Austin Animal Shelter Outcome Prediction

John Tazioli

14 June 2022

Abstract

The goal of this project was to use a classification model to predict whether new dogs admitted to the Austin Animal Shelter in Texas were more likely to be adopted or euthanized. This would allow the shelter to allocate limited resources towards overcoming euthaniasia trends or prioritizing the dogs more likely to be adopted. The data is publicly available on kaggle.com. The data was already clean but feature engineering was required to make it usable. A logistic regression was used as a preliminary model which was then exceeded by a Naive Bayes ensemble.

Design

The project idea originates from my love of dogs and desire to help shelter dogs find homes. The target variable the outcome with only two options; euthanasia or adoption. The feature variables are a mix of categorical and continuous values. The categorical features required dummy variables for the logistic regression and the features had to be split between a Guassian and a Bernoulli Naive Bayes put in an ensemble with a voting classifier.

Data

The dataset contains over 72,000 rows, each row being one animal taken into the shelter and it's outcome between October 2013 and October 2018. After filtering out all the cats and other wildlife, and the outcomes other than euthanasia or adoption, the dataset has 21,989 rows. There are 1411 euthanasia outcomes and 20578 adoption outcomes. In order to resolve the class imbalance, I oversampled the euthanasia outcomes at a ratio of 6:1 to make the minority class about 30% of the observations. This resulted in the training data set being 20507 observations with 14369 adoptions and 6138 euthanasia. The test data set was 6209 observations with 388 euthanasia outcomes.

Algorithms

0.1 Feature Engineering

1. The original dataset had over 40 columns and 70,000 observations. The redundant columns and unneeded observations were deleted in excel. The data was than exported as a CSV file and imported into a PANDAS dataframe in a Jupyter Notebook.

2. The shelter categorized the dogs into over 1400 breeds. This was too numerous to be useful as a feature so instead I calculated the euthanasia rate for each breed and used that as a feature.

3. The intake status recorded the condition of the dogs as they entered the shelter. This included categories like injured, normal, pregnant, or other. In order to be used in a Naive Bayes or a Logistic Regression model, these were converted to dummy variables.

4. The target variable is categorical with two options: 'Adoption' or 'Euthanasia'. I mapped 'Adoption' to 1 and 'Euthanasia' to 0.

5. The features were scaled in order to use the logistic regression model.

0.2 Models

For the Minimum Viable Product, a logistic regression model was used. This was improved upon with a Naive Bayes ensemble. A Gaussian Naive Bayes used the continuous features and the Bernoulli used the binary features and a voting classifier averaged the predictions.

0.3 Model Evaluation and Selection

The models were evaluated using F1 score. Accuracy and confusion matrices also provided insight.

- LogReg Baseline: F1 = 0.96
- LogReg Improvement: F1 = 0.97
- Bernoulli NB: F1 = 0.95
- Gaussian NB: F1 = 0.94
- Voting Classifier: F1 = 0.96

Tools

- Numpy and PANDAS for data manipulation
- Scikit-learn and imbalanced-learn for modeling
- Matplotlib and seaborn for plotting
- LaTex used for report writing.

Communication

A five minute power point presentation and the code used will be hosted on my github.