Final Project
Animal Shelter Survival

Astrid Fuentes, MS

Bellevue University

DSC 630: Predictive Analytics

Prof. Fadi Alsaleem

August 12th, 2021

Abstract

Every year, thousands of animals enter shelters across the United States either surrendered by owners or born and found in the streets. Some of these animals do not make it out of the shelter. They either die of a disease, are euthanized for different reasons, or transferred to other location. Some of them get adopted which is the happy ending I would like for all of them.

With a data set from the Austin Animal Center, this project is meant to identify the characteristics that contribute to a cat or dog surviving the shelter and being adopted.

Fortunately, most cats and dogs that entered the shelter between 2013 and 2016 did survive. However, I thought it was important to be able to predict the survival ahead of time and to be able to identify if those who did not survive had something in common. Using logistic regression, I have identified the top characteristics that contribute to the non-survival of dogs at the Austin Animal Center shelter. Pitbull mix dog breeds made it to the top of the chart which prompts some of my recommendations to improve their odds. Additionally, the logistic regression model built in this project can predict the survival of a cat or dog when they enter the shelter with an accuracy of over 90%. This can be used to advertise those animals with lower survival chances to help them survive the shelter. This can be done through different community events, adoption incentives, and other community programs.

In the next few pages, I will present a more detailed summary of my analysis, modelling technique, results, and recommendations with a more technical approach.

Introduction

For my project, I chose to work independently in an animal shelter data file that contains individual information about cats and dogs and the outcome of each.

Background

The data comes from the Austin Animal Center from October 1st, 2013 to March, 2016. All animals receive a unique Animal ID during intake. Outcomes represent the status of animals as they leave the Animal Center and include Adoption, Died, Euthanasia, Return to owner, and Transfer.

- Problem Statement

  Animals like cats and dogs frequently enter shelters across the US. I would like to use the Austin Animal Center dataset to build a model that will help us determine what characteristics in cats and dogs contribute to their survival in the shelter. If I can predict what animals are less expected to get adopted, I could make specific recommendations like advertising them more, bringing them to special adoption events, lowering their adoption fees, or creating other incentives to help them get adopted and survive the shelter.
- Scope

  The main objective of this project is to build a model based on determining characteristics of cats and dogs that will help predict whether the animal will survive the shelter or not. These characteristics might include age, color, health condition, and others. I would like to determine if there is a strong correlation between the variables collected in the data set and the outcome variable.

Methods

- Technical Approach

  Python and R have been used to clean-up and prepare the data and to perform preliminary analysis and modelling. Different statistics measures are being used to determine correlation and variable contribution to the model.

- Data sources or plan for data

  There are 3 files available in this data source, for the purpose of this project I have selected the train.csv file which contains the outcome variable. The test.csv file is preserved and might be used in the future for further prediction purposes.

  Data has been imported into Python and each variable has been analyzed. I have looked at things like minimum and maximum values, searched for nulls, duplicates, and outliers and handled them appropriately.

- Analysis

  With a clean data set, I began preliminary analysis by constructing different plots like histograms and boxplots that give us a better picture of our data and some initial insights. I then checked for correlation and covariance and then determine which model or models could be used for this type of data. At this stage I decided that a logistic regression model was the most appropriate to be able to determine survival rates.

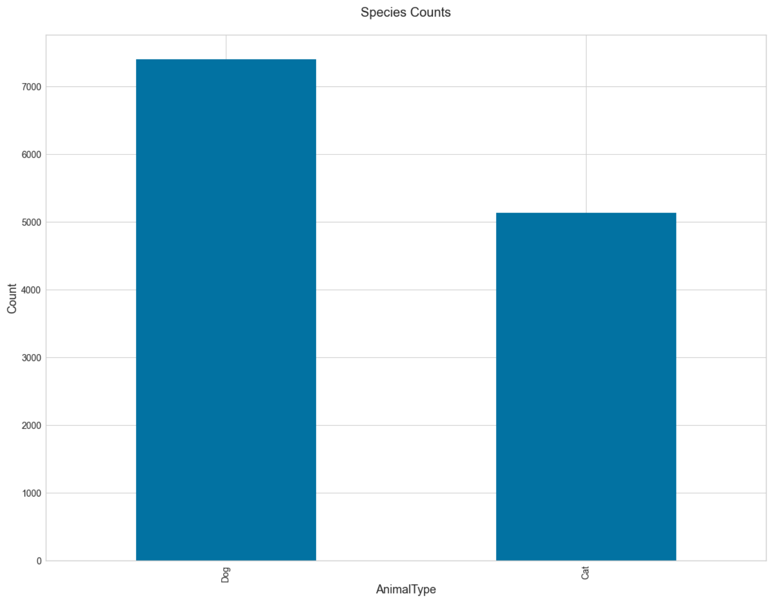
Initial Analysis

In my initial analysis, I limited the outcome variable to either Survived (1) or Died (0). I am grouping "Adoption", "Return to owner", and "Transfer" status into "Survived" = 1 and Died or Euthanasia into "Died" = 0.

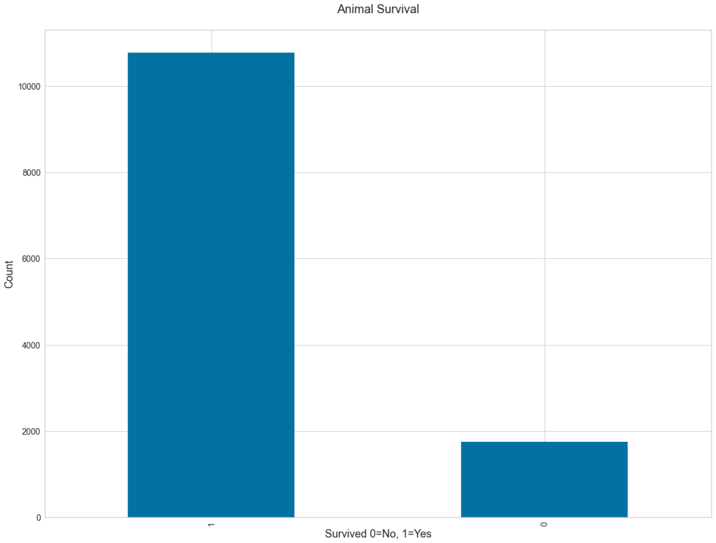
Additional variables are: Name, DateTime, OutcomeSubtype, AnimalType, SexuponOutcome, AgeuponOutcome, Breed, Color. Most of these variables are categorical with many different values, which makes initial analysis harder. For this project, I am grouping all animals with age of less than 1 year into "<1" and removing all "year" and "years" words from this column.

Some additional clean up includes checking for duplicates and null as well as renaming some variables for better handling and understanding. Null values for Age were filled in with a median age.

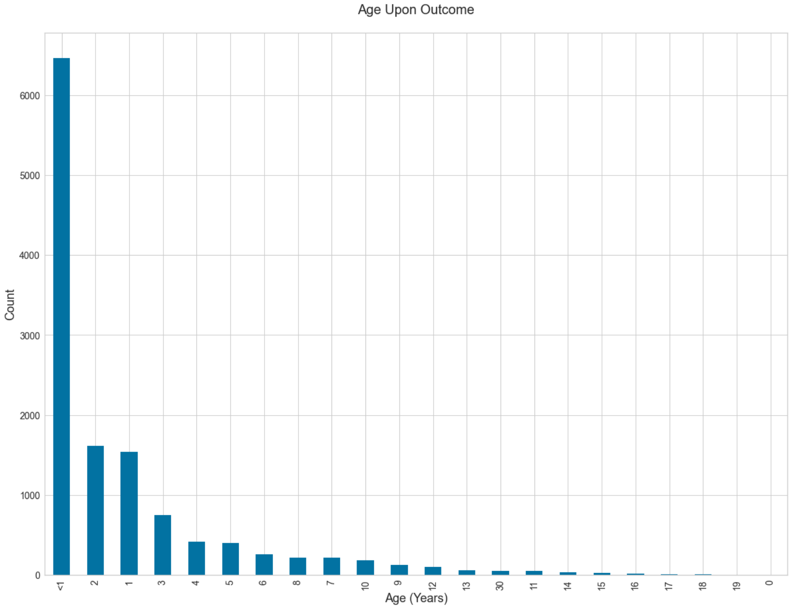
My initial analysis is based on exploration of the data. My data file contains 10 variables and 12521 records which represent one animal each. The animal type is divided in two major species: Cats and Dogs. There are 7392 dogs and 5129 cats in total.



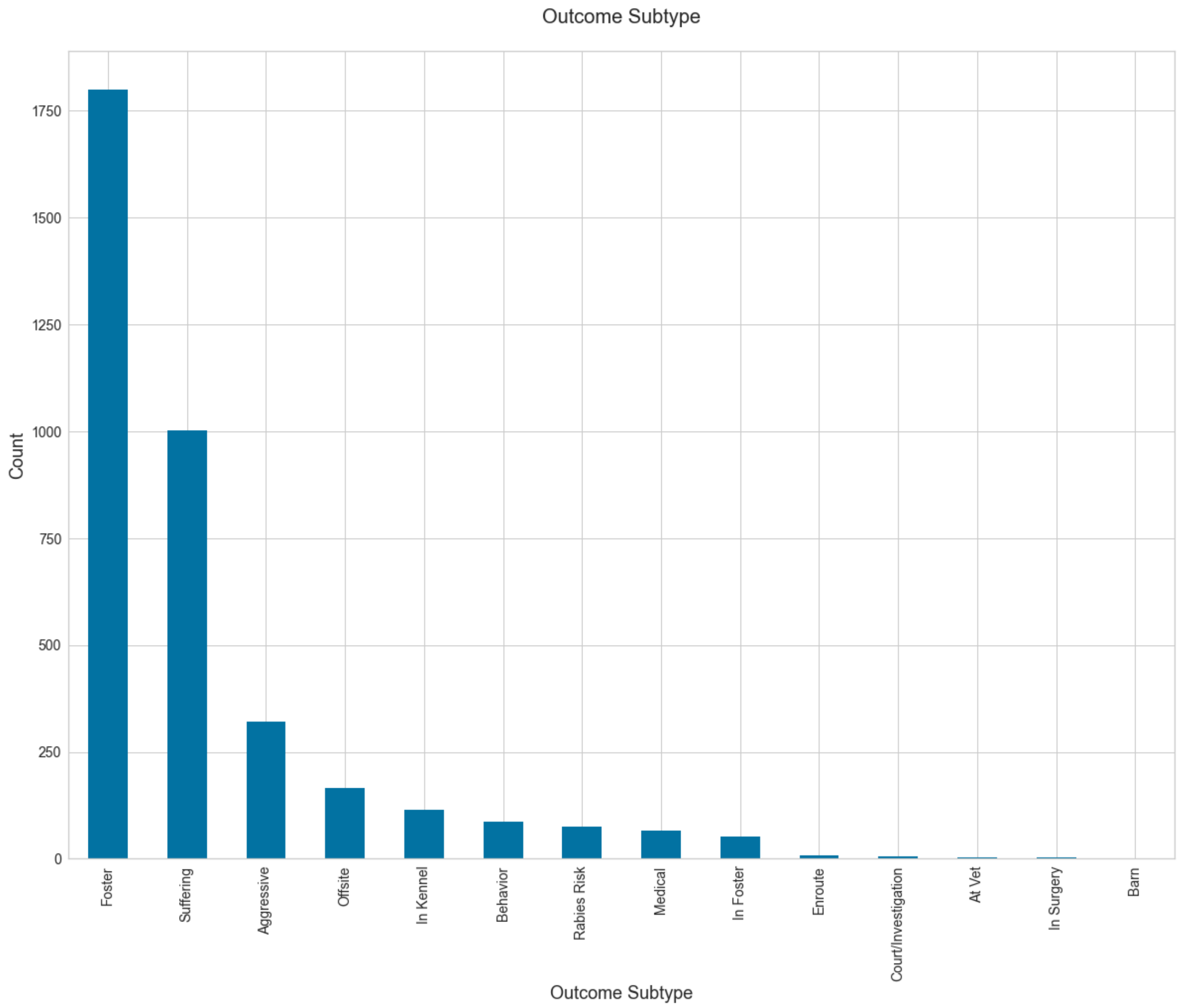
The outcome is also two groups: Survived (1) or Died (0). The number of animals who survived is considerably larger, 10769 animals versus 1752 animals that did not survive.



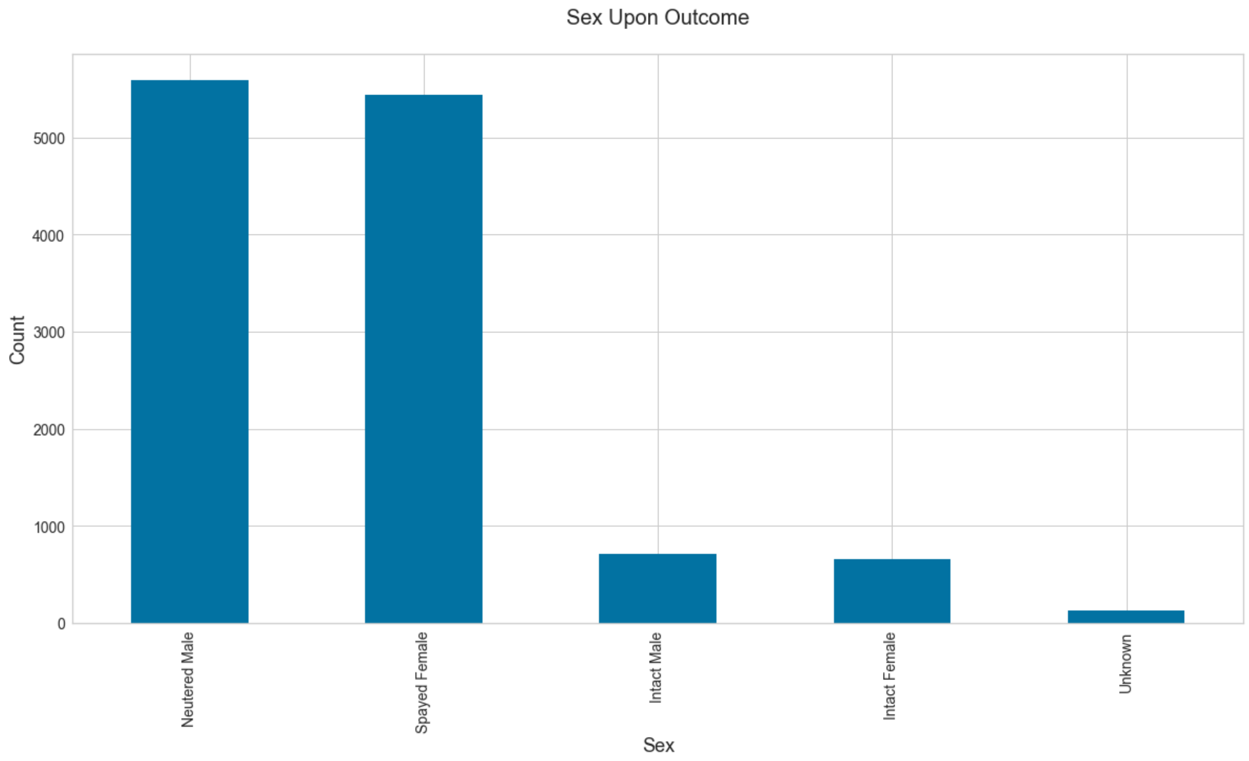
Most of the animals were under 1 year of age. This variable was recoded from the original file and contains values anywhere from 1 day to several weeks and up to 11 months. We can clearly see that older animals were less frequently seen at the shelter.



The Outcome Subtype variable contains a more specific reason for the outcome. The plot below combines survival and non-survival outcomes. Most of the animals who survived when to foster care while most of the animals who did not survive were euthanized due to suffering or aggressive behavior.



The plot below shows that most cats and dogs were neutered or spayed upon outcome. Under 1000 males and 1000 females were intact, most likely under the age/weight requirement for surgery.

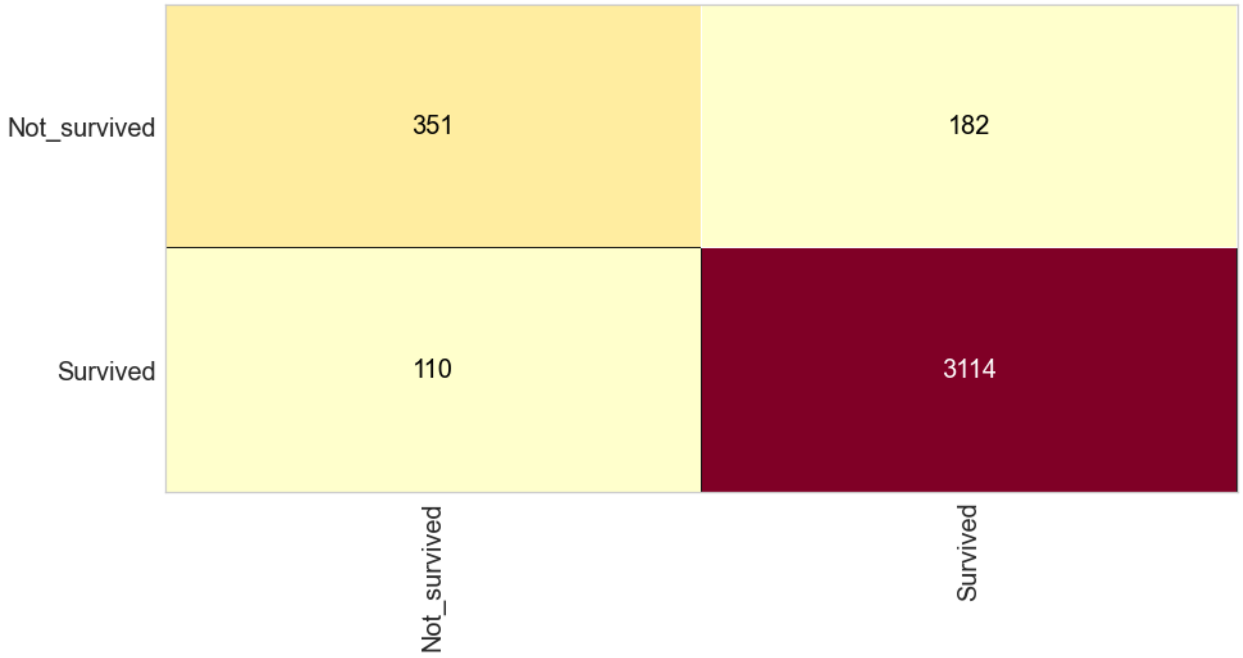


Results

My analysis of the data shows that 16% of the cats and dog that enter the shelter do not survive. Of those who do not survive, the majority are euthanized due to suffering or aggressive behavior.

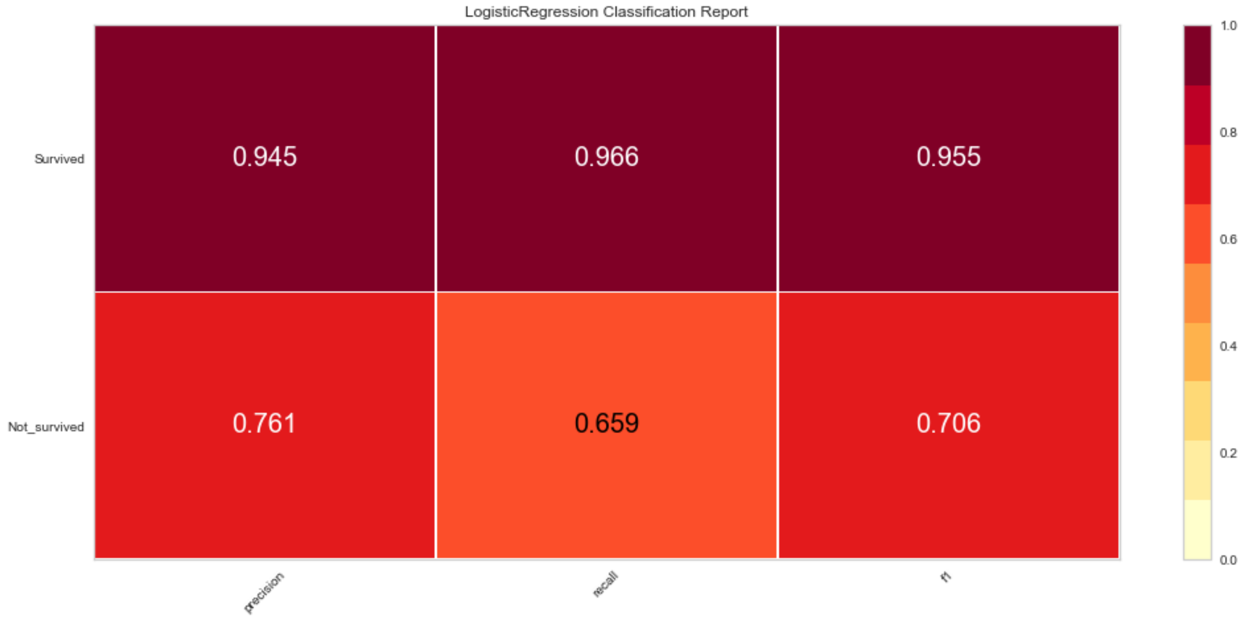
To create a logistic regression model, we converted some categorical variables to features. Animal Type was categorized as 1 for Cat or 2 for Dog. While Sex, Breed, and Color were converted to 1's and 0's in new features columns using pandas' get_dummies. These features were added to the data frame and used to model the outcome variables after splitting the dataset in train and test sections. The model chosen was a logistic regression classification model. The reason to choose this model is because the outcome variable is either survived or not, hence a logistic model makes sense. Classification was needed because we had several classifying variables or features.

After training our testing our model we built the below confusion matrix:

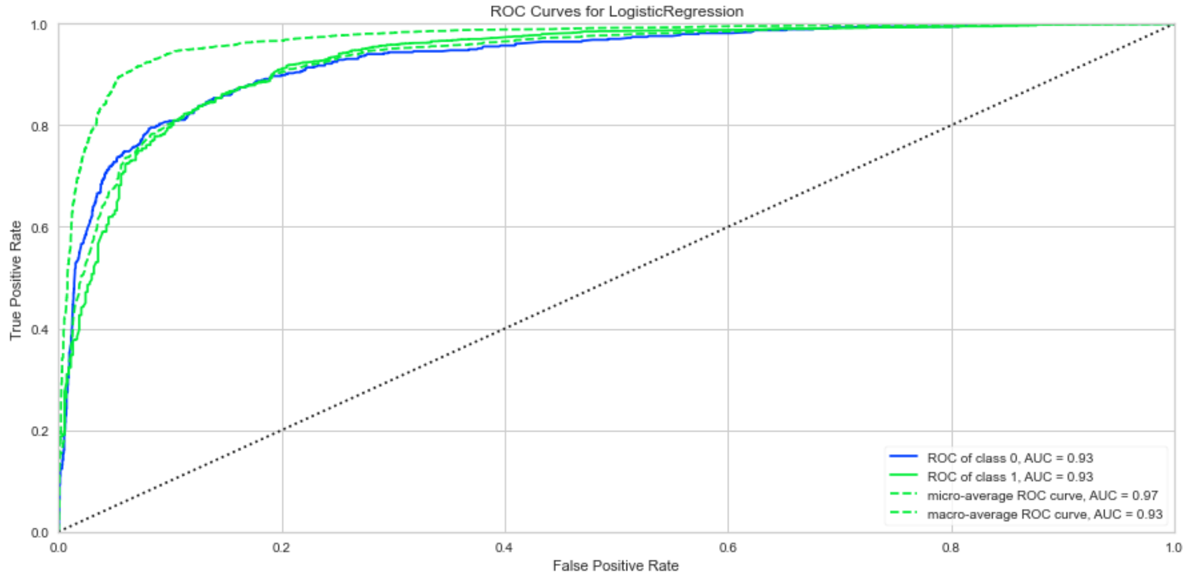


Based on the above confusion matrix, our model correctly predicted the survival of 3114 animals and the non-survival of 351 animals. The model incorrectly predicted the outcome of 292 animals. This model has an accuracy of 92.2% which is very good.

Model report:



ROC and AUC:

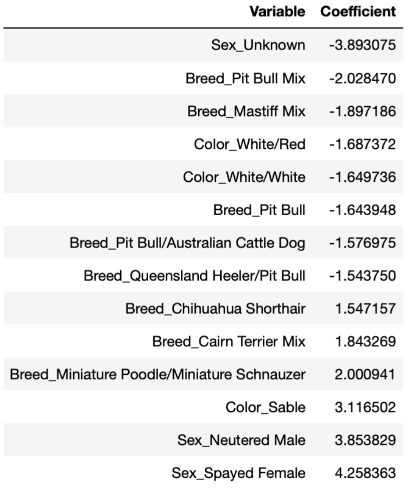


Based on the above, the model performed better at predicting the survivals with an accuracy of 94.5% while the non-survivals were predicted with a precision of 76.1%.

Our model was built to predict Outcome Type (survived or not) using Age and the categorical dummy variables of Sex, Animal Type, Breed, and Color. Previously, we only used the categorical variables of Sex and Animal Type. After introducing the Breed and color, the overall precision of the model increased from 92.1% to 92.2% which is not a huge improvement.

Now it is important to find out which coefficients contribute more to the outcome variable. Because this is a logistic regression mode and the outcome variable is 1 = Survive, 0 = Not survive, we can assume that a positive coefficient increases changes of survival while a negative coefficient decreases it. Also, taking the absolute value of the coefficient we can determine which characteristics have a larger contribution to the model.

After slicing out coefficients data frame by selecting only those coefficients with an absolute value greater than 1.5 we got the below 14 characteristics.



Variable	Coefficient
Sex_Unknown	-3.893075
Breed_Pit Bull Mix	-2.028470
Breed_Mastiff Mix	-1.897186
Color_White/Red	-1.687372
Color_White/White	-1.649736
Breed_Pit Bull	-1.643948
Breed_Pit Bull/Australian Cattle Dog	-1.576975
Breed_Queensland Heeler/Pit Bull	-1.543750
Breed_Chihuahua Shorthair	1.547157
Breed_Cairn Terrier Mix	1.843269
Breed_Miniature Poodle/Miniature Schnauzer	2.000941
Color_Sable	3.116502
Sex_Neutered Male	3.853829
Sex_Spayed Female	4.258363

Upon further inspection of the coefficients of the regression model, we found that:

- The Breed is a variable that highly contributes to the model. We can see that several variables related to Pit Bull breeds contribute negatively to the survival of the animal.
- Miniature Poodle and Miniature Schnauzer breeds as well as animals with Sable color have more chances of survival based on the variable coefficient positively contributing to the model.
- Animals color Sable as well as the Neutered/Spayed variables might only be contributing to the model because they are seen more frequently in the data set. This might need further analysis. For example, all animals that leave the shelter must be spayed or neutered. It might be wise to remove these characteristics from the model and blend them with just Sex being Female, Male or Unknown. Animals of color red and white also showed up as top contributors, we would need to further analyze these.

Discussion and Conclusion

The logistic Regression model built can successfully predict the survival of a cat or dog that enters the shelter with an accuracy of over 90%. It also helps us determine some of the main characteristics that contribute to this outcome, being breed one of the main.

Pitbull mix breed of dogs has been identified as a high contributor to non-survival at the shelter. Giving special attention to these animals could improve their survival rate by giving them better changes of being adopted. Some of my recommendations include:

- Additional advertisement campaigns that promote adoption for Pittbul Mix breeds at half the regular adoption rate.
- To provide Spay/Neuter surgery for free for any Pitbull Mix dog owners in the community. This will help avoid over population of this breed.
- To provide or improve low-cost veterinary attention in the community.
- To provide low-cost behavioral therapy after adoption. This would help adopters bring their dogs back regularly to work on their training and behavior concerns preventing the animal from being returned/surrender at the shelter.
- Educate the community about Pitbull Mix breeds to prevent fear due to bad reputation of this breed.

I would like to continue my analysis to identify further characteristics that help determine whether a cat or dog survives the shelter. Currently, sickness is not in the top contributors to non-survival list, however, I suspect this needs further research.

Additionally, I think it would be helpful to create separate models for cats are dogs. With breed being such an important characteristic for a dog survival rate, the model might not be performing as well for cats.

Acknowledgments

First and foremost, I would like to thank our professor Fadi Alsaleem who guided me in doing this project. He provided me with helpful instructions and feedback throughout the course and in each milestone.

I would also like to thank my classmates for providing wonderful motivation, interesting and relevant discussions, and great advice for this project and other class topics.

References

https://www.kaggle.com/c/shelter-animal-outcomes/data?select=train.csv.gz

Eberly College of Science. Penn State University. *Logistic Regression*.
https://online.stat.psu.edu/stat501/lesson/15/15.1

Li Susan, September 28, 2017. *Building A Logistic Regression in Python, Step by Step*. Towards Data Science. https://towardsdatascience.com/building-a-logistic-regression-in-python-step-by-step-becd4d56c9c8