Machine Learning Engineer Nanodegree

Capstone Proposal

Brice Yokoyama
June 5th, 2018

Shelter Animal Outcomes

Domain Background

According the American Society for the Prevention of Cruelty to Animals ("ASPCA") 6.5 million companion animals enter United States animal shelters every year. Approximately 3.3 million are dogs and 3.2 million are cats. The number of animals entering animal shelters appears to be declining as it is estimated that 7.2 million animals entered animal shelters in 2011.

Upwards of 3 million animals are adopted annually, but it is estimated that 1.5 million of shelter animals are euthanized every year. While the trend of euthanized animals seems to be dropping it is still a problem.

More information regarding the topic can be found here: https://www.aspca.org/animal-homelessness/shelter-intake-and-surrender/pet-statistics

I could find no research paper specifically applied to shelter animal outcomes. Here is a research paper comparing different classifiers: https://datajobs.com/data-science-repo/Supervised-Learning-[SB-Kotsiantis].pdf

Problem Statement

The problem presented by this data is one of maximizing desired outcomes. One idea for a solution is to use a classifier to identify trends in the outcomes based on a set of recorded features. If the classifier is able to successfully predict the outcome for any given animal it could help the animal shelter decide where to put their resources in order to maximize adoptions. The idea being that the shelter can put more resources towards animals who are not predicted to have a desired outcome.

Datasets and Inputs

The dataset is provided by the Austin Animal Center and was hosted by Kaggle as a "Playground" competition. The training data consists of instances of individual animals and the outcomes they reached while at the shelter. After some very basic exploration of the data it is found that there are 26729 instances. Each consists of 8 features and 2 outcomes.

Features of the dataset are:

AnimalID – A unique ID given to each animal

Name – Name of the animal. Value is NaN if no name.

DateTime – Date and time that the outcome occurs.

AnimalType – Cat or Dog

SexuponOutcome – Intact Male, Intact Female, Neutered Male, or Spayed Female

AgeuponOutcome – Age of animal at time that outcome occurs.

Breed – Breed of the animal

Color – Color of the animal

The two outcome labels are OutcomeType and OutcomeSubtype. Within each Outcome there are multiple sub-outcomes that can occur.

Possible OutcomeType values – Return_to_owner, Euthanasia, Adoption, Transfer, Died

Possible OutcomeSubtype – Suffering, Foster, Partner, Offsite, SCRP, Aggressive, Behavior, Rabies Risk, Medical, In Kennel, In Foster, Barn, Court/Investigation, Enroute, At Vet, In Surgery

For this project only the OutcomeType labels are important because the goal is to minimize euthanized and dead animals. The additional information given by the OutcomeSubtype does not change the outcomes.

More information from the Kaggle competition website can be found here: https://www.kaggle.com/c/shelter-animal-outcomes

Solution Statement

One way to make these predictions is to train a classifier to predict the outcome based on the features by training it with a large number of data points. The test data has a set of features and labels as stated in the previous section so this would be a good opportunity to use supervised learning. The types of models to be used will be mentioned in the following sections. The model should then be evaluated on a set of test data points. After training and optimizing the classifier it can be determined from the accuracy score whether or not a meaningful prediction is being made.

Benchmark Model

The benchmark model will be a tuned Random forest classifier. The random forest classifier is expected to perform well so I believe it will serve as a good benchmark model in this situation.

Evaluation Metrics

Per the Kaggle competition submission are evaluated using the multi-class logarithmic loss.

$$logloss = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}\log(p_{ij}),$$

Where:

N = number of animals in the test set

M = number of outcomes

yij = 1 if observation I is in the outcome j and 0 otherwise

pij = predicted probability that observation I belongs to the outcome j

Project Design

The first step will be to analyze the data and possibly drop out any of the features that don't have any bearing on the outcomes (DateTime comes to mind, but who knows before doing any analysis). I will also want to look for any missing values in the data and figure out a way to handle them. I'll have to do some research to figure out how those are handled. Next, some of the training data will have to be split off to be used as a validation set. From here the algorithms can be trained and validated.

I would like to implement multiple classifiers and compare them to each other along with the benchmark model. Besides the random forest classifier, adaboost, gradient boosting and a neural network come to mind as classifiers to compare (I will also look into XGBoost and LightGBM as suggested by the previous review). The classifiers will be tuned using the sklearn gridsearchCV. The neural network will be tuned using strategies from this website: https://machinelearningmastery.com/grid-search-hyperparameters-deep-learning-models-python-keras/.

I would like to play around with the neural network, perhaps making a grid of "# of hidden layers" v. "# of nodes" to see which one performs the best through the validation set.

After the models are optimized the test data will be run through them and then checked against the actual Outcomes. Finally the models will be compared based on their performance on the test set.