Data Science Final Project Proposal

Sabrina Riebe

Every day there are many animals given up as unwanted to U.S. shelters, while other animals are rescued from cruelty situations, or are found wandering the streets. This results in approximately 7.6 million pets and companions ending up in U.S. shelters each year. Many of the animals in the shelters are adopted and find loving families to take them home, however, many are not so lucky. Those that are not adopted cannot stay in the shelter forever and are euthanized. 2.7 Million cats and dogs are euthanized every year in the U.S.

The goal of this project is to create a model to understand trends in animal outcomes. These insights could benefit shelters by helping them focus their attention on those animals that are predicted to need a little extra help finding a new home.

The **purpose** of this project is to predict whether animals will be adopted or not.

The data for this project is from Austin Animal Center from October 1st, 2013 to March 2016:

- Starting with 10,000 animal profiles
    - Approximately 450 different breeds of animals
    - Approximately 5800 dogs
    - Approximately 4100 cats
    - Approximately 4000 female cats and dogs
    - Approximately 4500 males cats and dogs
- The dataset is split into a train and test dataset

The dataset includes the following information about the animals:

- Anima lD
- Name
- Date & Time
- Outcome Type
    - *Adopted*
    - *Returned to owner*
    - *Transferred (to another shelter)*
    - *Euthanized*
    - *Died*
- Outcome Subtype: additional information on the outcome
- Animal Type (cat or dog)
- Sex upon Outcome: gender and whether the animal has been spayed/neutered, or not.
- Age upon Outcome
- Breed
- Color

*All data reflects the status of the animal at the time of **leaving** the shelter

I will further extract information on whether the animal is spayed or neutered, or not, the date that the animal left the shelter, a numeric value for the animals age upon leaving the shelter, and also whether the animal is purebred or a mixed breed. Furthermore, I will use the ensemble method to create new variables from the original data which I will then use in a final prediction model.

By analyzing the information, I will predict which animals are more likely to be adopted, euthanized, returned to their owners, or transferred to a different shelter.

The proposed approach to the project:

In order to achieve this goal, I will create a balanced prediction model (python backend) which will use multiclass classification to predict the various animals' outcomes based on the provided information. I will firstly clean the data and then create a prediction model with five algorithms – logistic regression, decision tree, neural network, XGBoost, and random forest. I will analyze each model to determine whether the necessary data is included in the output, and thereafter deploy the best performing algorithm and score it against the test prepared dataset (which will in the interim be prepared to meet the requirements of the new columns).

What I will NOT predict:

I will not focus on predicting which animals are likely to die firstly due to the small amount of data provided on this (only 65 records in 3 years), secondly because the original dataset shows that it was only the very young and/or very old animals that died, and lastly, because the focus of this project is on getting animals adopted.

I will not be focusing on the animals' breed as a main factor for determining whether they were adopted or not. I will instead be focusing on the animals' age, gender, sterilized status, month of adoption, and whether or not the animal was a purebred.