**Apply Machine Learning in Animal Adoption**

Report for Capstone Spring 2020

Master of Data Science @ George Washington University

Xi Zhang, Siqi Jiang

Group 7

801 22nd St NW, 351

Washington, DC, 20052

GitHub: <https://github.com/xzhangfox/Pet-Adoption-Prediction>

**Table of Contents**

[**Introduction**](#_e2e9eyqvoe3o) **3**

[Introduction/Background](#_jxy47w5rneaj) 3

[Problem Statement](#_569s7yj9pwnh) 3

[Problem Elaboration](#_r5s0jl5nd40k) 3

[Motivation](#_46diqkq65a1z) 4

[Project Scope](#_y442ln5c0uaj) 4

[**Literature Review**](#_fxslw7qvhnzo) **4**

[Relevant Research](#_50df4eegc3bt) 4

[**Methodology**](#_cbg3ov78nm94) **5**

[Dataset Description](#_wyvrqq2bzulg) 5

[Data Collection](#_wyvrqq2bzulg) 6

[Data Preprocessing and Feature Engineering](#_llop967jqb0n) 7

[Data Modeling & Visualizations](#_wyvrqq2bzulg) 8

[Baseline Models](#_6zh4njlzuemy) 8

[Turi-Create Introduction](#_cub1l5eibyhw) 9

[Model Selection and Model Structure](#_hbljimmmublt) 10

[**Results & Analysis**](#_7bo4rhrpvc8f) **11**

[Feature Importance](#_x59ybz15irov) 11

[Adoption Prediction](#_g1wza4528twh) 12

[**Conclusion**](#_r4y8e44jybxs) **14**

[Conclusion](#_8z0q8kqvh7m6) 14

[Project Limitation](#_8z0q8kqvh7m6) 16

[Future Research](#_8z0q8kqvh7m6) 17

[**References**](#_m8ze30u60o3g) **18**

[**Appendix**](#_rntoaclmlhob) **19**

# 

# Introduction

## Introduction/Background

As ASPCA(2020) mentioned in Pet Statistics, more than 70 million stray animals live in the United States, and only six to eight million of those cats and dogs have a chance to be sent to 3,000 shelters across the country each year. About 1.6 million dogs and 1.6 million cats are adopted each year. Instead, about 710,000 lost animals were returned to shelters. Of these, 620,000 are dogs and only 90,000 are cats. Most of the stray animals that did not make it to the shelter died of hunger, cold and disease, while the pets that did not make it to the shelter, after a certain period, were euthanized.

## Problem Statement

The goal of our capstone project is to build a machine learning tool to predict how fast a pet is adopted so that shelters/adoption agencies can improve their pet profiles’ appeal, reduce animal suffering and euthanization, and better focus on their resources to help pets to find new homes.

## Problem Elaboration

Based on the characteristics of stray animals, linear and tree-based models are used to predict adoption rates and to find ways to increase adoption rates. Since cuteness cannot be accurately quantified by a limited set of features, it may be more accurate to use deep learning to process animal images directly. Exploratory Data Analysis provided the deep insights by visualizing. Users can enter animal features or upload images directly to get a prediction of the animal's adoption. According to the result, we may apply pre-trained neural network models as a data augmentation tool or struct a demo for entertainment.

## Motivation

Since we are unable to make any impact on the total number of stray animals in this project, we hope to use our knowledge to provide animal rescuers with a broader vision and thinking. We hope to help more stray animals in the shelter to find suitable adopters. Or by using deep learning models to predict the popularity of pets at the shelter, giving practitioners some direction, the adoption process can be made more efficient.

## Project Scope

In our project, we mainly focus on two parts: Identify the features by structuring classification models, and apply the deep learning models to extract the characteristics from animals’ images for locating adoption speed. After the basic project structure is settled, we keep adjusting the model and data preprocessing to optimize the prediction results.

# Literature Review

## Relevant Research

The previous study includes that by the organization called PetRescue and they focus mainly on creating innovative programs and are eager to deliver them for free to help thousands of rescue pets. By developing and increasing the capability of the artificial intelligence platform, pet care professionals are able to analyze animal photos on social media and find similar matches available for adoption. Recently, big data is continuing to play a big role in managing shelters. China’s top searched engine company, Baidu, designed and manufactured has studied to utilize mass data to improve the shelter profiling system, which they invested in the deep learning methods to profile the potential future owners - based on lifestyle, income, living space, family status, personal needs - and match them with pets will help older animals find homes. Into more technical aspects, the neural network has become one of the most popular tools/models to train and classify the images. James Ng (2019) performed image classification models with convolutional neural networks to identify animal information like breeds and ages for use by the Society for the Prevention of Cruelty to Animals in Singapore.

# Methodology

## Dataset Description

The datasets are provided by Kaggle. There are 14993 pets in the training set and 3948 pets in the test set. The target variable ‘AdoptionSpeed’ represents the window of time in which a pet was adopted after being listed — ranging from 0 to 4 — with 0 signifying the pet was adopted the same day the pet was listed and 4 signifying the pet was not adopted after 100 days of being listed. (Figure 1)

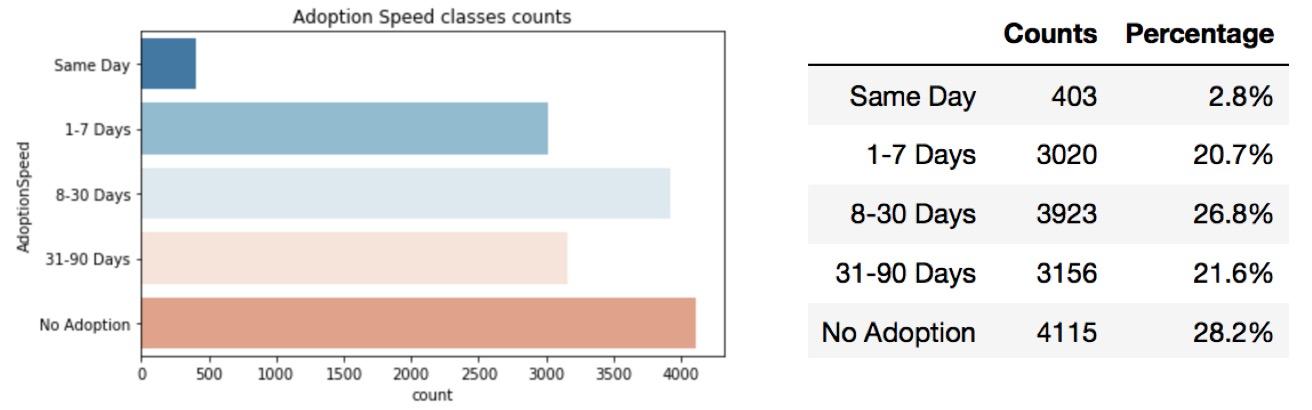


Figure1: Adoption Speed Classes Counts and Percentage

The independent variables include all the different information on the profile for the pet such as breed, color, and gender. Also, it includes image data on the uploaded photos for the pets and sentiment data for the description of the pets. The sentiment data was generated with Google’s Natural Language API.

## Data Collection

There are two types of JSON files in the dataset folder. One is for the sentiment information about the text description of the pet and the other gave us features about provided images. What we did for them is to extract the data provided by the files and compute aggregations on each extracted feature.

For the images files, we applied the flow\_from\_directory method in the Keras ImageDataGenerator. It is very convenient that it creates the pipeline to directly take the path to directories and generates batches of images data. To utilize it, we manipulate the images directory to the path of the directory that contains the sub-directories of the respective classes, like the structure in figure (figure 2). Each subdirectory is treated as a different class. The name of the class is inferred from the subdirectory name.

For the external data, we retrieved the demographics information about Malaysia.

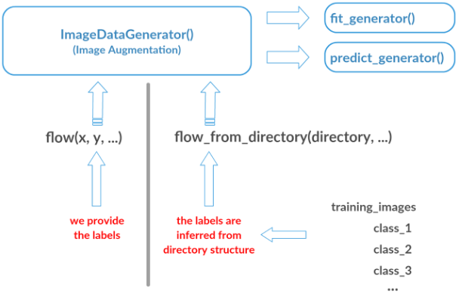


Figure 2: flow\_from\_directory() infers the labels from directory structure

## Data Preprocessing and Feature Engineering

For tabular data, several steps have been made to preprocess the dataset. Below is the highlight of our works.

* **Breeds**: we create dummies to indicate if the animal is a pure breed. Whether it is a domestic or mixed breed. Since the Breed feature has high cardinality, we grouped the lowest occurring levels into an “other” breed and then applied target encoding on them.
* **Age**: we create a dummy feature to indicate the months if the animal is newborn and younger than one-year-old.
* **Sterilized, Dewormed, Vaccinated, and MaturitySize:** we computed the average target for each level of features and changed the order so that higher values match with numbers, which reflects the higher correlation between values and their labels.
* **Multicollinearity**: we checked on the multicollinearity of the variables and filter and drop out the high correlated variable.
* **Geographical Labels**: concept hierarchy generation for geographical labels.

For Images data, the training set is unbalanced, which the smallest class only weighs 2.8% of the total dataset. We implemented image augmentation to solve it. The training set is added with more training samples for the under-represented classes. The augmentation was carried out by changes in scale, rotations, shearing, and horizontal & vertical flips, etc. The augmented images were created using the ImageDataGenerator API in Keras. It is the data augmentation that we ensure that our network, when trained, see new variations of our data each and every epoch. We also created the class\_weight for each category for our models, So, the under-represented classes get more attention. (figure3)

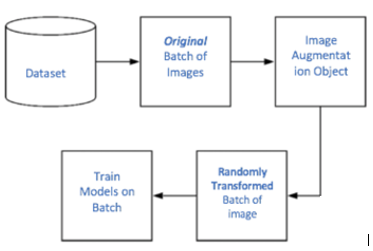


Figure 3: Image Augmentation using Keras

For sentiment data, we derived three variables from the text descriptions:

* **Document sentiment score:**  ranges between -1.0 (negative) and 1.0 (positive) and corresponds to the overall emotional leaning of the text.
* **Document sentiment magnitude**: overall strength of emotion within the given text.
* **The total number of letters** in the description.

These features are derived because we think the pet’s attractiveness is impacted.

## Data Modeling & Visualizations

### Baseline Models

We decided to test out two tree-based classifiers and test them on the validation set to see the baseline performance for the adoption speed projection. The two classifiers are XGBoost and LGBM. The XGBoost is a decision-tree ensemble Machine Learning algorithm that uses a gradient boosting framework, which improves upon the base GBM framework through systems optimization and algorithmic enhancements. The LightGBM is built on the GBDT algorithm, but with the feature of parallel training, lower memory consumption, and generally better accuracy. Once we had our baseline scores we then used Grid Search Cross-Validation to test out different parameter values for each classifier to tune them. We once again ran predictions on the validation sets to see the new improved scores. The baseline Kappa scores and then the tuned scores are displayed. (Table1). We have also made the predictions of the best models using tabular and sentiment data.(Figure 4)

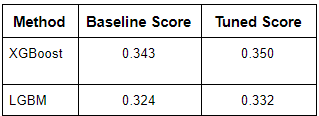
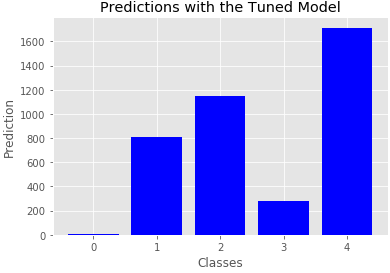
 

Table 1:Baseline Score vs Tuned Score Figure 4: Prediction with Best Result

### Neural Network Model

Turi-Create Introduction

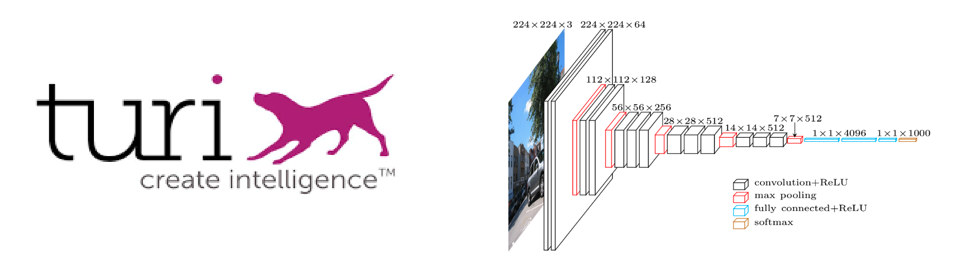


Figure 5: Main Structure of Turi-Create

We trained our image dataset by **Turi-Create** for adoption speed prediction. Turi-Create simplifies the development of custom machine learning models. It has strong computer power and speed. If you have a MacBook Pro, you get access to leverage the Mac GPU resource. No AWS or GCP needed for this library. The core of Turi-Create is ResNet50 which is strong and deep enough to do image classification.

Model Selection and Model Structure

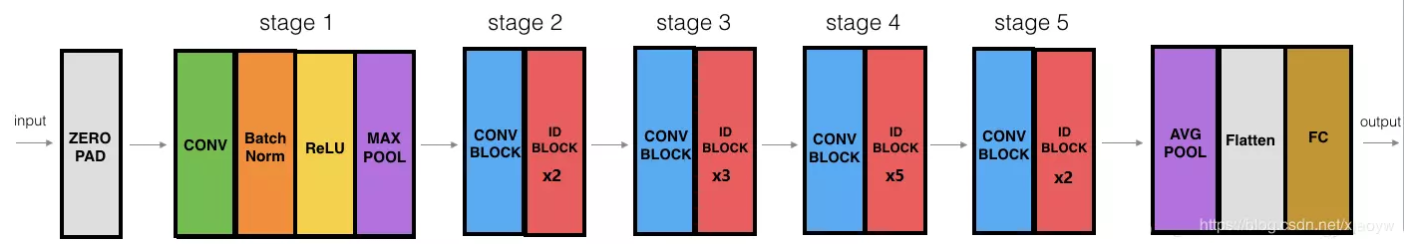


Figure6:: Layers of Turi-Create

ResNet50 has two basic blocks, one is Identity block, the input and output dimensions are the same, so you can concatenate multiple; The other basic block is Conv block. The input and output dimensions are different, so it cannot be concatenated continuously. Its purpose is to change the dimension of the feature vector.

Because CNN finally is to the image a little bit of the convert into feature map is small but the depth is deep, the general formula is unity with the smaller kernel (such as VGG is 3 \* 3), but with the increase of the depth of the network, the output channel also increases (learned more and more complexly, so it is necessary to enter the Identity Block before use Conv Block changes our dimensions, so behind can pick Identity continuous Block.

By using 1x1 convolution, sparse information is compressed and computational force is effectively utilized, so the efficiency is higher. The network is clear, the structure is simple, very standard, and there are different layers to choose from, there is no fixed limit to the size of the input, the application is very wide.

ResNet50 computes faster than ResNet101. Compared with squeezenet\_v1.1, the accuracy of ResNet50 was higher. So all things considered, ResNet50 is the most suitable model for our project. In the Resnet50 of our project, the shape includes batch, in\_height, in\_width, and in\_channels. Since we set our batch size as 64, the input shape will be [64,224,224,3] in this model and after processing from the fc1000 layer the final output shape would be: [64, 1000].

# Results & Analysis

## Feature Importance

In XGBoost we can see the impact of each feature on adoption rates. In order to increase adoption rates, we may be able to do this by changing some features. But as can be seen from the chart, many characteristics are innate in animals and cannot be changed artificially. For example, species, age. But fortunately, the number of promotional photos, whether or not the animal is sterilized, and the length of its hair is still crucial to adoption success. The number of photos is especially important here. Therefore, we strongly suggest that the staff of the animal shelter can allocate more time to take photos and publicize the new animals. Next careful sterilization and grooming.

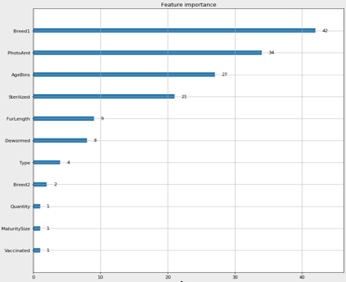


Figure 7: Feature Importance

## Adoption Speed Prediction

# 

Figure 8:Analysis with Cat and Dog

We labeled our images into fast or slow and the dividing line is 30 days. In shelters in many countries, 30 days means the end of the animal. And as you can see from these two tables, whether the data in animal types or adoption speed, the subsets are balanced.

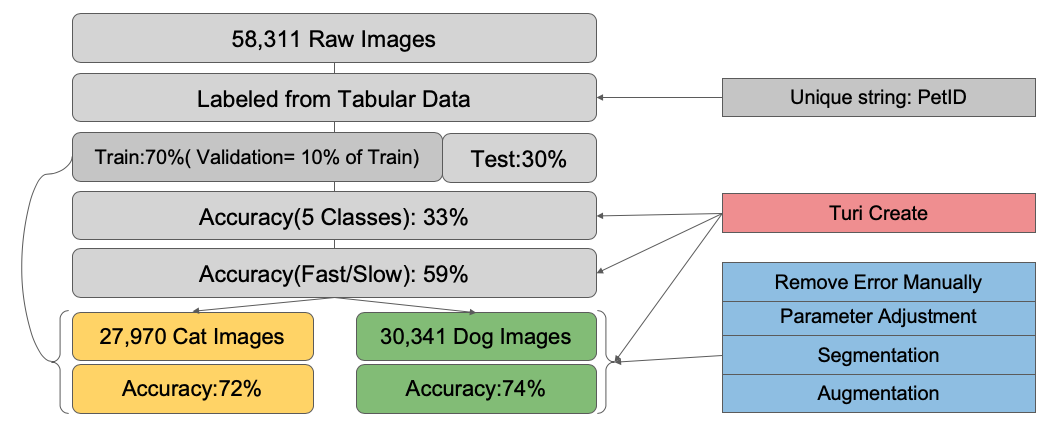


Figure 9: Complete Model Flow

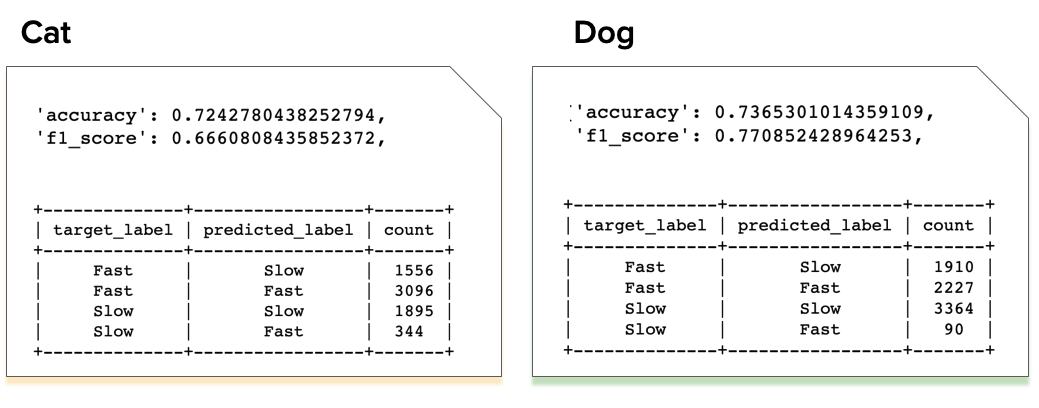
This flowchart shows my complete flow from data to results. (Figure 9) The raw images are not labeled but they are named by Pet ID, which is also contained in our tabular data. Then we leverage this connection to label the images.

Since our dataset is large, we believe a 70% train set would be enough, and in the training set, we set aside 10 percent for validation.

We tried to train the data with 5 classes which are set from tabular data originally. But the result was bad. 33% accuracy means nothing in this case. Then we merged the classes, split it from 30 days, into fast and slow, which became a binary classification issue and it is still reasonable.

However, 59% accuracy for binary classification. The performance didn’t improve.

I separated cat and dog images, then concentrated on the preprocessing part, augment data, did the segmentation, removed some wrong photos manually, and adjusted the hyperparameters of Turi-Create. Finally, the accuracy of cat adoption speed approached 72% and dog 74%. This suggests that we can indeed use image classification to predict adoption rates.



Here are our confusion matrices. If the accuracy feedback is not comprehensive enough, you can see our f1 scores are not bad as well. In our Confusion Matrices, you can see that the Numbers in Slow/Fast are much lower than the other ones. This is the result of incomplete data cleansing. Theoretically, the error data should be relatively evenly distributed throughout the data set. But our approach to data cleansing is time-consuming and laborious. We pulled out all the results of the prediction errors and compared them with the error messages. For example, the wrong label of dog and cat, the wrong label of the group photo, etc. But it also includes data that are normal, but the predictions are wrong. So the current result is only the result after we manually delete part of the obviously wrong data. We should figure out how to eliminate this type of error by automatically traversing. We already have a model to find out what's wrong with the species, but group photos require object detection, and we don't have time to do that right now. But we will continue to do so in the future.

# Conclusion

## Conclusion

As American Society for the Prevention of Cruelty to Animals stated in 2020: about 6.5 million pets enter U.S. animal shelters each year. About 3.3 million of those are dogs and 3.2 million are cats. We estimate that the number of dogs and cats entering U.S. shelters each year has declined from about 7.2 million in 2011. The biggest decline was in the number of dogs (from 3.9 million to 3.3 million).

About 1.5 million shelter animals are euthanized each year (670,000 dogs and 860,000 cats). The number of cats and dogs euthanized each year in the United States has fallen from about 2.6 million in 2011. The decline can be partly explained by an increase in the proportion of adopted animals and the number of stray animals successfully returned to their owners.

About 3.2 million shelter animals (1.6 million dogs and 1.6 million cats) are adopted each year. About 710,000 lost animals have been returned to shelters. Of these, 620,000 were dogs and only 90,000 were cats.

As we illustrated in the overview of the issue, we hope that through our program, we can increase the number of animals admitted to shelters each year, increase the number of animals adopted from shelters, and reduce the number of animals abandoned and euthanized.

Our prediction accuracy is over 70 percent, and I believe it can get even better after that. Ideally, our program could be embedded in the equipment of animal protection personnel looking for stray animals in the wild. By taking pictures of the animals they find, conservationists can know in advance the likelihood that the animals will be adopted, and adjust their strategies by sorting them into different groups.

In shelters, staff can also predict adoption rates by analyzing photos in advance. For faster animals, they can be placed in a more prominent position so that they can be adopted quickly. For slower animals, they can improve their adoption rates by posting more photos and grooming on the site. I believe that with the anticipation of information, they will have more strategies to increase the adoption rate.

For animal owners, they can find out the adoption rate by uploading photos of their pets to our project website. For the vast majority of loving and compassionate people, seeing that their pets may or may not be adopted by someone else can also reinforce their determination not to abandon their pets and try to overcome real difficulties.

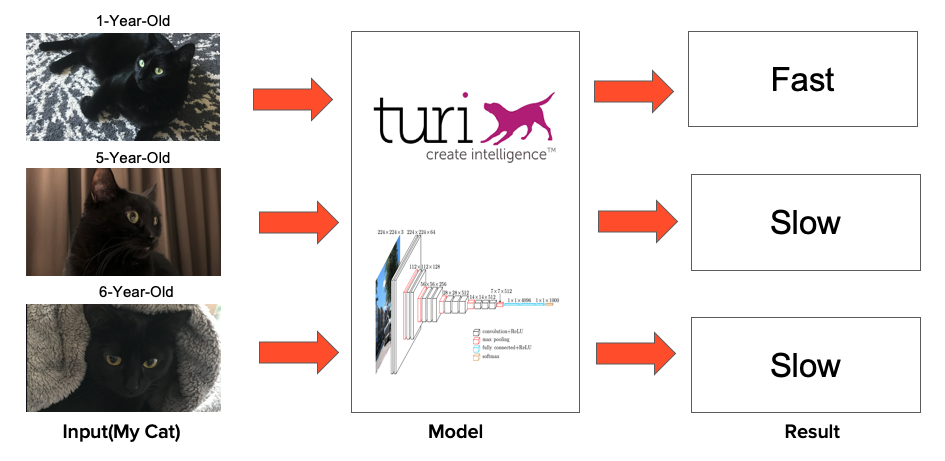


Figure 10: Demo Display

Here my cat is 1-year-old, 5-year-old and 6-year-old. (Figure 10) I input the photos for the demo practice. If we make our project as a real product, the logic should be like this. The result showed only my cat at 1-year-old could be adopted fast. Looking back to our feature importance part, the model improved that age plays an important role in adoption speed. This result looks reasonable, but we still have a lot of space to improve.

## Project Limitation

In order to test our model’s capability, we extracted a subset of the data for identifying the species. The accuracy approaches almost 100% in this case. This shows that our model is very powerful in the task of image classification. So we believe some of the characteristics that

affect adoption rates are not accounted for (The timing of the animals being taken to shelters, resistance to political and economic factors).

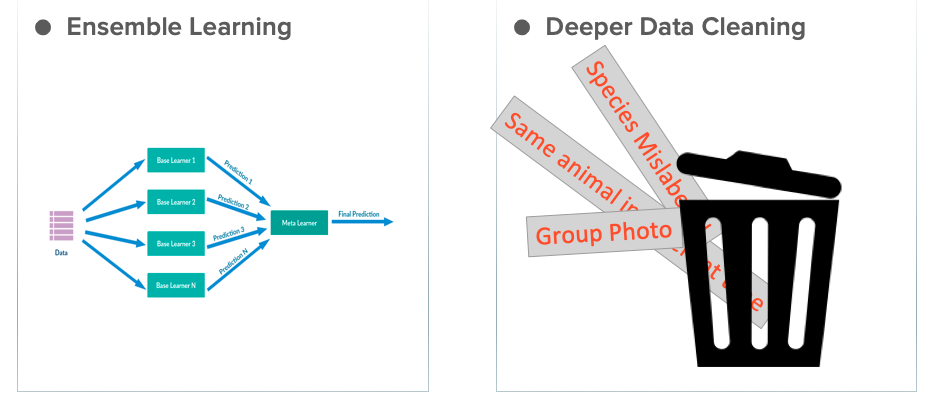
Furthermore, when we apply the species classifier into the whole dataset, the accuracy decreased, then we found some images are mislabeled



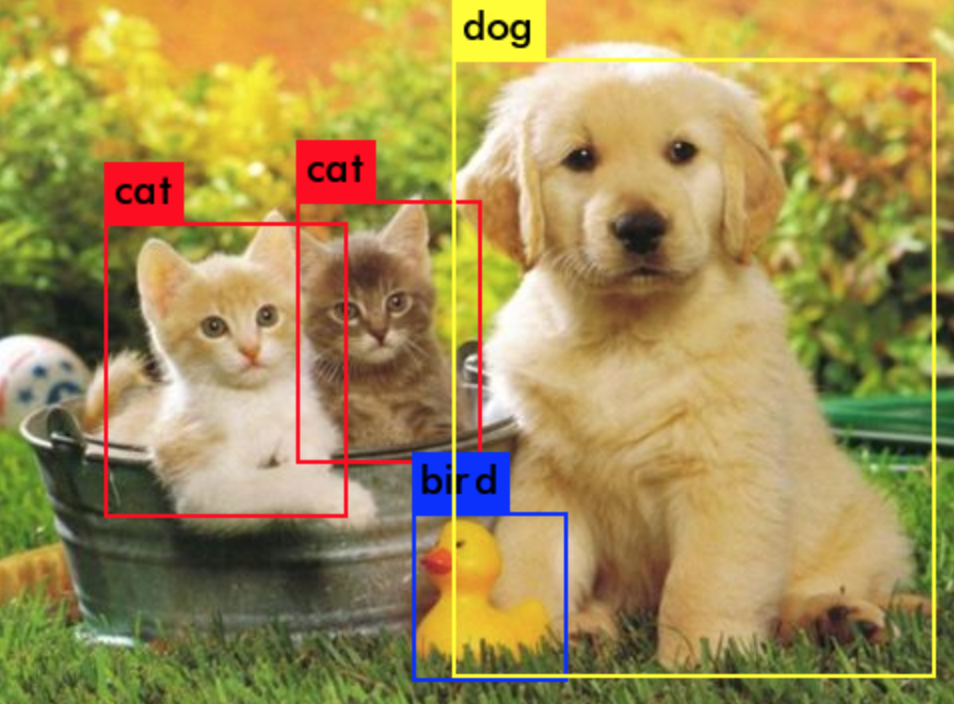
Figure11: Mislabeled Photos

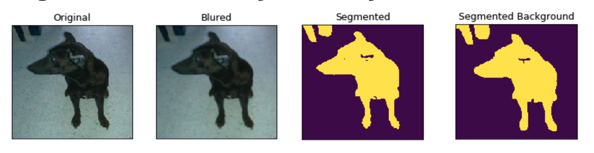
Besides dog and cat mislabels, other errors like group photos but with a single label, same animal but at different times also cause trouble to feature extraction. I manually removed some kinds of data like that and the result did improve.

## Future Research



For further research, we plan to apply ensemble learning to improve machine learning results. However, the improvement from algorithms will be very limited at this status. We must discover, fix, or remove more anomalies.



In the early stage of image data and processing, we found that it was difficult for the same model and parameters to perfectly cut images of different background types, different colors, and different types. After the data cleaning is relatively complete, we can further optimize by tuning the segmentation model.

For dealing with the group photos, because all photos have a single label, we regard all group photos as mislabel. However, the object detection model is required to identify several bodies in the photo. We don't have time to build it this time, so we'll do it in a follow-up study.

# References

Pet Statistics. (n.d.). Retrieved from <https://www.aspca.org/animal-homelessness/shelter-intake-and-surrender/pet-statistics>

Apple. (2020, April 30). apple/turicreate. Retrieved from <https://github.com/apple/turicreate>

Keras. (2017, December 12). ResNet-50. Retrieved from <https://www.kaggle.com/keras/resnet50>

PetMe uses AI to drive dog adoptions. (n.d.). Retrieved from <https://www.ddb.com.au/petme-uses-ai-to-drive-dog-adoptions>

Liao, S. (2019, February 11). Baidu made a smart cat shelter that uses AI to tell cats and dogs apart. Retrieved from <https://www.theverge.com/2019/2/11/18220606/baidu-smart-cat-shelter-ai-facial-recognition-dogs>

Build software better, together. (n.d.). Retrieved from <https://github.com/topics/object-detection>

# Appendix

Exhibit 1: Columns Details

