Gradient Tree Boosting + Support Vector Machines for Animal Shelter Outcomes

Sabirah Shuaybi, Mount Holyoke College 2019

Abstract + Motivation

According to pet statistics in the US, every year, approximately 7.6 millions animals end up in shelters. Mostly abandoned by owners or rescued from cruelty.

Motivation: Implementing **classification** models of **XGB** and **SVM** to predict the outcome of an animal given their various attributes and conducting **variable importance analysis** will provide shelters with insight into animal shelter trends and enable them to focus their energy on which kinds of animals need extra help to find a new home.

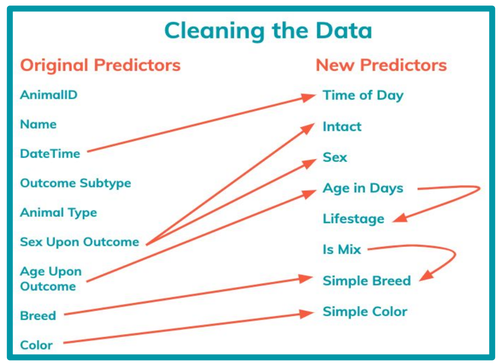
Data Overview

26729 observations

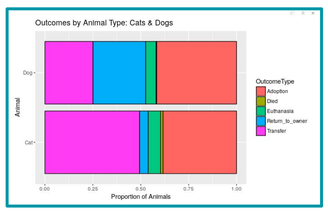
Variables: **Animal ID, Name, Date and Time, Outcome Type, Outcome Subtype, Animal Type, Sex upon Outcome, Age upon Outcome, Breed**, and **Color**

Outcome → response variable. Five animal outcomes: **adoption, return to owner, transfer, died, euthanasia**

Cleaning the Data

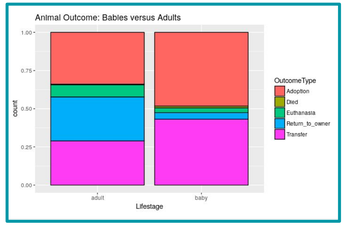


Data Exploration



To assess which predictors are most relevant in predicting animal outcome

Dogs more likely to be returned to owner than cats. Baby animals more likely to be adopted than adult. Are also more likely to be transferred and to have died.



Gradient Tree Boosting

An **ensemble** method; predictors selected *sequentially*

Conversion of **weak learners** → **strong learners**

After evaluating the first tree, we increase weights of observations that are difficult to classify and lower **weights** for those that are easy to classify

Each new tree = modified

New predictors are learning from mistakes committed by previous predictors → it takes **less time/iterations** to get close to final predictions

- ✓ AnimalType
- ✓ Lifestage
- ✓ IsMix
- ✓ Sex
- ✓ SimpleColor
- ✓ SimpleBreed

Predictors in Final XGB Model

Experimentation with Tuning Parameters

Nrounds → # of component models
Tried: 5, 10, 20 to 100

Learning rate → if too high, can result in overfitting.
Tried: 0.5, 0.6, and 0,7

Subsample → prop of obs used to grow each tree.
Tried: 0.5, 0.9, 1.

Max depth → deeper trees more complex, shorter pref.
Tried: values from 1 to 5.

Results and Test Set Performance

Predictions of final ensemble model = **weighted sum** of predictions made by previous tree models.

After experimenting with parameters, model with highest cross validation accuracy selected:

nrounds	20
max_depth	4
eta	0.6
gamma	0
colsample_bytree	1
min_child_weight	1
subsample	0.9

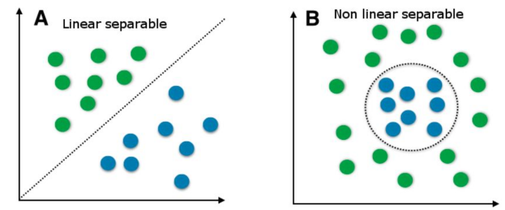
Test Set Error Rate: 0.378

Support Vector Machines

A classification technique that generates **hyperplanes** to **separate** and **classify** data into different regions

Given labeled training data (supervised learning), outputs an optimal hyperplane which categorizes new points

Computational benefits of using **kernels** vs physically enlarging feature space



Support Vector Machines can do both, **linear** and **non-linear** classification

Method and Implementation

Radial Kernel →

$$K(x_i, x_{i'}) = exp\left(-\gamma \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2\right)$$

Training observations far from x* play ~ no role in the predicted class label for x*. Thus, radial kernel has very **local behavior**; only nearby training observations have an effect on class label of a test observation.

- Used '*e1071*' pkg to train SVM on animal data
- Outcome ~ AnimalType + Lifestage + IsMix + SimpleBreed + SexuponOutcome + SimpleColor
- **One-vs-One** approach used for multiclass-classification with k > 2 levels (k = 5)

```
Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  radial
       cost:  1
      gamma:  0.00390625

Number of Support Vectors:  15175

 ( 3829 1244 5446 4498 158 )


Number of Classes:  5
```

Results

Test Set Error Rate: 0.387

Variable Importance Analysis (for final XGB model)

Why? Provides transparency and insight to modeling process by revealing which variables were most **influential** in predicting animals' outcome.

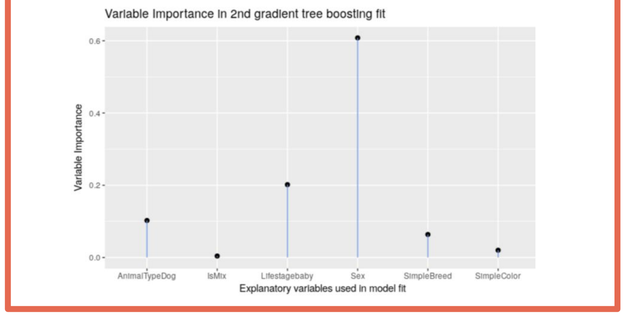
Because motivation for this project is to to identify which areas animal shelters can focus on to ensure better outcomes for these animals, conducting this type of variable importance analysis will lead one step closer to this goal.

How is Variable Importance Measured?

Package → caret::**VarImp**()

Sums up prediction accuracy over each boosting iteration. Uses three factors to measure a variables importance in the model.

1. **Gain** - improvement in accuracy achieved by feature to branches it's on
2. **Cover** - relative quantity of observations concerned by a feature
3. **Frequency** - # of times a feature is used in all generated trees



Discussion: Limitations and Future Work

Explore a **polynomial kernel** with a positive degree for a more **flexible decision boundary**. Assess if this results in a better svm fit.

Conduct further analysis as to which *levels* of these variables were most influential. In other words, establishing that the variable Life Stage is influential does not in itself indicate which of the **levels** relates to a certain outcomes. Are baby animals more likely to be adopted? Or adults? Are adults more likely to be euthanized vs babies? This is a **limitation** of the current variable analysis and would be an area of future enhancement.

How? Select one level out of the variables deemed important and isolate all data points equal to that level and evaluate/predict the outcome for that level to see if there was an observed difference in shelter outcome.Ideally, would repeat this process over all the important predictors to acquire a more **meaningful** and in depth understanding of the **distribution of importance** across the various levels of the predictors.