Ilene Tam and Javier Liu

Math 457 Introduction to Statistical Learning

Due: December 13th, 2021

Dataset Used: Shelter Animal Outcomes

Final Report

Introduction

In this dataset, we will be investigating data provided by the Austin Animal Center, which is one of the largest no-kill animal shelters in the US with over 18,000 animals in their shelter. Their mission revolves around protecting and caring for abandoned, at-risk, and surrendered animals. Especially because every year in the US there are approximately 7.6 million animals placed into these animal shelters. And of the 7.6 million animals, 2.6 million dogs and cats are unluckily euthanized due to several reasons. While some animals were euthanized due to behavioral issues, others had to be put down due to the overcrowding/overpopulation of shelters. These were statistics recorded from the early 2010s, since then new data has been collected in 2019 that says these numbers have been significantly reduced. In 2019, the number of animals entering the shelter was recorded at about 6.3 million and only 920,000 sheltered animals were euthanized. Given the significant decrease, from 2.6 million to 920,000, we can credit this to the increase of animal adoptions and returning of stray animals to their owners. We can further analyze this success rate and continue reducing the number of animals in shelters by understanding these adoption patterns. Thus, the following dataset from the Austin Animal Center was quite helpful. This collective dataset provides a wide range of information including the animal's name, date of birth, animal breed (either a dog or cat), color (animal's fur), sex (female/male), and outcome type (including adoption, transfers of ownership such as to other shelters/partners, return to owner or euthanasia), and age at the time of outcome.

The most obvious questions to be asked about the dataset pertain to the outcome of the animals; which animal is the most likely to be newly adopted, which feature is the most influential in being adopted, and other variables revolving around adoption successes or failures of each animal. Another interesting observation we would like to gather information on is which day of the week and time of the day do adoptions most frequently occur. These are all significant factors to consider for animal shelters to look into in order to decide where to allocate their

resources and effort towards. By doing so, it would help maximize the number of animals being adopted in the shelter and the probability of each animal being adopted. Lastly, we would like to test our hypothesis that cats and dogs have a different distribution of feature importance that affects their outcomes. Consequently knowing this information can influence the owner's decision. Another use of this information is for shelters when they are considering euthanizing their animals. That is only in the case that they have a very low likelihood of becoming adopted and are under severe financial hardship.

Methods

Given these multiple features, we extracted particular features in our dataset that were relevant to our research questions that would have influenced the outcomes of the adoptions for the animals, as well as removing strongly correlated noise variables such as "outcomeSubtype". To further investigate this dataset and answer our research questions, we proposed using the following methods to analyze the dataset: k-nearest neighbors (KNN), logistic regression, linear and quadratic discriminant analysis (LDA and QDA), support vector methods (SVM), random forests, boosting, and cross-validation. We were able to use logistic regression to model the categorical variables and identify their correlation with their outcomes. Additionally, random forest trees and boosting acted as an ensemble method to also rank the importance of extracted features in predicting the outcome of the animal. Lastly, we conducted a 70-30 training and testing split through cross-validation to refit our generated model, measure and improve the performance of our models, and establish empirical distribution functions for our data statistics. Model performance was measured using an accuracy score, which provided information on how many predictions were correct, and cross-entropy loss. Cross-entropy loss was used to penalize both types of errors, but especially the predictions which are confident and wrong.

Data Preparation

Before conducting any analysis, there were some challenges we encountered including the need to cleanse the dataset. The first step was to preprocess/clean up the data, like removing irrelevant features such as the animal's name, where some fields were left as an "unknown" and that was perceived as an actual name. This step also included removing data that isn't useful and can't provide any insight, such as the animal ID and outcomeSubtype field. In addition, this

dataset consisted of a majority of categorical variables, where many of the features required some data processing work. Here, we used label encoding and dummy/one-hot encoding to counter this issue. Furthermore, the age of animals during the outcome ("Age Upon Outcome") are scaled at different units. While some outcomes for the animals occurred at several years old, some took place when they were only a few months old or even to the extent of being a few weeks old upon adoption. Thus, there is a need to have a uniform unit for this variable, in our analyses, we calculated their age in terms of the months. Once we had full numerical data, we could finally train and test a classifier. Lastly, some more challenges that we had encountered when working with our training and test data was the support vector classification for the linear kernel. Any value larger than the default max_iter = 100 would not complete runtime, and with max iterations being too low, it led to a convergence error from the innate lbfgs solvers' algorithm provided by the package that we used, the Scikit-Learn Library. From discussing our challenges, the following are our output results and analysis that we came up with.

Results

Upon performing some analyses on the overall dataset, we found that the most frequent outcomes for dogs were adoption, return to owner, transfers; whereas, for cats, their most frequent outcomes were only adoption and transfers. Further analysis on random forest and boosting provided similar results in regards to the importance of features that shows that people have a slight preference upon adopting dogs over cats. And that age played a huge role in the dataset, that it is the most important feature people consider when they want to adopt an animal at the shelter. In addition to this, we split the dataset based on animal types; one being dogs and the other being cats to investigate whether there was a distribution difference between them. Our results show that there was a slight difference between them, especially when it came to the importance of features, there is a great impact on cats' adoption rates depending on whether they were given a name or not.

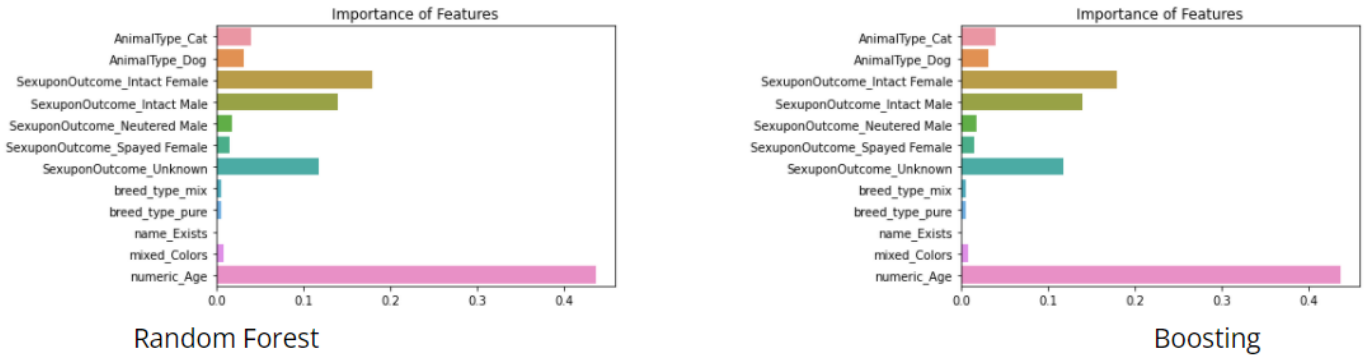

Figure 1: Importance of features for random forest and boosting classifiers, respectively

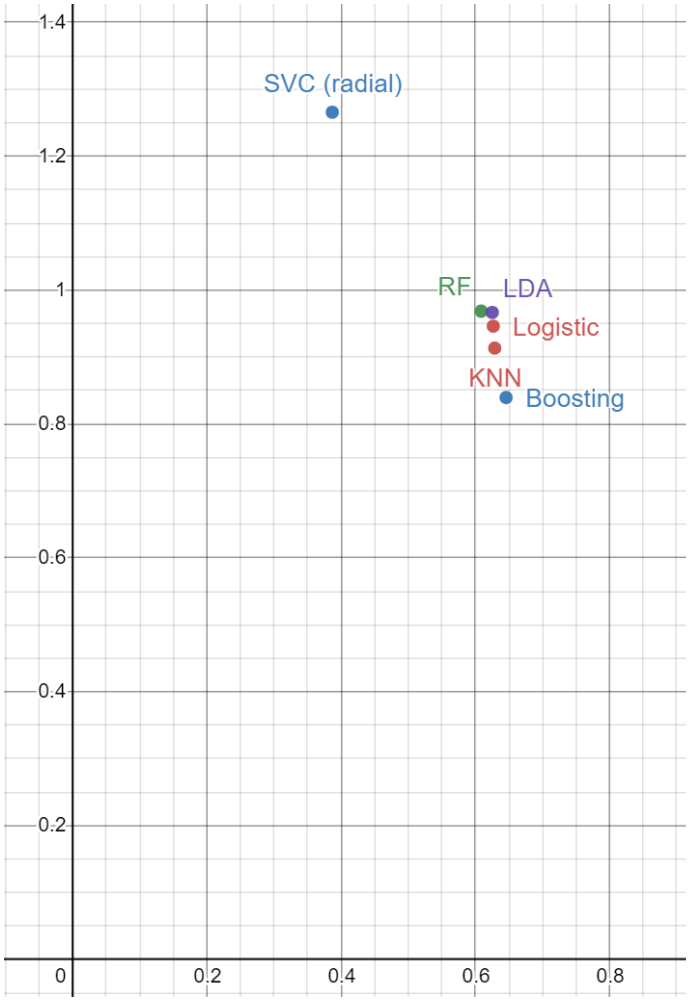

Figure 2: Model performance with accuracy and cross-entropy (log loss)

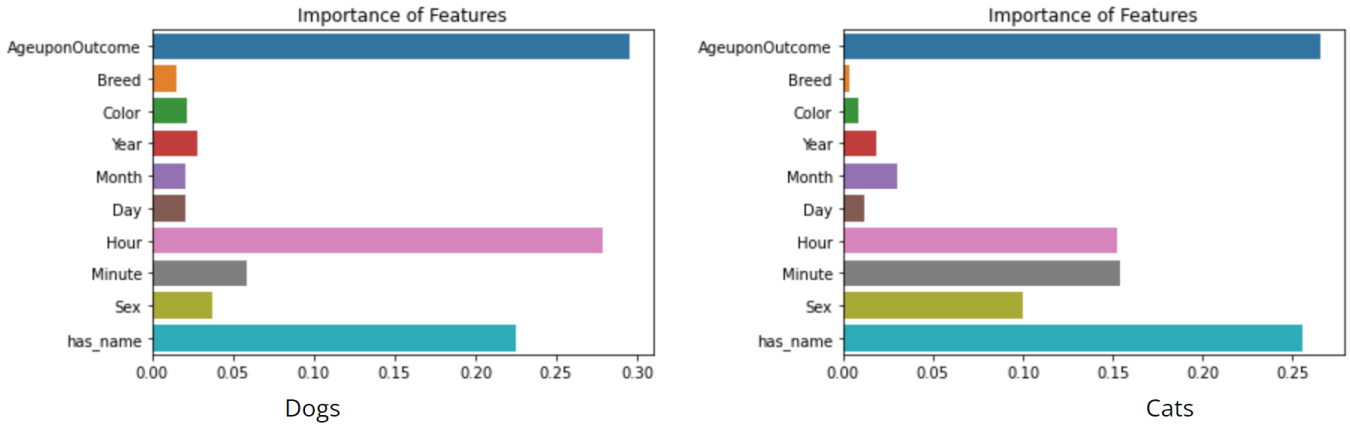

Figure 3: Random forest model performance for separately trained dogs and cats data

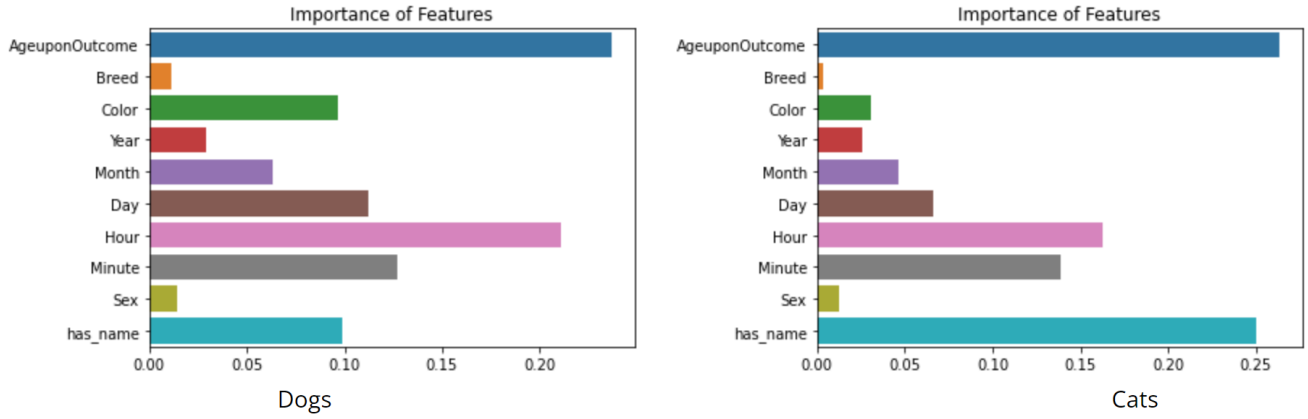

Figure 4: Boosting model performance for separately trained dogs and cats data

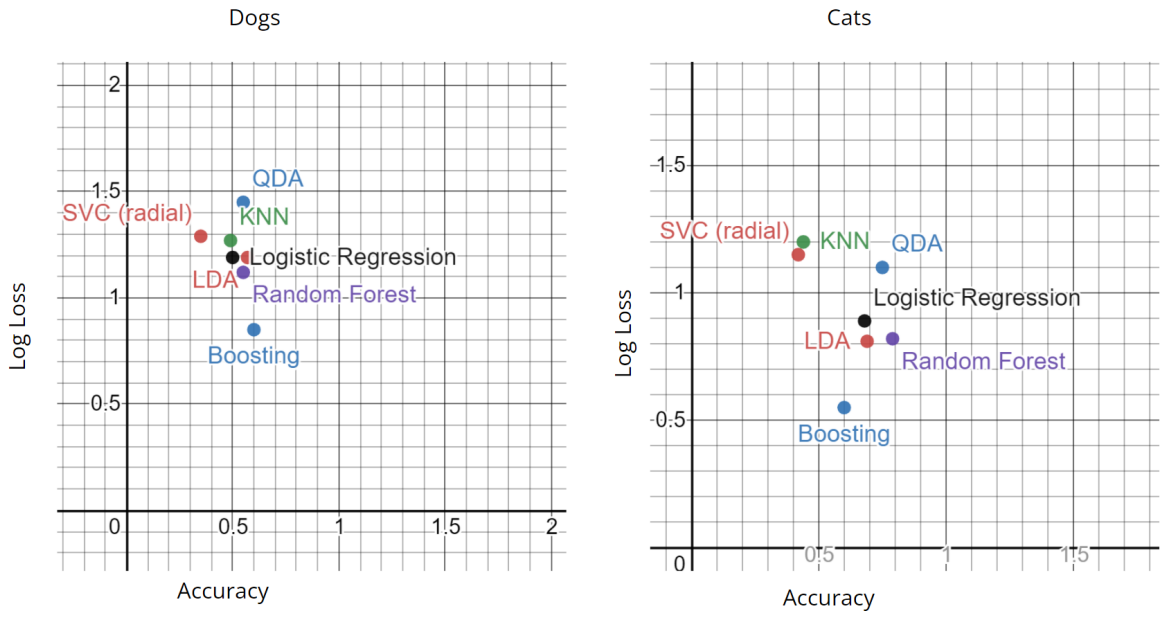

Figure 5: Model performance for separately trained dogs and cats data

Conclusions

Our overall analysis helped us better understand the adoption dataset and gain insight into the dataset to help the broader audience who would be interested in improving the animal shelter system and for those interested in adopting animals. This information and analysis are beneficial to animal shelters that want to maximize their efforts and resources in finding a home for every animal they have. Although it is unlikely that it can remove all animals in their shelter, it is definitely possible to reduce the number of animals in the shelter every year. In understanding what features are relevant to owners that are interested in adopting, shelters now know where to allocate their resources to, in order to help those that are less likely to be adopted and save them from being put into euthanasia. Although our analysis was quite limited, we were able to analyze and interpret relevant data that'd be useful for the animal shelters. With that being said, the most important features in determining their outcome are based on the type of animal, their sex, whether they have a name, if they are multi-colored, if their breed is mixed or pure, and their age. To conclude our research question, we found that dogs and cats are similarly adoptable/transferable in shelters; however, they do have a different distribution. Also, cats were found to have a higher prediction accuracy in this specific data set. Lastly, our findings show that while dogs' outcomes were more dependent on the color of their fur; cats' outcomes were more dependent on the existence of their name's. In future studies, there are many other possible tweaks to be made to improve the model. One is that there can be further research done on the features of animals such as looking into whether each specific breed of the animal, weight, age, and more has an impact on the animal being adopted. Another possible improvement would be that there are more ways to transform and interpret our categorical data into numerical data. By providing further research and analysis it provides more data for shelters to use for reference to make more decisions.