Machine Learning Engineer Nanodegree

Capstone Proposal

Ben Griffith
August 6, 2019

Domain Background

Every year, approximately 7.6 million companion animals end up in US shelters. Many animals are given up as unwanted by their owners, while others are picked up after getting lost or taken out of cruelty situations. Many of these animals find forever families to take them home, but just as many are not so lucky. 2.7 million dogs and cats are euthanized in the US every year.

My personal motivation for investigating a particular problem in this domain is due to the enjoyment animals bring to my life and the lives of others. I am a former and current animal owner and the reality of so many abandoned companion animals is heartbreaking.

Related Information (former Kaggle competition): https://www.kaggle.com/c/shelter-animal-outcomes/overview

Problem Statement

Every year, approximately 7.6 million companion animals end up in US shelters. Many of these animals find families to take them in, but just as many are not as lucky. Approximately 2.7 million dogs and cats are euthanized in the US every year.

Datasets and Inputs

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

Two CSV files were provided: train.csv and test.csv. The data used for train.csv and test.csv were randomly split.

- train.csv is composed of roughly 26,700 observations along with 8 features and 1 target variable.
- test.csv is composed of roughly 11,400 observations and 8 features.
- The features include ID, Name, DateTime, Animal, Sex, Age, Breed and Color.
- The target variable has five potential values (Adoption, Died, Euthanasia, Return to owner, and Transfer).

- Both train.csv and test.csv contain missing values and a mixture of data (categorical, numerical, date time).

Information including breed, color, sex, and age of each animal is extremely relevant to the problem of forgotten companion animals. These and other features will help provide insight into trends among breed, color, sex and age.

Reference: https://www.kaggle.com/c/shelter-animal-outcomes/data

Solution Statement

The dataset of intake information (breed, color, sex, age, etc.) will be used to help predict the outcome for each animal as they leave the Animal Center. The outcomes can also help animal shelters identify and understand trends in animal outcomes potentially helping shelters focus their energy on specific animals who need a little extra help finding a new home.

Benchmark Model

The benchmark model I plan to use will be a Logistic Regression classifier that will be trained/tested using the same data and conditions as my solution classifier.

Evaluation Metrics

The evaluation metric used will be multi-class logarithmic loss. For each animal, a set of predicted probabilities will be derived for each class.

For the logloss formula, N is the number of animals in the test set, M is the number of outcomes, log is the natural logarithm, yij is 1 if observation i is in outcome j and 0 otherwise, and pij is the predicted probability that observation i belongs to outcome j.

The probabilities for a given animal are not required to sum to one because they are rescaled prior to being scored.

Project Design

For my capstone project, I intend to follow the workflow outlined below.

Step 1: Problem Preparation
- Load libraries
- Load dataset

Step 2: Data Summarization

- Descriptive statistics such as .info(), .describe(), .head() and .shape
- Data visualization such as histograms, density plots, box plots, scatter matrix and correlation matrix

Step 3: Data Preparation

- Data cleaning such as handling missing values
- Feature preparation and data transforms such as one-hot encoding

Step 4: Evaluate Algorithm(s)

- Split-out validation dataset
- Test options and evaluation metric
- Spot check and compare algorithms such as Decision Tree, Naive Bayes or SVM

Step 5: Improve Algorithm(s)

- Algorithm tuning
- Compare selected algorithm against Ensembles such as AdaBoost or Random Forests

Step 6: Model Finalization

- Predictions on validation dataset
- Predictions on test dataset
- Save model for later use