PetFinder: Predicting Adoption Speed of Pets

Jinyi Liu

Business Analytics & Statistics

December 6, 2023

THE UNIVERSITY OF
TENNESSEE
KNOXVILLE

Outline

Outline

Motivation

1

Outline

Problem Statement

Aim

The training data contains 14993 pet profiles, each with 23 features (we don't use the description). For each pet, the outcome is a number from 0 to 4, indicating how quickly a pet is adopted. The task is to predict this number for each pet in the test set.

Problem Statement

Evaluation

An $\mathrm{N} \times \mathrm{N}$ histogram matrix O is constructed, such that $O_{i,j}$ corresponds to the number of adoption records that have a rating of i (actual) and received a predicted rating j.

An $\mathrm{N} \times \mathrm{N}$ matrix of weights, w, is calculated based on the difference between actual and predicted rating scores:

$$w_{i,j} = \frac{(i-j)^2}{(N-1)^2}.$$

Problem Statement

Evaluation

An $\mathrm{N} \times \mathrm{N}$ histogram matrix of expected ratings, E, is calculated as the outer product between the actual rating's histogram vector of ratings and the predicted rating's histogram vector of ratings, normalized such that E and O have the same sum. From these three matrices, the quadratic weighted kappa is calculated as:

$$\kappa = 1 - \frac{\sum_{i,j} w_{i,j} O_{i,j}}{\sum_{i,j} w_{i,j} E_{i,j}}.$$

What we want is a κ score as close to 1 as possible.

Outline

Experiment and Results

Model

Since the adoption speed to be predicted is categorical, we adopt a xgboost model. This model is a gradient boosting algorithm that is optimized for speed and performance. It is based on decision trees and is a popular choice for machine learning competitions.

Experiment and Results

Data Preprocessing

Since we adopt the Xgboost method, the missing value is not a problem as Xgboost can handle it. We combine the breed and color features into one feature, and add one more column which combines pet's type, gender, vaccinated, dewormed and sterilized status. Furthermore, there are some pets without names, so we add a column to indicate whether the pet has a name or not, and also the length of the name. Besides, the breed of the pet is also an important factor. We add some columns about the breed. For brevity, we leave the details in the code and the report.

Experiment and Results

Feature Engineering

To better train the model, we do a feature engineering. We briefly introduce the features we add in this section and leave the details in the appendix code. We add the following features: Color, Breed, State, Hard-Interaction, State-BreedFull-ColorFull, Name, RescuerID, etc. For each one, we calculate the mean, minimum, maximum and standard deviation. After that we drop the mentioned features and only keep the new features since they are highly correlated. After this, we get a new 136 numerical features.

Experiment and Results

Model Training — Classifier

Intuitively, we think the Xgboost classifier is a good choice for this problem since the predicted variable has 5 values. We use the Xgboost classifier to train the model. To find the best hyperparameters, we do a grid search and find the best hyperparameters. This model gives us an average of $\kappa = 0.40$ which is not good enough.

Experiment and Results

Model Training — Regressor

Since the Xgboost classifier does not give us a good result, we try the Xgboost regressor. We do a new grid search for the best hyper-parameters. In this case, we choose the best hyperparameters as follows:

- max_depth: 7
- subsample: 0.8
- colsample_bytree: 0.8
- α: 0.05
- η: 0.01

This gives us an average of $\kappa = 0.4694$ with a standard error 0.01058 for a 5-fold crossvalidation, which is better than the Xg-boost classifier.

Outline

Discussion and Implication

Future Direction

We do not use the image and the metadata in this project. We think the image and the metadata are important for the adoption speed. However, we do not have enough time to do the image analysis and the metadata analysis. We think the image analysis is a good direction for future research. Also, there is also a lot of information in the description of the pet. We think the description is also a good direction for future research. Though we don't use the image and the metadata, we think the result is good enough. A κ of 0.4694 is not bad since we only use 176 features to predict the adoption speed.

Discussion and Implication

Feature Importance

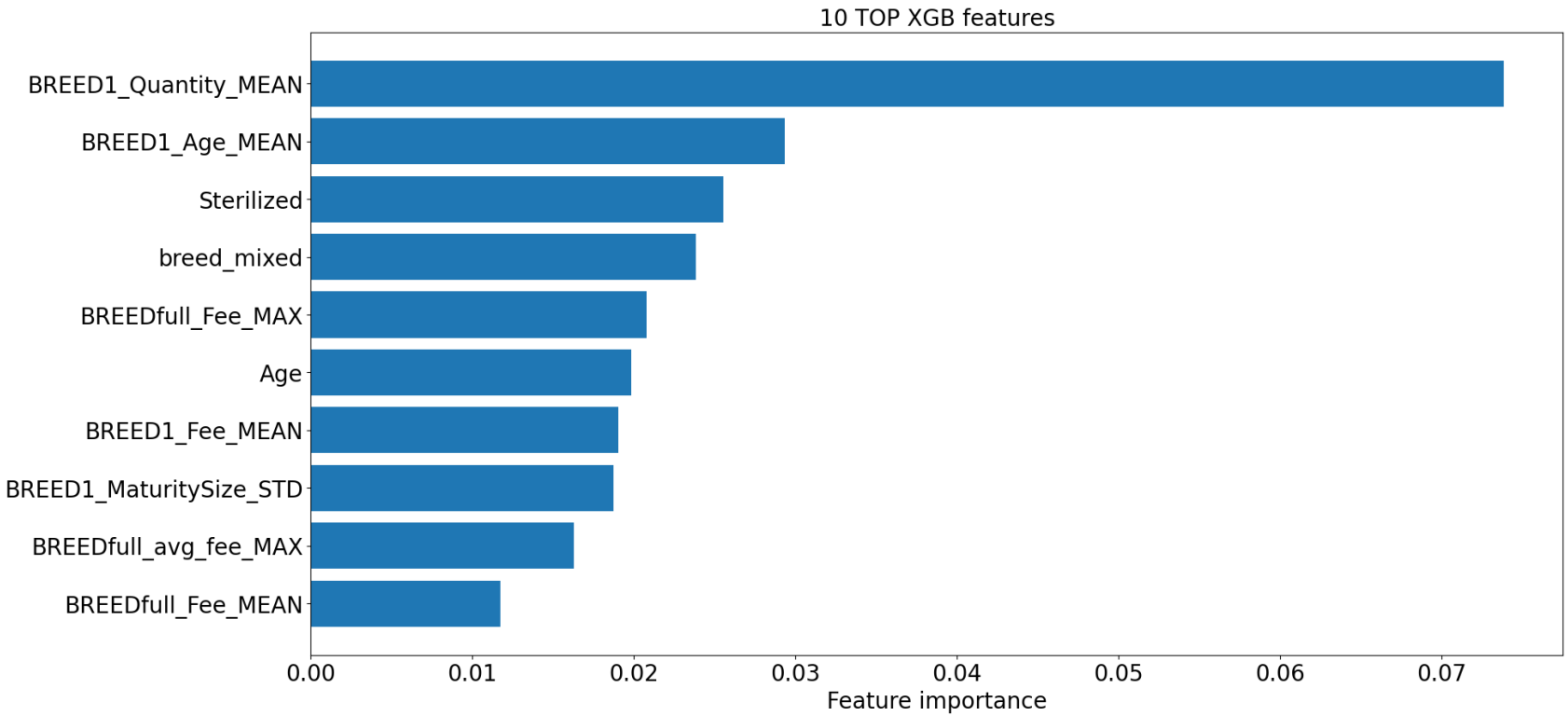

Figure: Top 10 XGB Features

Discussion and Implication

Implication

We can see that, the main breed, i.e., Breed1, is the most important factor affecting the adoption speed. Also, whether is sterilized, whether is breed is mixed and the age of the pets are also important factors. This is intuitive since in reality, people tend to adopt pets with a younger age. Also, the sterilized status is also important since people tend to adopt pets that are sterilized. The color of the pet is not important. This is reasonable since the color of the pet is not a good indicator of the adoption speed. We can also see that, the adoption fee also matters.

Discussion and Implication

Implication

Thus, to improve the adoption speed, we can do the following things:

- Highlight the main breed of the pet.
- Show mixed breed.
- Sterilize the pet.
- Highlight the age of the pet.
- Set a reasonable adoption fee.