 | 53% | 43 | 12% | 14,513 | 37% | 2,665 | 55% | 42,975 | 54% |
| **White-footed Mouse** | 187 | 27% | 515 | 22% | 24 | 25% | 2,133 | 25% | 5 | 0% | 643 | 31% | 136 | 33% | 3,643 | 25% |
| **White-tailed Deer** | 495 | 26% | 11 | 9% | 268 | 8% | 1,749 | 32% | 27 | 22% | 1,746 | 12% | 391 | 37% | 4,687 | 32% |
| **Woodchuck** | 48 | 50% | 36 | 39% | 216 | 9% | 1,107 | 44% | 5 | 0% | 1,074 | 24% | 161 | 55% | 2,647 | 44% |
| **Total** | **10,315** | **23%** | **20,119** | **55%** | **8,513** | **11%** | **96,072** | **47%** | **347** | **15%** | **86, 379** | **28%** | **13,472** | **50%** | **235,217** | **47%** |

### 4.2.4.3 Causes of admission and corresponding released rate for top 20 admitted reptiles

Reptiles were the lowest admitted number of animals and had the lowest release rate (40%) when compared with birds and mammals. Of the 16,165 reptiles (top 20 species) that were admitted, 7,666 were eventually released. Of these successfully released reptiles, 13, 016 had arrived at the facilities due to trauma, while 1,895 had arrived due to confiscation.
The Plains Garter Snake had the highest overall rehabilitation rate of all the reptiles evaluated with 96% of being released (413 admitted; 396 released).

For comparison purposes, 155 (19%) of the 820 Western Painted Turtle were successfully released. Western Painted Turtle represented 4.3% of the reptiles admitted for care and possessed the lowest release rate.

**Table 11. Intake reasons and corresponding released rate for top 20 admitted reptiles**

| | Confiscation | | Habitat Loss | | Infection | | Orphaned | | Toxicity | | Trauma | | Unknown | | Total | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Cases | % Rel. | Cases | % Rel. | Cases | % Rel. | Cases | % Rel. | Cases | % Rel. | Cases | % Rel. | Cases | % Rel. | Cases | % Rel. |
| Common Box Turtle | 11 | 82% | 7 | 29% | 5 | 20% | 10 | 60% | 1 | 100% | 180 | 29% | 36 | 78% | 250 | 38%% |
| Common Garter Snake | 9 | 22% | 18 | 44% | 29 | 24% | 19 | 37% | 2 | 50% | 267 | 33% | 40 | 72% | 384 | 36% |
| Common Map Turtle | 50 | 0% | 0 | - | 1 | 0% | 3 | 100% | 0 | - | 179 | 25% | 50 | 72% | 283 | 41% |
| Diamondback Terrapin | 94 | 0% | 24 | 50% | 5 | 20% | 44 | 55% | 0 | - | 97 | 25% | 51 | 94% | 315 | 55% |
| Eastern Box Turtle | 373 | 6% | 54 | 69% | 631 | 49% | 377 | 51% | 5 | 20% | 3,380 | 33% | 262 | 72% | 5,082 | 41% |
| Eastern Painted Turtle | 100 | 11% | 23 | 26% | 66 | 27% | 68 | 54% | 0 | 100% | 1,197 | 27% | 43 | 70% | 1,497 | 32% |
| Eastern Rat Snake | 12 | 25% | 14 | 57% | 18 | 28% | 8 | 62% | 4 | 0% | 182 | 48% | 18 | 83% | 256 | 50% |
| Florida Box Turtle | 25 | 4% | 3 | 100% | 7 | 57% | 8 | 50% | 0 | - | 93 | 32% | 5 | 60% | 151 | 48% |
| Florida Cooter | 20 | 0% | 2 | 100% | 4 | 25% | 3 | 33% | 0 | - | 81 | 44% | 8 | 50% | 118 | 43% |
| Florida Soft-shell Turtle | 30 | 0% | 2 | 100% | 2 | 50% | 7 | 71% | 0 | - | 181 | 40% | 15 | 67% | 237 | 50% |
| Gopher Tortoise | 88 | 3% | 26 | 88% | 4 | 25% | 111 | 50% | 0 | - | 973 | 35% | 49 | 73% | 1,251 | 41% |
| Midland Painted Turtle | 15 | 0% | 0 | - | 13 | 31% | 10 | 30% | 1 | 0% | 151 | 36% | 12 | 67% | 202 | 41% |
| Painted Turtle | 46 | 0% | 6 | 17% | 39 | 28% | 21 | 81% | 0 | - | 508 | 36% | 12 | 83% | 632 | 42% |
| Peninsular Cooter | 22 | 0% | 1 | 100% | 0 | - | 8 | 62% | 0 | - | 199 | 36% | 3 | 100% | 233 | 43% |
| Plains Garter Snake | 0 | - | 9 | 90% | 0 | - | 393 | 97% | 0 | - | 11 | 72% | 0 | - | 413 | 96% |
| Red-eared Slider | 123 | 8% | 29 | 34% | 36 | 31% | 130 | 83% | 3 | 67% | 509 | 23% | 164 | 51% | 994 | 42% |
| Snapping Turtle | 652 | 2% | 204 | 67% | 123 | 33% | 259 | 66% | 8 | 37% | 3,625 | 29% | 305 | 67% | 5,175 | 36% |
| Three-toed Box Turtle | 21 | 0% | 5 | 80% | 46 | 41% | 19 | 42% | 0 | - | 336 | 26% | 31 | 65% | 458 | 32% |
| Western Painted Turtle | 157 | 4% | 0 | - | 30 | 10% | 4 | 25% | 0 | - | 623 | 18% | 6 | 50% | 820 | 19% |
| Yellow-Bellied Slider | 47 | 83% | 12 | 0% | 15 | 47% | 18 | 33% | 1 | 0% | 274 | 45% | 47 | 70% | 414 | 49% |
| Total | 1,895 | 60% | 438 | 60% | 1,074 | 41% | 1,520 | 68% | 25 | 44% | 13,046 | 31% | 1,167 | 68% | 19,165 | 40% |

# 4.4 Machine Learning

## 4.4.1 Experimental Plan

Since the final disposition is given, we assumed this is a supervised machine learning classification problem. In machine learning, classification is a supervised learning approach in which the computer learns from one or more given inputs, makes observations, and then predictions on one or more outcomes. To be more specific, it concludes the input data or training data and predicts categories or class labels on the output data or given testing data. Some examples of classification problems are spam detection, image segmentation, sentiment analysis, digit recognition, and others. There are various classification algorithms, for instance, Decision tree, Logistic Regression, Gradient Boosting, Support Vector Machine, Random Forest, Naive Bayes, neural networks, and k Nearest Neighbors.

There is not an algorithm better than others; It all depends on the data we are trying to build our model on and the trade-off between performance and explainability.

To assist rehabilitators in deciding which species and factors contribute to the predicted outcomes, it was fundamental to implement an interpretable model. Because of this, three different supervised machine learning algorithms were used: Random Forest, Logistic Regression, and Gradient Boosting (XGBoost).

To implement the algorithms, we'll use hyperparameter tuning via gribsearch. Every machine learning model has different hyper-parameters that can be tuned in to optimize the model performance. The task is achieved by controlling the learning process. This optimization technique is used to iterate through different combinations of values of hyper-parameters until it finds the combination of parameters that give the best performance. The dataset was partitioned into train and test sets. The train set was used in K-fold validation to find the best parameter, and the test set was used to test the model performance.

We later evaluate the performance of each classifier using a classification report, accuracy classification score, AUC score, and confusion matrix to evaluate the accuracy of the classification, evaluate predictions on the test data set, and compare classifiers' performance. The classification metrics report shows the precision, recall, F1, and support scores for the model. The recall score is the proportion of correct classifications from positive cases. The precision score is the proportion of correct classifications from cases that are predicted as positive. Since the classes are uneven, we will focus on the f1-scores. The f1 - score is a combination of the precision and recall scores, measures the model accuracy, and takes both false positives and false negatives into account. The closer the f1- score is to 1, the more predicting power the model has. The AUC curve (Area Under the Curve) represents the degree of separability between classes (euthanized and released). The higher the AUC score, the better the model is at predicting 0s as 0s and 1s as 1s (Note that class 0 = Euthanasia and class 1 = Released). All data reduction, machine learning, and evaluation of predictive models were performed within the python environment.

### 4.4.2 Random forest

Random Forest is a combination of multiple decision trees. In other words, it is an ensemble tree-based learning algorithm. An ensemble algorithm combines more than one algorithm for classifying objects then decides the class label (if we use the classifier) through majority voting. This supervised learning method can be used for classification and regression.

The random forest classifier creates a set of decision trees from a randomly selected subset of the training set.

Some of the advantages of using this algorithm are that it tends to be very accurate, does not need feature scaling, is efficient on large datasets, less sensitive to outliers, and can handle a large number of input variables.  Some of the disadvantages are that it is computationally expensive, and the trees can be hard to visualize.

The table below shows the different parameters that were used to tune the model during cross-validation. The same hyper-parameters were used for the birds, mammals, and reptiles models.

Table 12. Table of hyperparameters for random forest classifier

| Hyper-parameter | Description | Values |
|---|---|---|
| **n_estimators** | The number of trees in the forest. | 100, 300, 500, 800 |
| **max_depth** | Maximum depth of the trees. | 15, 25, 30,50 |
| **min_samples_split** | Minimum number of samples required to split an internal node. | 2, 5, 10 |
| **min_samples_leaf** | The minimum number of samples required to be at a leaf node | 1, 2, 5 |

#### 4.4.2.1 Predicting birds outcome

The best parameters for the birds model were: max_depth: 25 ;min_samples_leaf: 1, min_samples_split:10; and n_estimators: 800. After outputting the best parameters, those parameters were fit on the entire training set, and then made predictions from the model on the testing data.

The figures below show the classification report, confusion matrix, and ROC curve plot. The classification report shows the performance metrics for the model, and the confusion matrix shows the number of correctly and incorrectly classified observations. We later computed the feature importance plot. The most important features for predicting a bird's outcome are the Duration of Care (in days), Anatomical site of Injury_Forelimb/wing/shoulder, Life Stage_Hatchling, and Circumstances of Rescue_Orphan / Parents not available.

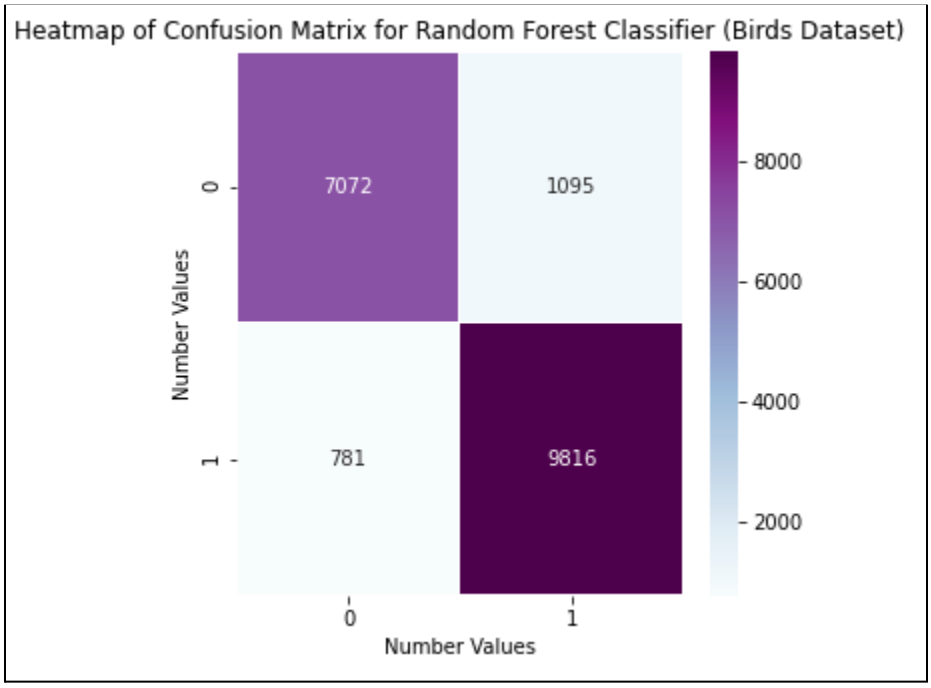
Figure 13. Classification report for random forest (birds dataset)

=== Classification Report ===

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.90 | 0.87 | 0.88 | 8167 |
| 1 | 0.90 | 0.93 | 0.91 | 10597 |
| accuracy | | | 0.90 | 18764 |
| macro avg | 0.90 | 0.90 | 0.90 | 18764 |
| weighted avg | 0.90 | 0.90 | 0.90 | 18764 |

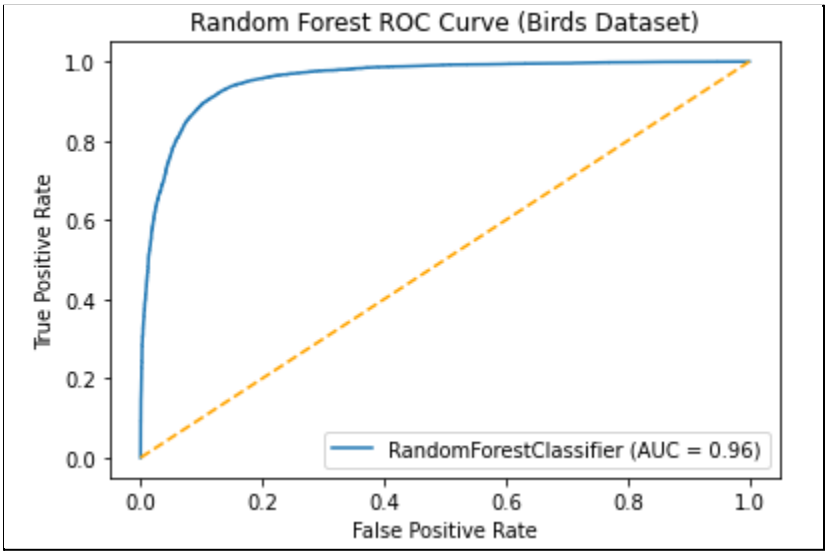
The accuracy score after fitting the model on the testing set was 90%. This means that we were able to correctly predict the probability of euthanizing and releasing a bird 90% of the time. Overall, the model had a good generalization performance. Let's recall what the precision, recall, and f1-score means. The recall score means the proportion of correct classifications from positive cases. For instance, of the 8,167 birds euthanized, 7,105 were classified correctly or were euthanized. The precision score means the proportion of correct classifications from cases that are predicted as positive. For instance, of the 10,597 birds that were classified as released, only 9,537 were released. The model is much better at predicting a bird being released than it is at predicting a bird being euthanized. The F1 score for the euthanasia class is 0.88, and the F1 score for the released class is 0.91.

Figure 14. Confusion matrix for random forest (birds dataset)



The confusion matrix gives us a better idea of which outputs were erroneously classified. For instance, 1,095 euthanizations were classified as released when they were euthanizations.

**Figure 15. ROC for random forest (birds dataset)**




Above is the ROC curve, which shows the curve in blue and a reference curve in dotted red for a "random guessing" model. The AUC score is 0.96, which is fairly close to 1, so the model is fairly good at distinguishing between the two classes

#### 4.4.2.2 Predicting mammals outcome

The best parameters for the model were: max_depth: 50 ;min_samples_leaf: 1, min_samples_split:10; and n_estimators: 300. Similar to what we did on the birds model, the parameters were fit on the entire training set, and then made predictions from the model on the testing data.

The figures below show the classification report, confusion matrix, and ROC curve plot. The classification report shows the performance metrics for the model, and the confusion matrix shows the number of correctly and incorrectly classified observations. We later computed the feature importance plot. The most important features for predicting a mammal's outcome are the duration of care, Life Stage_Infant, and Anatomical site of Injury_CNS-central/brain.

Figure 16. Classification report for random forest (mammals dataset)

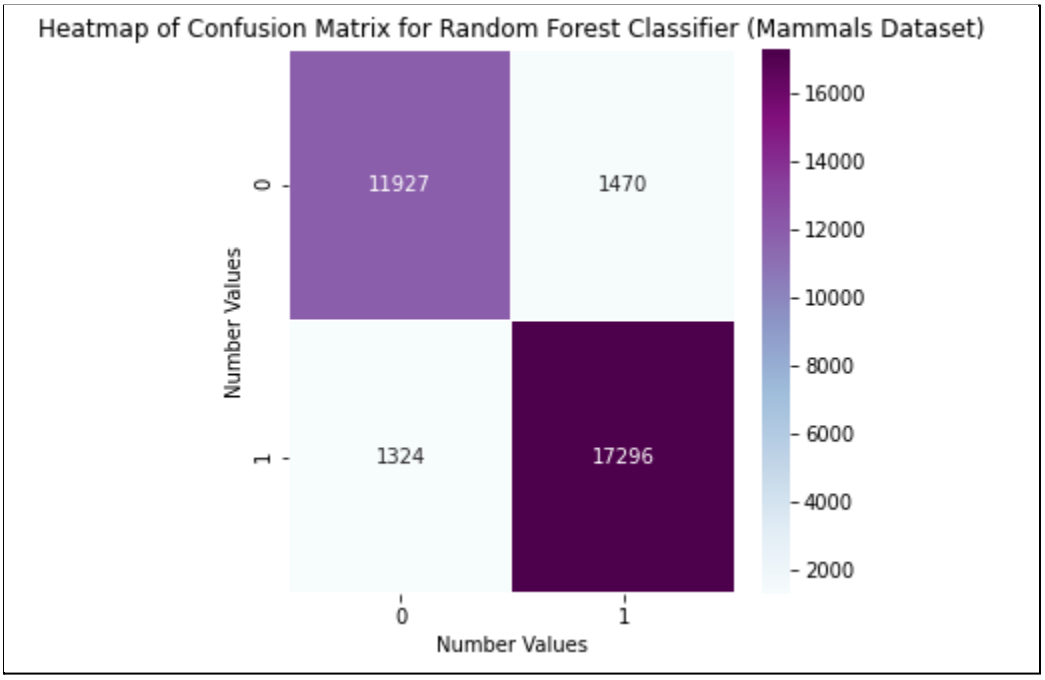
=== Classification Report ===

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.90 | 0.89 | 0.90 | 13397 |
| 1 | 0.92 | 0.93 | 0.93 | 18620 |
| accuracy | | | 0.91 | 32017 |
| macro avg | 0.91 | 0.91 | 0.91 | 32017 |
| weighted avg | 0.91 | 0.91 | 0.91 | 32017 |

Similar to the performance of the birds model, the mammals model was much better at predicting the probability of mammals being released than being euthanized. This might be explained by the fact more birds and mammals were released than euthanized. The accuracy score after fitting the model on the testing set was also higher for the mammals model ( 91.7% accuracy score). This means that we were able to correctly predict the probability of euthanizing and releasing a mammal  91.7%% of the time. Overall, the model had a good generalization performance.

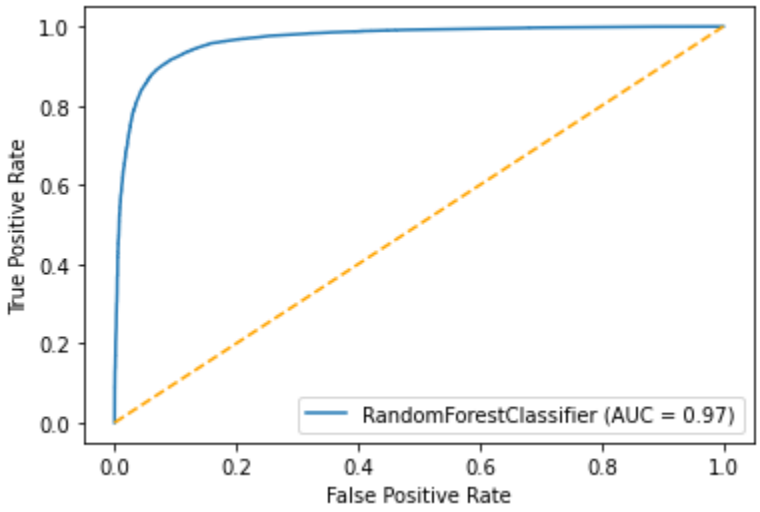
The F1 score for the euthanasia class is 0.90, and the F1 score for the released class is 0.93 Both scores are very close to 1, which is a good indicator of the production power of the model.

Figure 17. Confusion Matrix for random forest (mammals dataset)



The confusion matrix gives us a better idea of which outputs were erroneously classified. 1,470 euthanizations were classified as released when they were euthanizations. In addition, 1,324 released were classified as euthanizations when they were released.

**Figure 18. ROC for random forest (mammals dataset)**




Above is the ROC curve, which shows the curve in blue and a reference curve in dotted red for a "random guessing" model. The AUC score is 0.96, 0.1 higher compared to birds, so the model is fairly good at distinguishing between the two classes.

#### 4.4.2.3 Predicting reptiles outcome

The best parameters for the reptiles model were: max_depth: 50 ;min_samples_leaf: 1, min_samples_split:10; and n_estimators: 500. Similar to what we did on the birds and mammals model, the parameters were fit on the entire training set, and then made predictions from the model on the testing data. The most important features for predicting a mammal's outcome are the duration of care in days, collision with moving objects, and abduction with intent of rescue.

The figures below show the classification report, confusion matrix, ROC curve plot, and feature importance plot.. The classification report shows the performance metrics for the model, and the confusion matrix shows the number of correctly and incorrectly classified observations. Similar to the birds and mammals dataset, we computed the feature importance plot. The most important features for predicting a reptile's outcome are the duration of care, Circumstances of Rescue_Collision / Moving object, Life Stage_Egg , and Anatomical site of Injury_CNS-central/spine.

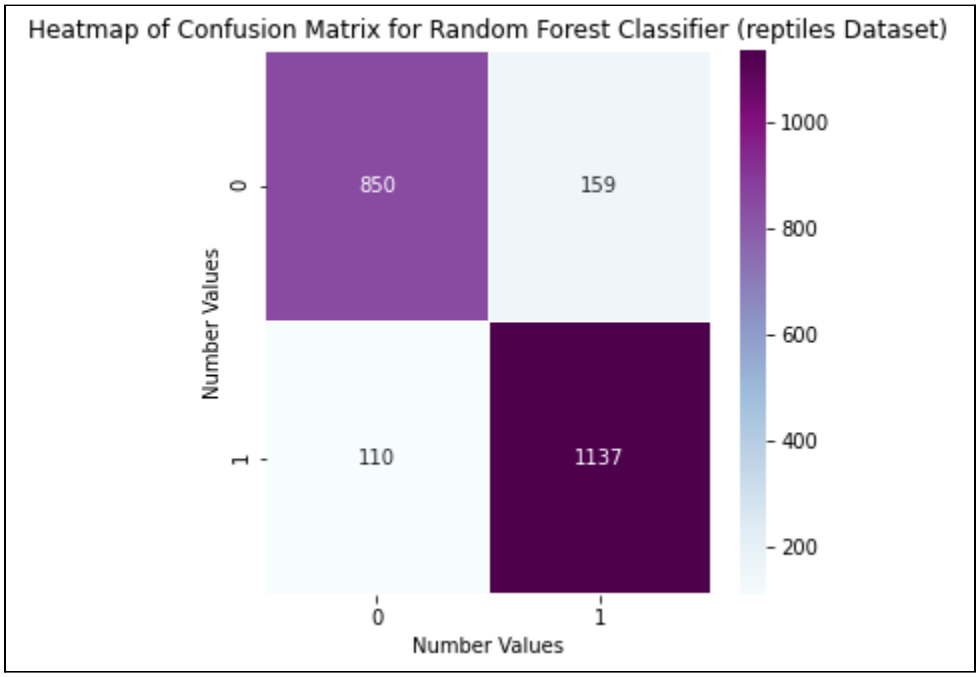
Figure 19. Classification Report for random forest (reptiles dataset)

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

           0       0.89      0.84      0.86      1009
           1       0.88      0.91      0.89      1247

    accuracy                           0.88      2256
   macro avg       0.88      0.88      0.88      2256
weighted avg       0.88      0.88      0.88      2256
```

Compared to the mammals model, the reptiles model performance decreased. This might be attributed to the reduced size of the dataset. Similar to the previous models, this model was much better at predicting the probability of an animal being released than being euthanized. The accuracy score after fitting the model on the testing set was also higher for the mammals model ( 88.1% accuracy score). This means that we were able to correctly predict the probability of euthanizing and releasing a reptile  88.1%% of the time. Overall, the model still had a good generalization performance.

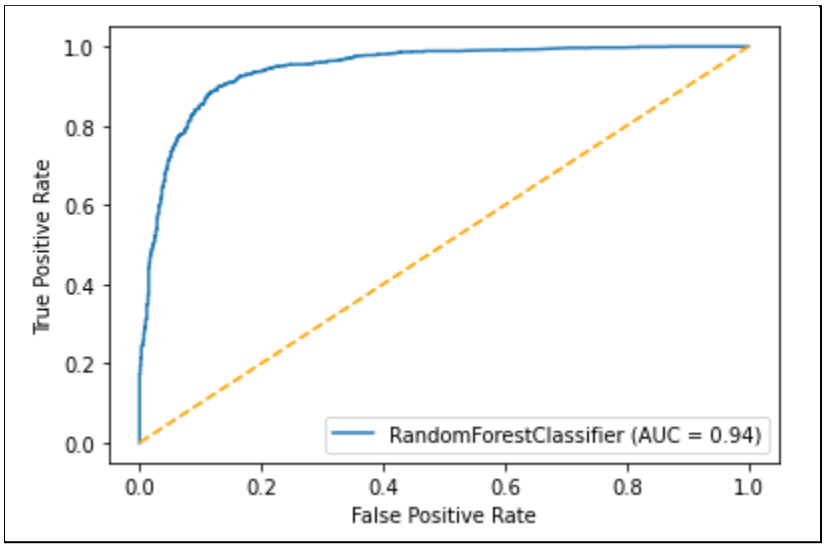
The F1 score for the euthanasia class is 0.86, and the F1 score for the released class is 0.89 Both scores are very close to 1, which is a good indicator of the production power of the model.

Figure 20. Confusion matrix for random forest (reptiles dataset)



On the testing dataset, 159 euthanizations were classified as released when they were euthanizations. In addition, 110 released were classified as euthanizations when they were released.

**Figure 21. ROC for random forest (reptiles dataset)**




Above is the ROC curve, which shows the curve in blue and a reference curve in dotted red for a "random guessing" model. The AUC score is 0.94, still high besides the decrease in the accuracy score, so the model is fairly good also at distinguishing between the two classes.

### 4.4.3 Logistic Regression (Ridge Regularization)

Logistic regression is a supervised learning algorithm used to predict the likelihood of an event happening. It is used for binary classification. Logistic regression works by testing to see if a variable's effect on the prediction is significantly different from 0. If not, it can mean that the variable doesn't help the prediction. This model uses the logistic function, which has the shape of an S, to fit the output between 0 and 1. Regularization is a technique used to reduce or prevent overfitting by adding a regularization term to the equation. Ridge regularization or L2 function works by forcing the weights toward zero but it does not make them exactly zero.

The advantages of Logistic Regression Classifier are: it is easy to implement, interpret, and efficient to train; it is not computationally expensive, and it makes no assumption about the distribution of the classes.  Some of the disadvantages are: it prefers a large dataset; is sensitive to outliers; and might not choose the correct features if they are strictly correlated.

#### 4.4.3.1 Predicting birds outcome

Compared to Random Forest, Logistic Regression had a reduced number of parameters to tune. To select the correct penalization, the penalty parameter was used. This was accomplished by using the hyperparameter tuning grid-search along with K-fold cross-validation.

The figures below show the classification report, confusion matrix, and ROC curve plot. The classification report shows the 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 = released). We later computed the feature importance plot. The top features were different compared to the Random Forest Classifier. The most important features for predicting a bird's outcome are the Anatomical site of Injury_Generalized, Anatomical site of Injury_Hypothermia, Circumstances of Rescue_Entrapment / Trap / Fishing Tackle, and Circumstances of Rescue_Entrapment / Non-trap / Sporting/landscaping netting.

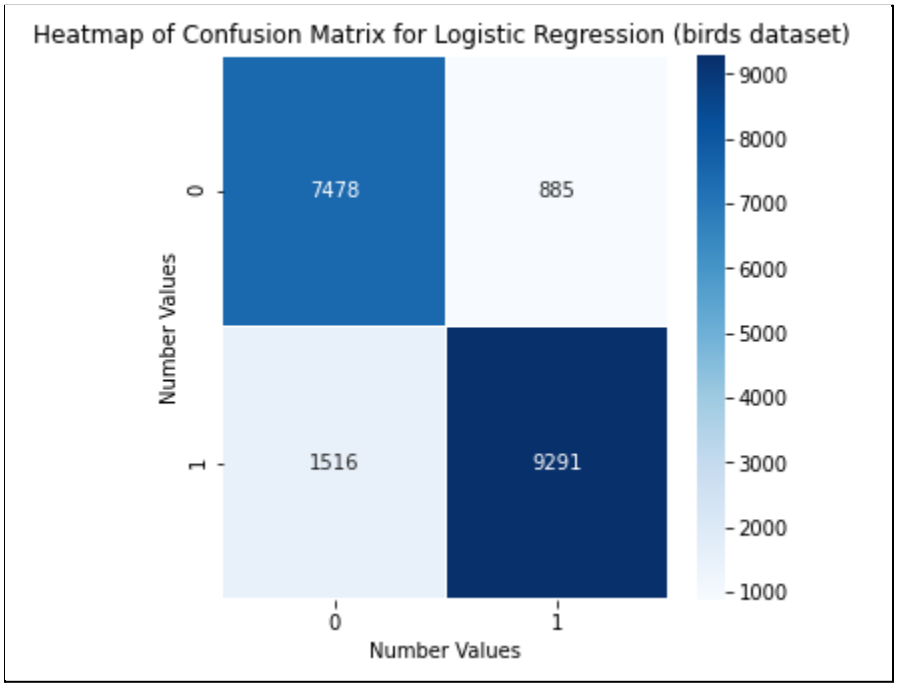
Figure 22. Classification report for logistic regression (birds dataset)

=== Classification Report ===

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.83 | 0.89 | 0.86 | 8363 |
| 1 | 0.91 | 0.86 | 0.89 | 10807 |
| accuracy | | | 0.87 | 19170 |
| macro avg | 0.87 | 0.88 | 0.87 | 19170 |
| weighted avg | 0.88 | 0.87 | 0.88 | 19170 |

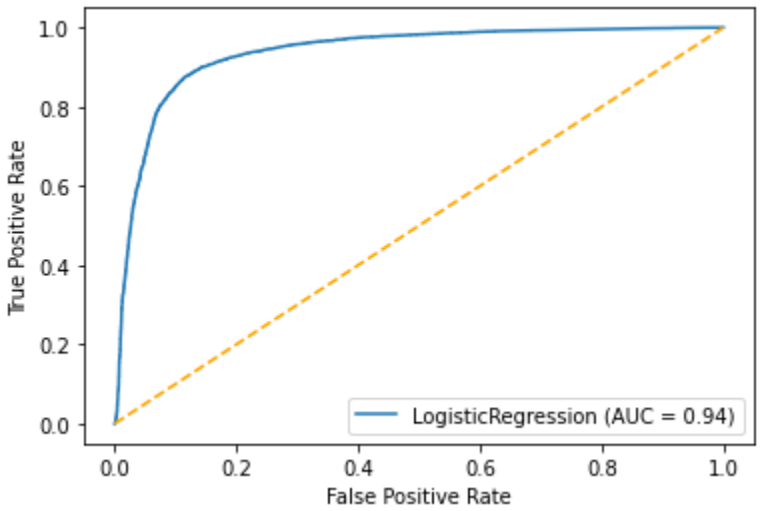
The results for this model were lower compared to the Random Forest. The accuracy score after fitting the model on the testing set was 87.5%. This means that we were able to correctly predict the probability of euthanizing and releasing a bird 87.5% of the time. Overall, the model had a good generalization performance, but lower than expected. Again, the model is much better at predicting a bird being released than it is at predicting a bird being euthanized. Tough, for the first time, the recall score is higher for euthanized animals than released ones. The F1 score for the euthanasia class is 0.86, and the F1 score for the released class is 0.89.

Figure 23. Confusion matrix for logistic regression (birds dataset)




The confusion matrix gives us a better idea of which outputs were erroneously classified. We can observe that the model was more efficient at classifying euthanizations compared to Random Forest. For instance, 885 euthanizations were classified as released when they were euthanized. However, 1,516 released were classified euthanizations when they were released.

Figure 24. ROC for logistic regression (birds dataset)

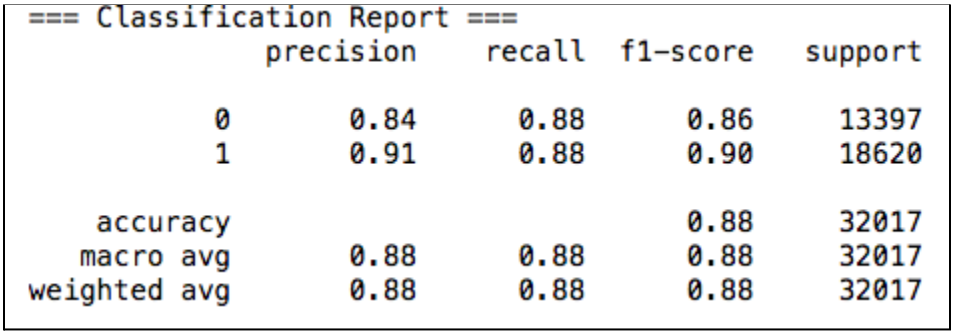


Above is the ROC curve, which shows the curve in blue and a reference curve in dotted red for a "random guessing" model. The AUC score also decreased as expected compared to the Random Forest classifier. The score is 0.94, still high besides the decrease in the accuracy score, so the model is fairly good also at distinguishing between the two classes.

#### 4.4.3.2 Predicting mammals outcome

Next, the logistic regression was used to predict mammal outcomes. The figures below show the classification report, confusion matrix, and ROC curve plot. The top features were different compared to the Random Forest Classifier. The most important features for predicting a mammal's outcome are the Anatomical site of Injury_Generalized, Anatomical site of Injury_Hypothermia, Circumstances of Rescue_Entrapment / Trap / Humane/Cage Trap, and Anatomical site of Injury_Dehydration.

Figure 25. Classification report for logistic regression (mammals dataset)

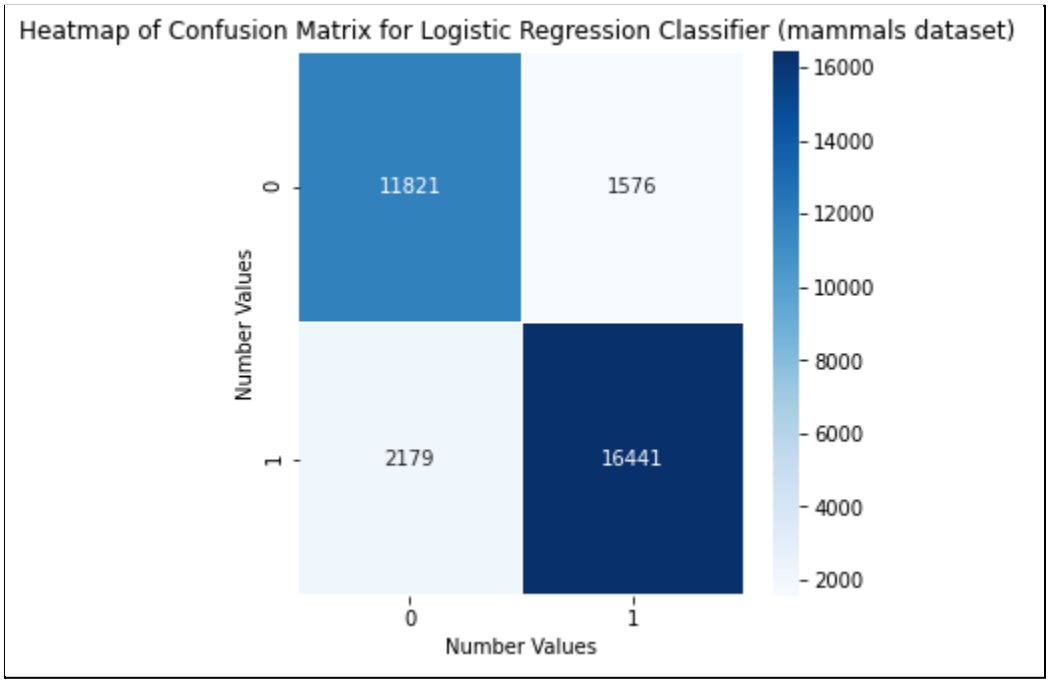


=== Classification Report ===

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.84 | 0.88 | 0.86 | 13397 |
| 1 | 0.91 | 0.88 | 0.90 | 18620 |
| accuracy | | | 0.88 | 32017 |
| macro avg | 0.88 | 0.88 | 0.88 | 32017 |
| weighted avg | 0.88 | 0.88 | 0.88 | 32017 |

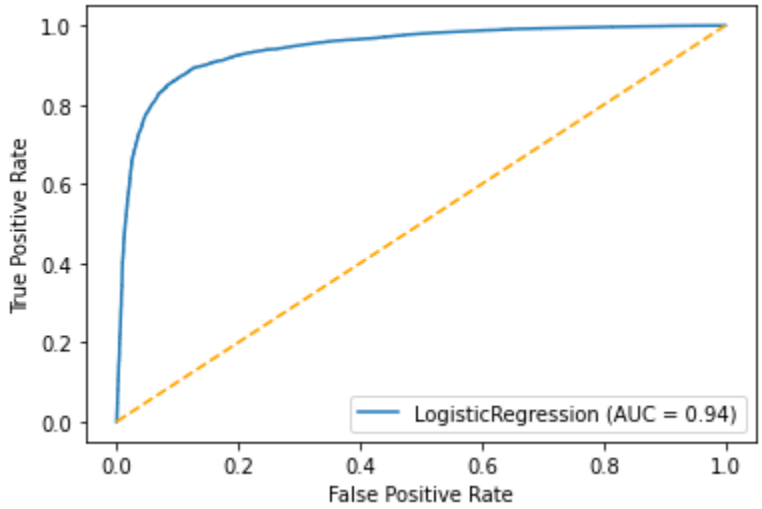
Again, the results for this model were slightly lower compared to the Random Forest. The accuracy score after fitting the model on the testing set was 88.3%. Overall, the model still had a good generalization performance, but lower than expected. Once more, the model is much better at predicting a mammal being released than it is at predicting a mammal being euthanized. The F1 score for the euthanasia class is 0.86, and the F1 score for the released class is 0.90.

Figure 26. Confusion matrix for logistic regression (mammals dataset)




The results of the confusion matrix were also lower compared to those of Random Forest. For instance, 1,576 euthanizations were classified as released when they were euthanized. However, 2,179 released were classified euthanizations when they were released.

Figure 27. ROC for logistic regression (mammals dataset)




Below is the ROC curve, which shows an AUC of 0.94. This is 0.2 points lower compared to Random Forest.

#### 4.4.3.3 Predicting reptiles outcome

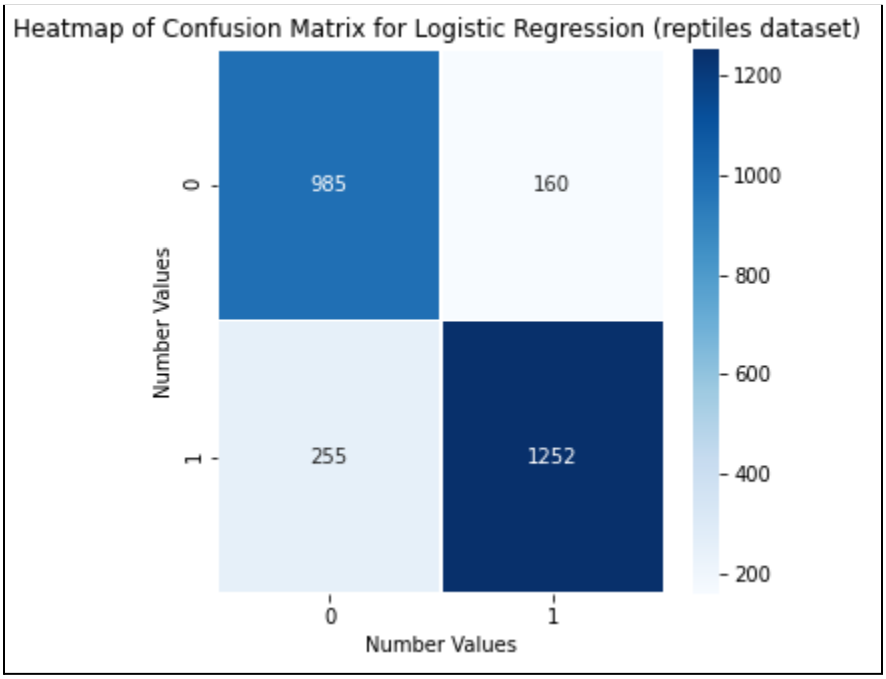
The logistic regression was used to predict reptiles outcomes. The figures below show the classification report, confusion matrix, and ROC curve plot. The most important features for predicting a reptile's outcome are Circumstances of Rescue_Orphan / Parents not available, Anatomical site of Injury_Auditory, and Circumstances of Rescue_Behavioral Stranding.

**Figure 28. Classification Report for logistic regression (reptiles dataset)**

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

           0       0.79      0.86      0.83      1145
           1       0.89      0.83      0.86      1507

    accuracy                           0.84      2652
   macro avg       0.84      0.85      0.84      2652
weighted avg       0.85      0.84      0.84      2652
```

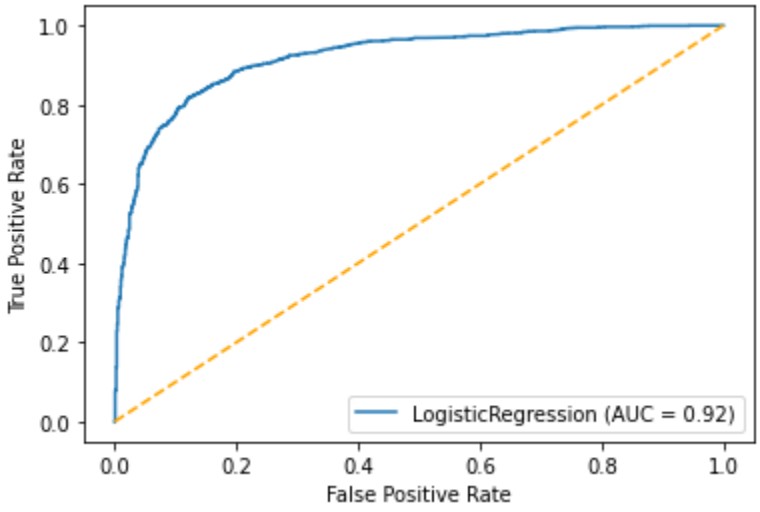
Once more, the results for this model were lower compared to the Random Forest. The precision score for the class 0 decreased by 0.10 points. The accuracy score after fitting the model on the testing set was 84.3%. Almost a 4 points difference in comparison to Random Forest. Overall, the model still had a decent generalization performance, but lower than expected. The F1 score for the euthanasia class is 0.83, and the F1 score for the released class is 0.86.

**Figure 29. Confusion Matrix for logistic regression (reptiles dataset)**



Below is the ROC curve, which shows an AUC of 0.92. This is 0.2 points lower compared to Random Forest.

**Figure 30. ROC for logistic regression (reptiles dataset)**




### 4.4.4 XGBoost

Similar to Random Forest, XGBoost is a decision-tree-based ensemble algorithm. The main difference between these two algorithms is that XGBoost uses a gradient boosting framework. Gradient Boosting combines simple individual models so together they create a more powerful new model. The idea is that when combined with previous models, the new model minimizes the overall prediction error.

One of the main advantages of using XGBoost is speed. Similar to Random Forest, XGBoost is flexible, supports regularization, and has a good generalization performance. Its disadvantages are that it is difficult to tune due to a large number of parameters, difficult to interpret and visualize.

The table below shows the different parameters that were used to tune the model during cross-validation. The same hyper-parameters were used for the birds, mammals, and reptiles.

Table 13. Table of hyperparameters for XGBoost

| Hyper-parameter | Description | Values |
|---|---|---|
| **min_child_weight** | Minimum sum of instance weight (hessian) needed in a child. | 1, 5, 10 |
| **learning_rate** | Use to control the weighting of new trees added to the model. | 0.05, 0.10, 0.15 |
| **colsample_bytree** | Subsample ratio of columns when constructing each tree | 0.6, 0.8, 1.0 |
| **max_depth** | Maximum depth of a tree. | 3, 4, 5 |
| **n_estimators** | Control de number of trees to be used. | 50,100,400 |

#### 4.4.4.1 Predicting birds outcome

The same process implemented in random forest and logistic regression was used for XGBoost. A grid search along with K-fold cross-validation was set to find the optimal parameters. The objective parameter was also implemented to satisfy the condition of binary classification. This parameter was set to 'binary: logistic'.

The best parameters for the birds model were: max_depth: 25 ;min_samples_leaf: 1, min_samples_split:10; and n_estimators: 800. After outputting the best parameters, those parameters were fit on the entire training set, and then made predictions from the model on the testing data.

The figures below show the classification report, confusion matrix, and ROC curve plot. The classification report shows the 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 = released). Similar to Random Forest, the most important features for predicting a bird's outcome are Duration of Care (in days), Common Species Name_Brown Pelican, and Circumstances of Rescue_Entrapment / Trap / Fishing Tackle.

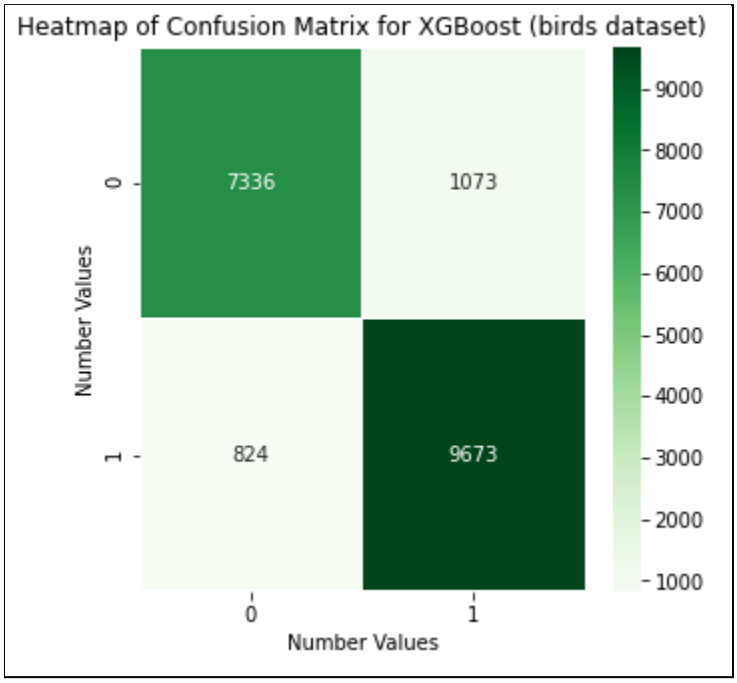
Figure 31. Classification report for XGBoost (birds dataset)

=== Classification Report ===

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.90 | 0.87 | 0.89 | 8409 |
| 1 | 0.90 | 0.92 | 0.91 | 10497 |
| accuracy | | | 0.90 | 18906 |
| macro avg | 0.90 | 0.90 | 0.90 | 18906 |
| weighted avg | 0.90 | 0.90 | 0.90 | 18906 |

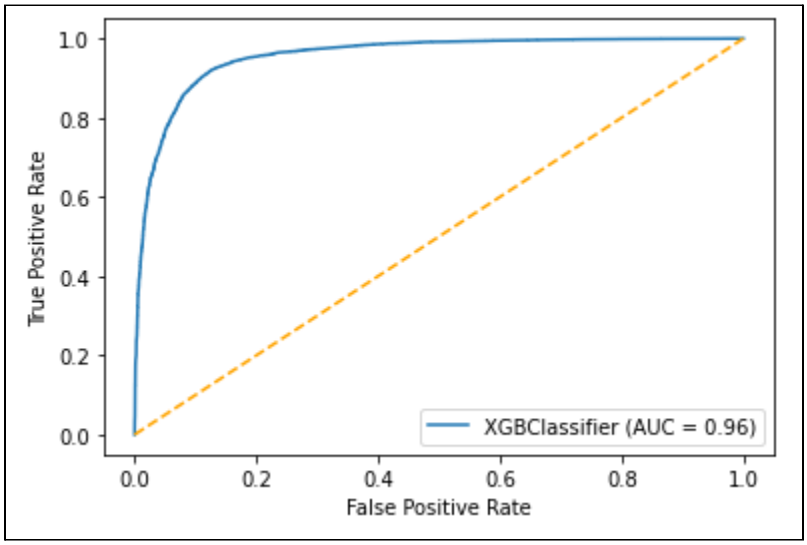
The model performed similarly to Random Forest and this might be due to similarities of both algorithms. The accuracy score after fitting the model on the testing set was 89.97%. This means that we were able to correctly predict the probability of euthanizing and releasing a bird 90% of the time. Overall, the model had a good generalization performance. The F1 score for the euthanasia class is 0.89, and the F1 score for the released class is 0.91.

Figure 32. Confusion Matrix for XGBoost (birds dataset)




1073 euthanizations were classified as released when they were euthanizations. In addition, 824 released were classified as euthanizations.

Figure 33. ROC for XGBoost (birds dataset)




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

#### 4.4.4.2 Predicting mammals outcome

XGBoostwas also used to predict mammal outcomes. The same hyper-parameters that were used in the birds model were also used for the mammals model. The optimal parameters for the colsample_bytree , learning_rate, max_depth, min_child_weight, n_estimators were 0.6, 0.05, 5, 1, 400 respectively. After outputting the best parameters, those parameters were fit on the entire training set, and then made predictions from the model on the testing data. The figures below show the classification report, confusion matrix, and ROC curve plot. The most important features for predicting a mammal's outcome are the Anatomical site of Injury_CNS-central/spine, Anatomical site of Injury_CNS-central/brain, and Duration of Care (in days).

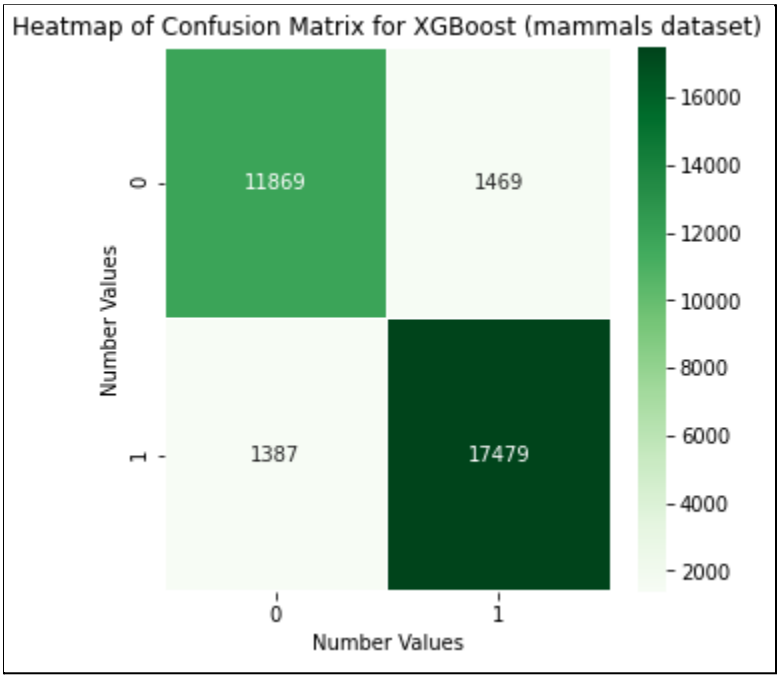
Figure 34. Classification report for XGBoost (mammals dataset)

=== Classification Report ===

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.90 | 0.89 | 0.89 | 13338 |
| 1 | 0.92 | 0.93 | 0.92 | 18866 |
| accuracy | | | 0.91 | 32204 |
| macro avg | 0.91 | 0.91 | 0.91 | 32204 |
| weighted avg | 0.91 | 0.91 | 0.91 | 32204 |

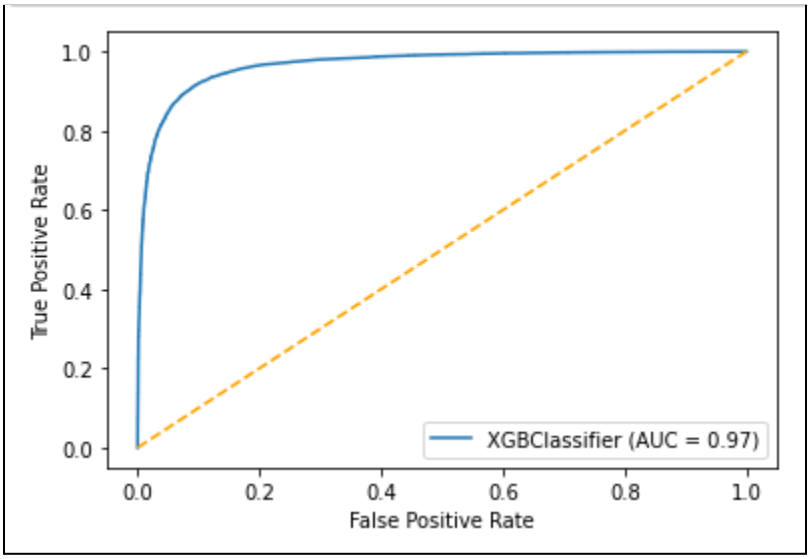
Once again, the model performed similarly to Random Forest. The accuracy score after fitting the model on the testing set was 91.13%. This means that we were able to correctly predict the probability of euthanizing and releasing a mammal is  91.13% of the time. Overall, the model had a good generalization performance. The F1 score for the euthanasia class is 0.89, and the F1 score for the released class is 0.92.

**Figure 35. Confusion Matrix  for XGBoost (mammals dataset)**




The confusion matrix gives us a better idea of which outputs were erroneously classified. We can observe that the model performance is almost identical to Random Forest.  For instance, 1,469 euthanizations were classified as released when they actually were euthanized. However, 1,387 released were classified euthanizations when they were released.

Figure 36. ROC for XGBoost (mammals dataset)




Above is the ROC curve, which shows the curve in blue and a reference curve in dotted red for a "random guessing" model. The AUC score is 0.97, so the model is fairly good also at distinguishing between the two classes.

#### 4.4.4.3 Predicting reptiles outcome

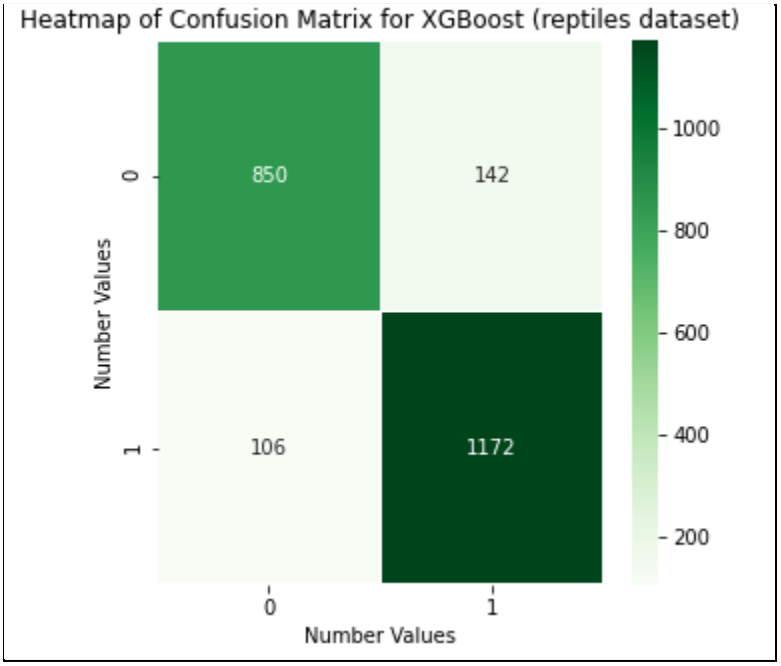
The best parameters for the reptiles model were: colsample_bytree: 0.6, learning_rate: 0.1, max_depth: 3, min_child_weight: 1, n_estimators: 400.Similar to what we did on the birds and mammals model, the parameters were fit on the entire training set, and then made predictions from the model on the testing data. The most important features for predicting a reptile's outcome are the Duration of Care (in days), Circumstances of Rescue_Collision / Moving object / Car/truck/motorcycle, and Life Stage_Egg.

Figure 37. Classification report for XGBoost (reptiles dataset)

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

           0       0.89      0.86      0.87       992
           1       0.89      0.92      0.90      1278

    accuracy                           0.89      2270
   macro avg       0.89      0.89      0.89      2270
weighted avg       0.89      0.89      0.89      2270
```

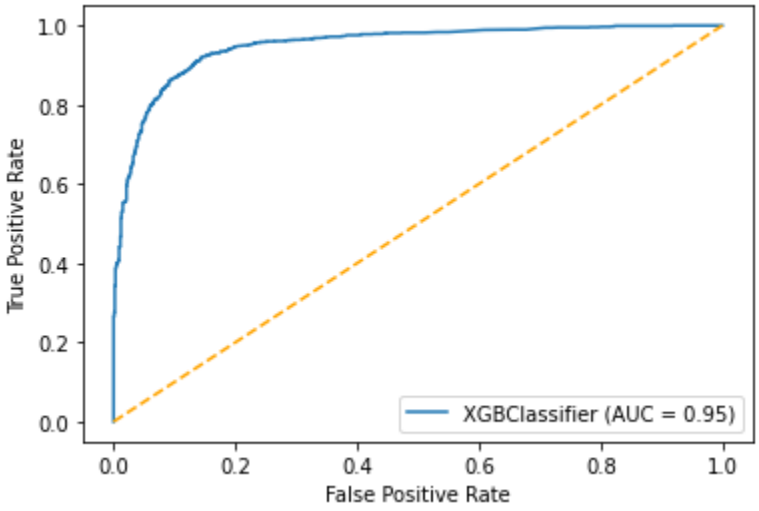
This model performed slightly better compared to Random Forest and Logistic Regression. The accuracy score after fitting the model on the testing set was 89.7%. The F1 scores were also higher compared to previous models, the euthanasia class was 0.87 and the F1 score for the released class was 0.90.

**Figure 38. Confusion Matrix for XGBoost (reptiles dataset)**




The confusion matrix gives us a better idea of which outputs were erroneously classified. For instance, 142 euthanizations were classified as released when they actually were euthanizations.

**Figure 39. ROC for XGBoost (reptiles dataset)**



Above is the ROC curve, which shows an AUC of 0.95. This is 0.1 higher compared to random forest.

## 4.3.7 Assessment

### 4.3.7.1 Model Comparison

A total of three machine learning algorithms were computed. These models included: (1) random forest (RF), (2) logistic regression with ridge regularization (LG), (3) XGBoost. To evaluate the model generalization power and performance, we interpreted the accuracy score, F1-score and AUC score. As displayed in the picture and table below, Random Forest had the best performance and generalization power overall followed by XGBoost and logistic regression. The random forest model for the mammal's dataset had the highest accuracy and F1- score.

**Figure 40. Comparison of Algorithms**

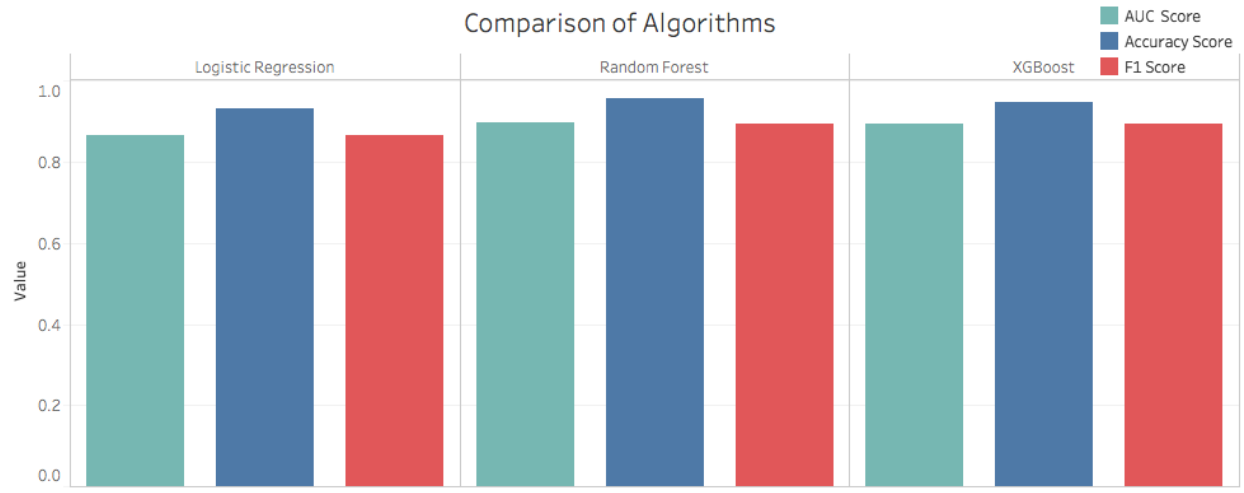



**Table 14. Comparison of Machine Learning Models**

| | Birds | | | Mammals | | | Reptiles | | |
|---|---|---|---|---|---|---|---|---|---|
| | Logistic Regression | Random Forest | XGBoost | Logistic Regression | Random Forest | XGBoost | Logistic Regression | Random Forest | XGBoost |
| Accuracy Score | 0.87 | 0.90 | .90 | 0.88 | 0.92 | 0.91 | 0.84 | 0.88 | 0.87 |
| F1-Score | 0.86 E<br>0.89 R | 0.88 E<br>0.91 R | 0.89 E<br>0.91 R | 0.86 E<br>0.90 R | 0.90 E<br>0.93 R | 0.89 E<br>0.93 R | 0.83 E<br>0.86 R | 0.86 E<br>0.89 R | 0.87<br>0.90 |
| AUC Score | 0.94 | 0.96 | 0.96 | 0.94 | 0.97 | 0.97 | 0.92 | 0.94 | 0.95 |

**E = Euthanized or class 0**

**R = Released or class 1**

# 5.Discussion

Rehabilitation centers play a multidisciplinary role in different aspects, from offering care, public outreach, education, advocacy, and nursing. The rehabilitators also play a key role in the ultimate disposition of each animal they come in contact with, from deciding on which animals they should spend their limited resources, effort, and time to selecting an appropriate release location. The collection and proper report of information about the animals they care for are of equal importance. The data contributes to the understanding of local and national populations, as well as the well-being of the habit shared between the community and wildlife.

## 5.1 Areas of improvement and future work

There were several limitations to the project.  While the data present all wildlife rehabilitation cases over a 9 year spend, we were able to identify reporting errors and inconsistencies both within and between rehabilitation centers. Some of the inconsistencies ranged from spelling errors, duplicate species names, and failure to report distress causes. The major limitation of this project was that some of the fields used to analyze and predict the outcome contained information that didn't provide a clear insight into why those animals were brought in for care. For instance, the variable Anatomical site of Injury had a significant number of 'Clinically Healthy' animals. This term doesn't provide a clear foundation for regulatory management. The models would likely perform better if these fields contained more specific information. As a consequence, the results should be maintained with some degree of skepticism. In addition, a decent number of animals were admitted with evidence of animal or human-inflicted wounds, but there was no specification of the inciting species.

There is still plenty of room for improvement regarding the analytical and predictive goals. For instance, we can study the relationship between the intake of some species and the hunting season. We could also dig deeper into the animals that come due to habitat loss and constructions near the zip code where those animals were found. The results of these possible improvements might assist rehabilitators with educational and staffing purposes. Another improvement that could help gain a better understanding of the common species admitted is grouping them into family categories.

As done by the research work of Melissa Handon, birds can be aggregated into Columbiformes, raptors,  passerines, waterfowl, and other), mammals (large mammals, small herbivores, and small carnivores), reptiles (turtles, snakes, lizards), rabies vector species (bats, raccoons [Procyon lator], and striped skunk [Mephitus mephitis]), and amphibians into (frogs and salamanders). This can help find patterns in family categories that could be more inclined to survive than others.

## 5.3 Reflections

One of the key takeaways from this project is that proper data documentation is fundamental. It is accurate that more data leads to more reliable models, generalization power and performance, and therefore better results, but as long as the data is veracious. It is more recommended to use fewer data rather than more volume but lacking information. Another key takeaway is that domain knowledge is extremely important. Some of the most relevant fields used to build the models used terms that required content knowledge. With decent domain knowledge, we could have reduced the number of categories, aggregated the very similar ones, and identified the ones that were duplicates, and just different terms were used. One more key detail is that a proper and in-depth exploratory data analysis makes the difference. It is very important to be curious about the data, find patterns and recognize the relationships that will improve the model.

# 6. Conclusion

The end goal is to find relevant trends that help explain which reasons for admission and diagnosis have the greatest impact on predicting the animal outcome (released/ transferred or euthanized). The results of these models will help the client gain a better understanding of the factors that can potentially optimize the chances of animal survival. They will also be used as a tool for their multiple educational programs.

The exploratory data analysis phase led to many relevant findings. First, the top dispositions were released, died, and euthanized, accounting for 40%, 30 %, and 30%, respectively. The number of incidents increased almost every year, except for 2019, in which the number of incidents decreased. May, June, and July are the busiest months overall. Collision was the top reason for euthanizing animals. Most euthanizations with 0 and 1 day of admission. Additionally, the top causes for intake were trauma, orphaning, and habitat loss.

We discovered that the busiest states were Florida, Ohio, Wisconsin, and Virginia. We tested the gender hypothesis and found gender did not improve the outcome. The top dispositions in order were released, died, and euthanized.

The three classification algorithms used were random forest, XGBoost, and logistic regression. We implemented the algorithms on the top 3 classes admitted (birds, mammals, and reptiles). Random forest and XGBoost had the best performance, and this can be attributed to the nature of both models. For the birds model, the accuracy score, F1 score, and AUC were 0.90, 0.90, and 0.96, respectively. For mammals, the accuracy, F1 score, and AUC were roughly 0.91, 0.92, and 0.97. For reptiles, the accuracy, F1 score, and AUC were roughly 0.88, 0.97, and 0.94 Some of the top predicting features overall were duration of care in days, Collision with moving object car, truck or motorcycle, and life stage.

There were several limitations to the project. We were able to identify reporting errors and inconsistencies both within and between rehabilitation centers. Some of the inconsistencies ranged from spelling errors, duplicate species names, and failure to report distress causes. There is still room for improvement regarding the analytical and predictive goals. For instance, we can study the relationship between the intake of some species and the hunting season. We could also dig deeper into the animals that come due to habitat loss and constructions near the zip code where those animals were found. We can also aggregate the variable Anatomical Site of the Injury into few categories to use on the predictive models The results of these possible improvements might assist rehabilitators with educational and staffing purposes.

## 6.1 References


Brownlee, Jason. “A Gentle Introduction to XGBoost for Applied Machine Learning.” *Machine Learning Mastery*, Machine Learning Mastery, 16 Feb. 2021, machinelearningmastery.com/gentle-introduction-xgboost-applied-machine-learning/.

Hanson, Melissa, Nicholas Hollingshead, Krysten Schuler, William F. Siemer, Patrick Martin, Elizabeth M. Bunting et al. “Species, Causes, and Outcomes of Wildlife Rehabilitation in New York State.” BioRvix, 19 Nov. 2019.

“Sklearn.ensemble.RandomForestClassifier¶.” *Scikit*, scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html.

Swaminathan, Saishruthi. “Logistic Regression - Detailed Overview.” *Medium*, Towards Data Science, 18 Jan. 2019, towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc.

Data Contributed to the WILD-ONe database (Wildlife center of Virginia, Waynesboro, VA USA;www. wild-one.org)