Animal Adoption from Shelter

Dequan Zhang

Data Science Institute

Final Project Report

https://github.com/DQZ25/Animal_Adoption_Project.git

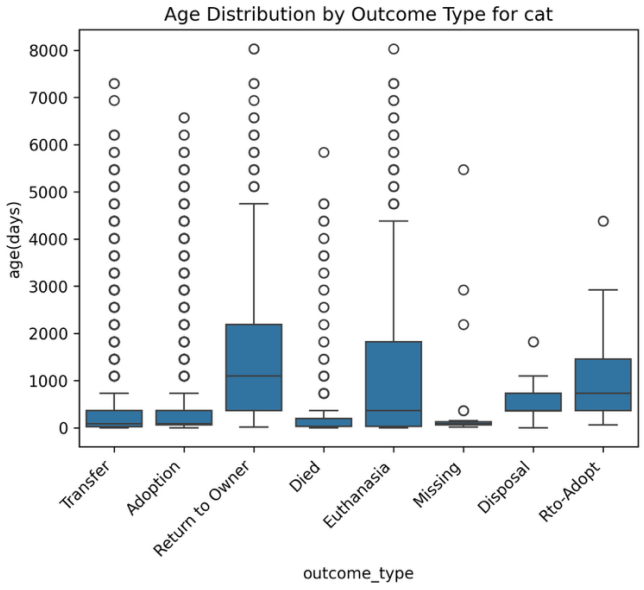
1. Introduction

This report is about shelter animal adoptions, and there are one main topic and one minor topic included. The main topic is, what is the reason will cause an animal to leave the shelter, in data science word, it is can we make any predictions for what will happened to an animal in shelter. As we know, there are many stray cats, dogs or other homeless animals are sent to animal shelters every day. If no one choose to do the adoptions for these animals, some of them may can stay in the original shelter or be transferred to another shelter if it is a “no-kill” shelter. However, if these animals are in a “kill” shelter, these shelters are often forced to euthanize animals based on their duration of stay so they will have enough cage space available to accept all animals.[2] So, for better saving these animals’ life, the prediction mentioned above can be used to make the public learn more about what may happen to shelter animals and promote people to do the animal adoption from shelter. The secondary topic is using the findings about the combination of breeds, colors, and age of different animal from the data analyze to inform the public what kinds of animals they can get from the shelter, which is also helpful for increasing the people’s interest in animal adoption.

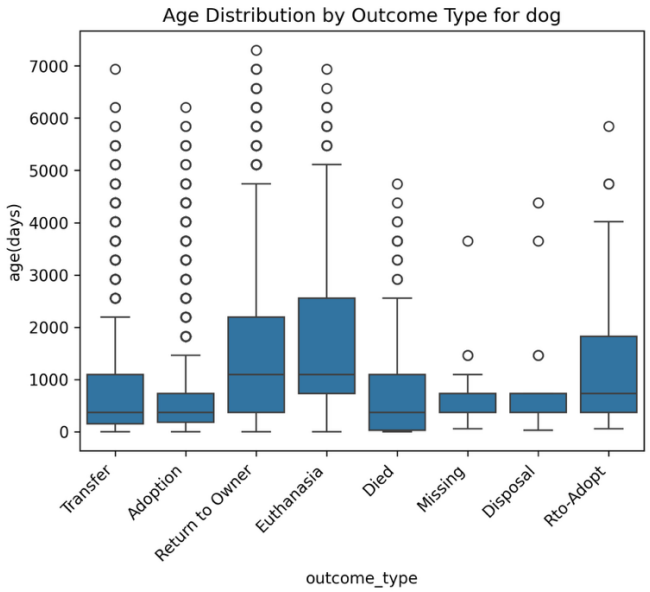
My question is a multiclass classification question. I collect the data from Kaggle, and the data name is Austin Animal Shelter Outcomes. [1] The dataset contains the age of animals, animal id number, animal types, breeds, colors, birth date information, names, outcome information, sex, and sterilization situation.

2. Explanatory Data Analysis

As we know, the most important thing for analyzing data is to understand what the dataset is telling us. I used sum and mean function to check the missing value information in the dataset and divided the data into cat and dog category to check how the target outcome would be different for them. Based on plots of Age Distribution by Outcome Type for cat and dog, the overall idea about what will happen to shelter cats or dogs according to their age were visually provided for us.

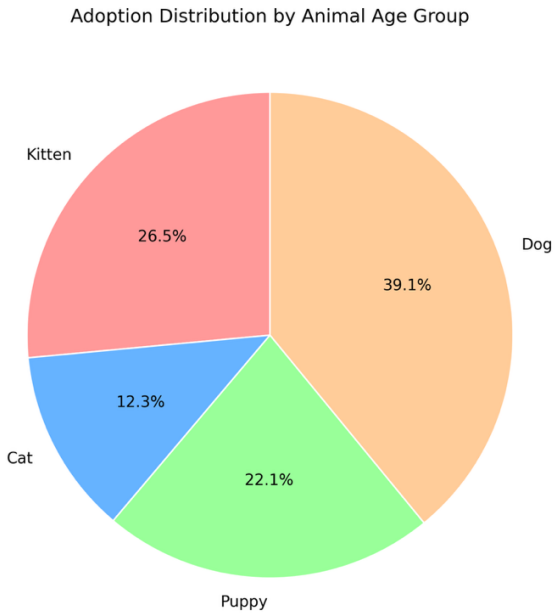


<Table 1: The Age Distribution by Outcome Type for Cat>



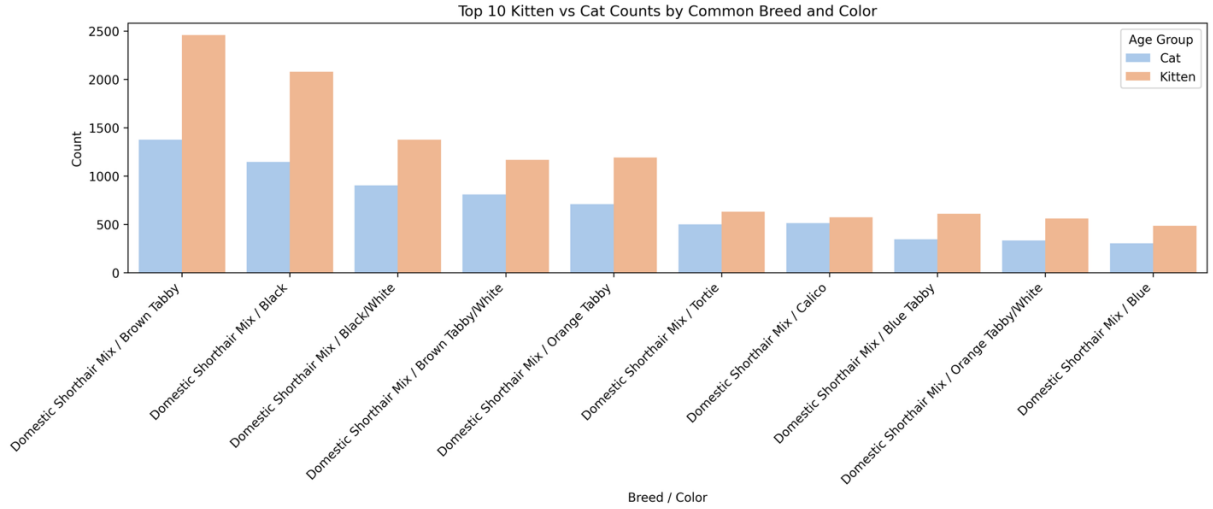
<Table 2: The Age Distribution by Outcome Type for Dog>

The pie chart about the adoption by animal age also tells us that it is usually the younger cats that are adopted, but for dogs, adult dog may be more chosen by people to adopt.

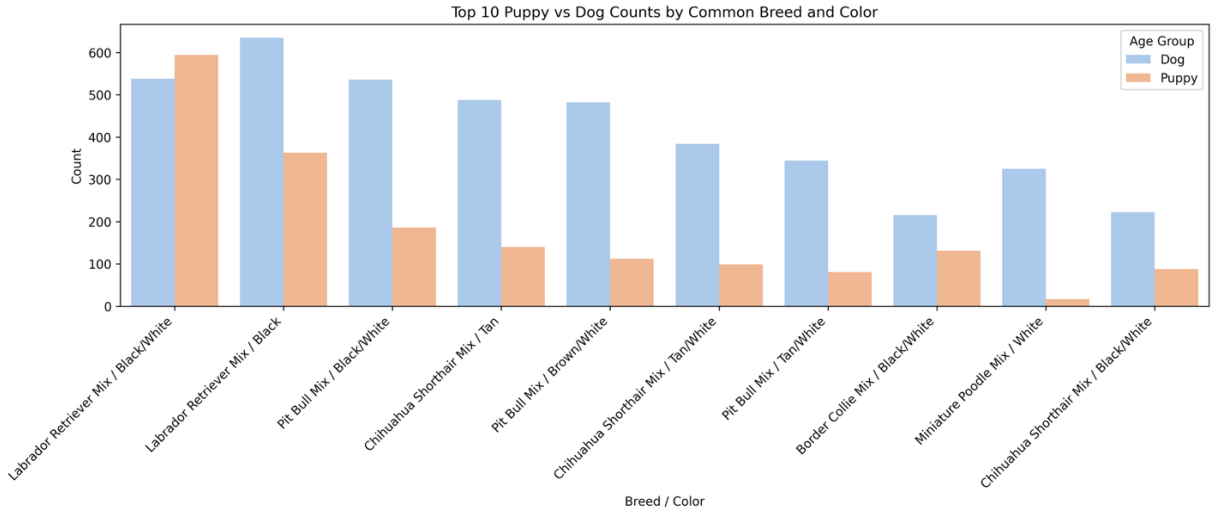


<Table 3: Adoption Distribution by Animal Age Group>

The plot for number of kitten and adult cat group by breed and color shows what are the most common types of cats in shelters. The plot for number of puppy and adult dog also has the same values. These two plots also provide people information about what they can choose from the shelter to adopt, which is definitely a way to improve people's interest to do the animal adoption from shelters.



<Table 4: Top 10 Cat Counts by Common Breed and Color>



<Table 5: Top 10 Dog Counts by Common Breed and Color>

3. Methods

To evaluate the prediction performance on the dataset, I first divide the dataset into a training + validation set (80%) and a test set (20%) by using the sklearn stratified split. Within the training + validation set, a 3-fold cross validation procedure is applied through GridSearchCV, where data is repeatedly split into three folds: in each iteration, two folds are used to fit the preprocessing pipeline and logistic regression model, and the remaining fold serves as a validation set. I repeat the full experiment with three different random seeds and compare the metric value during the process.

Because there are missing values in dataset, for the preprocessing part, I applied the SimpleImputer and used the string “unknown” to replace the missing values, which can help the encoding step to properly recognize and represent these missing categories for categorical variables. Also, these variables are then transferred using One-Hot Encoding. For numeric feature, I used the IterativeImputer to deal with the missing value and StandardScaler to place the feature on a comparable scale for the model.

For the metric selection, since the outcomes of target variable in my dataset is imbalanced and it is a multiclass classification question, I used the macro-averaged f1 score as the primary evaluation metric. The macro-averaged F1 score is computed using the arithmetic mean of all the per-class F1 scores.[3] I also include the accuracy score, but it was only used as a secondary metric because it might be misleading in imbalanced classification dataset.

I used four machine learning algorithms for analysis: Logistic Regression, SVM, Random Forest, and XGBoost. For XGboost, it also has an early stop round of 50. For each algorithm, the tuned parameters are shown in the below table.

Logistic Regression	**C: [0.01, 0.1, 1, 10]**
SVM	**C: [0.1, 1, 10]**
Random Forest	**n_estimators: [100, 200],** **max_depth: [5, 10, None],** **max_features: [0.1, 0.3, 0.5]**
XGBoost	**model_max_depth: [3, 5],** **model_learning_rate: [0.05, 0.1],** **model_reg_alpha: [0, 0.1]**

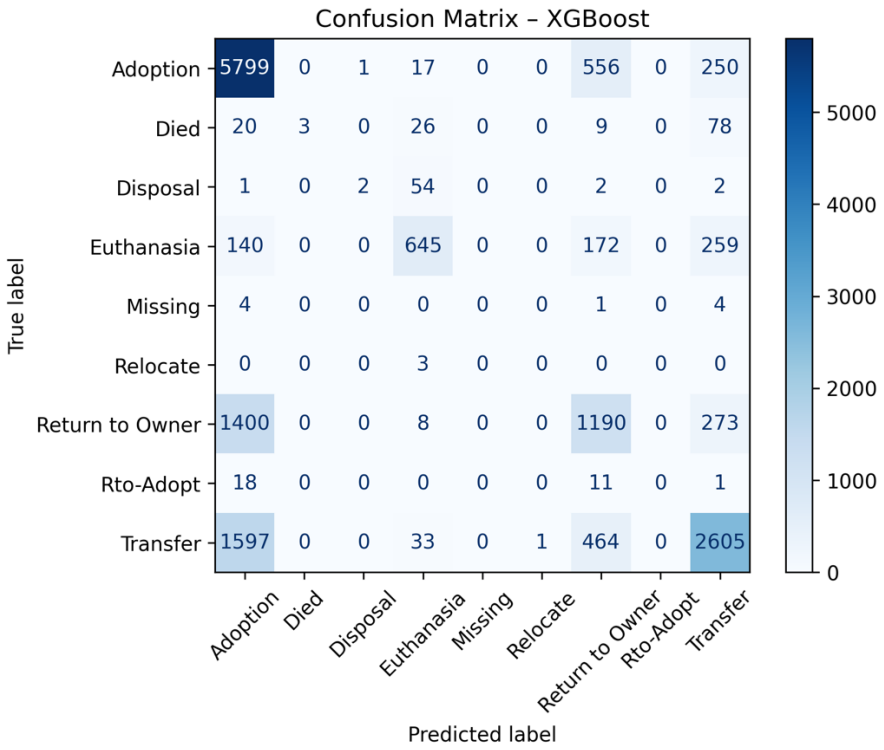
<Table 6: Tuned Parameter Value for Each Machine Learning Model>

4. Result

Model	F1_mean	F1_std	F1_std_above_baseline
Logistic Regression	0.283764	0.007463	29.170298
SVM	0.272230	0.003103	66.428299
Random Forest	0.260521	0.007088	27.434071
XGboost	0.288502	0.000785	283.443572

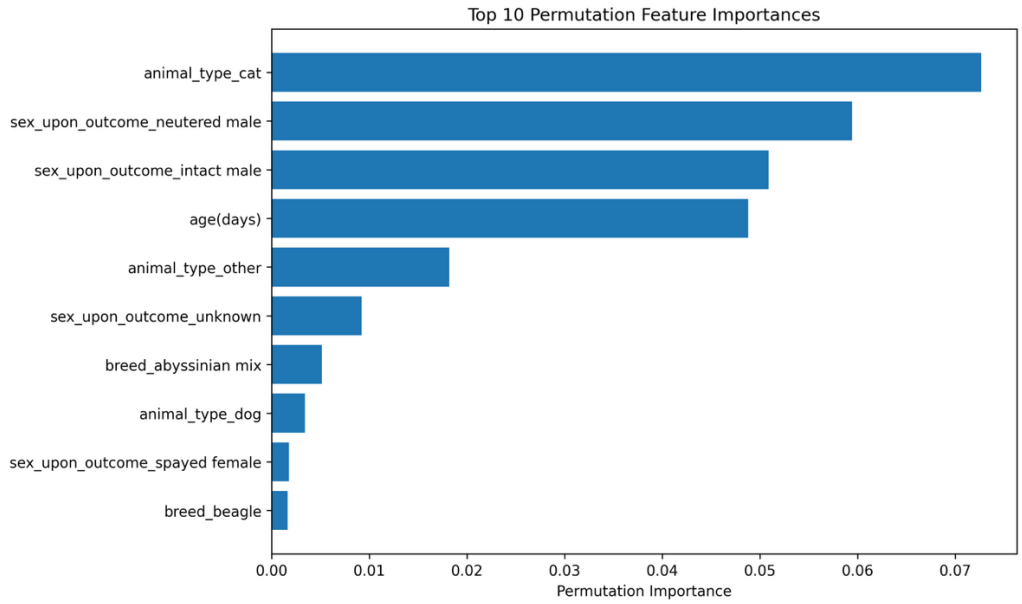
<Table 7: Score Results for Each Machine Learning Model>

The baseline F1 score I get is about 0.066. Based on the comparison, the result of all four Machine Learning models' F1 score are better than that of baseline. The XGboost model achieved the best performance among them with the F1 score 0.289.

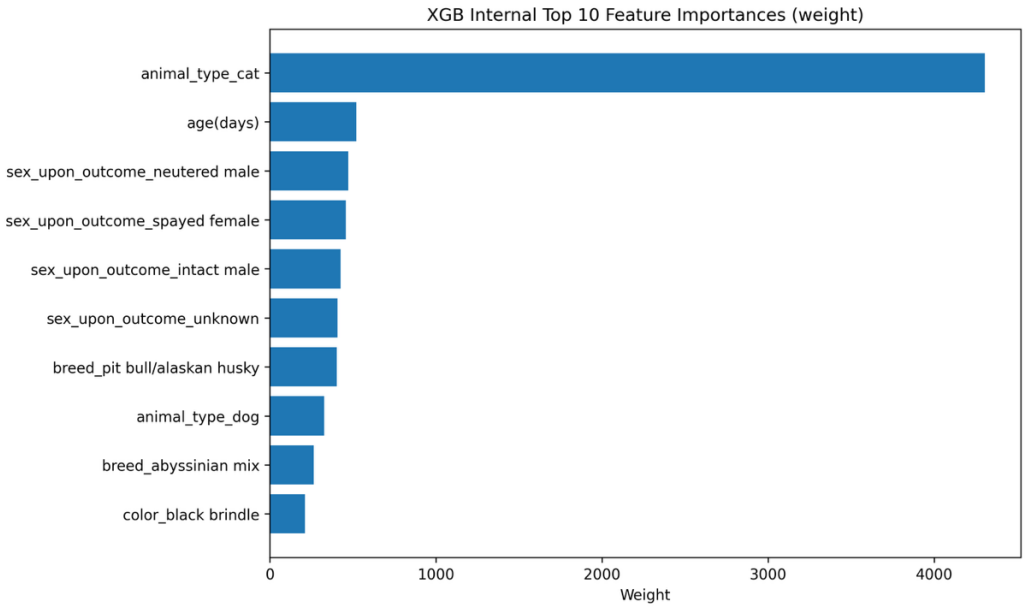


<Table 8: Test Set Confusion Matrix for Best Model XGB>

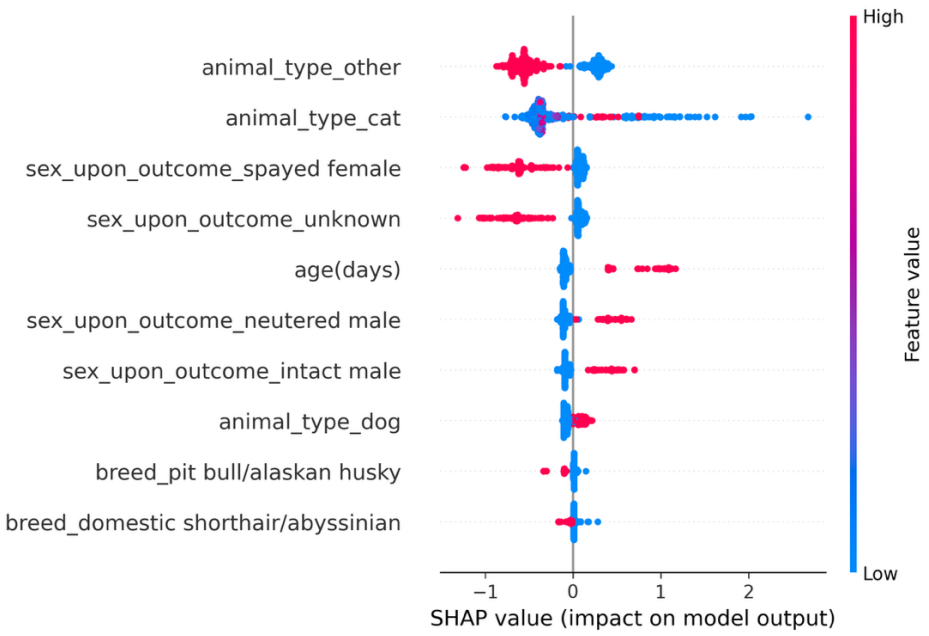
To inspect model behavior, I also created a confusion matrix for XGB, which shows the model make good prediction on the majority outcomes, like Adoption and Transfer. In contrast, the smaller classes, such as Died and Disposal, receive only a handful of correct predictions, which reflect sever class imbalance rather than model failure. Overall, the model is accurate for the classify classes but has limited predictive power for rare outcomes, which is expected when minority classes contain very little training data.



<Table 9: Top 10 Permutation Feature Importance using XGB>



<Table 10: XGB Internal Top 10 Feature Importance(weight)>



<Table 11: SHAP Summary Plot for XGB>

For interpretability, I use the permutation importance, which shows what features have the strongest impact on the XGB model's predictive performance when randomly shuffled. In this permutation plot, the animal_type_Cat is the clearly most influential feature and has the highest importance, meaning that shuffling this variable will cause the largest drop in predictive accuracy. The features neutered and intact males and age(days) also appear as the strongest predictors. Besides, the XGBoost internal feature importance are also used for better understanding what the model result wants to tell us. The XGBoost weight-based importance ranking broadly supports the permutation findings but also reveals some differences in how the model splits on features internally.

The SHAP summary plot provides deeper understanding by showing not only which features matter, but also how they influence predictions. For example, for the animal_type_other, we can get the information that having the animal type other than cats or dogs will strongly push the

prediction away from the positive class; for sex_upon_outcome_spayed female, we can have being a spayed female will also push the prediction toward to the negative class; and for age(days), the younger animals increase the prediction for the positive class. So, SHAP emphasizes feature importance and the directionality influence.

5. Outlook

To make improvement for next step, I plan to divide the sex_upon_outcome into two features, sex and sterilization condition, because I think with these two separated features, the prediction will be more accurate. For this time, I only use the half of my original dataset for the Machine Learning model and compare their Macro F1 score, because when I tried the whole dataset, it always ruins my computer, and I have to restart it and try everything again. So, for next step, I plan to solve the problem and use the whole original dataset to get a more detailed and ideal metric score for comparing machine learning models. Furthermore, I also plan to increase the number of folds in my cross validation, which can also help the result to be more accurate.

6. Reference

[1]Daoud, J. (2020). *Animal Shelter Analytics* [Dataset]. Kaggle. From https://www.kaggle.com/datasets/jackdaoud/animal-shelter-analytics

[2] Vocal For Pets. (N.A.). *Why kill shelters exist.* From https://vocalforpets.org/why-kill-shelters-exist/

[3]Kenneth L (2022, January 4). *Micro, Macro & Weighted Averages of F1 Score, Clearly Explained*. From https://towardsdatascience.com/micro-macro-weighted-averages-of-f1-score-clearly-explained-b603420b292f/

7. GitHub Repository

https://github.com/DQZ25/Animal_Adoption_Project.git