MLN Capstone Proposal

Domain Background

This proposal is based on the Shelter Animal Outcomes competition posted in Kaggle, as described here:

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

Animal shelters across the United States end up getting receiving a total of 7-to-8 million new animals each year. These shelters are often able to find new homes for their animals; however, a significant percentage of them ends up being euthanized, as these shelters are not able to find new homes for them. Utilizing historical data, it could be possible to predict animal outcomes based on certain features, which could help these shelters better focus their budgets and efforts to help the most needed segments of their animal population to find new homes.

Problem Statement

Approximately 35% of animal shelter animals, or above 2.5 million annually, end up being euthanized, since a new home cannot be found for them. By looking at historical data collected by Austin Animal Center (AAC), from October 2013 through March 2016, and utilizing Machine Learning algorithms to predict future animal outcomes, we could build a *prediction model* to help AAC and other shelters across the country to better identify segments of their animal population that need extra attention/efforts in trying to find new homes.

Datasets and Inputs

The data to be used for this project is available here:

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

The relevant files for this project are:

File Name	Purpose	Columns	Total rows
train.csv	Training data	10 columns: AnimalID, Name, DateTime, OutcomeType, OutcomeSubtype, AnimalType, SexuponOutcome, AgeuponOutcome, Breed, Color	26,729

File Name	Purpose	Columns	Total rows
test.csv	Test data	8 columns: AnimalID, Name, DateTime, AnimalType, SexuponOutcome, AgeuponOutcome, Breed, Color	11,456

NOTES:

1. The label to be predicted is OutcomeType. For the purposes of this project, OutcomeSubtype will be ignored, as the Austin Shelter goal is to only predict the main outcome.
2. DateTime and AgeuponOutcome will also be ignored, as these are not really features that will be known before making a future prediction. Those features were mistakenly included in the test data file, as explained in this post: https://www.kaggle.com/c/shelter-animal-outcomes/forums/t/22119/cheating-your-way-to-the-top-of-the-lb-remove-the-lb

All the datasets will be stored in *csv* files in a separate directory within the project solution main directory. These data files, along with any other project-related artifacts, will be stored in a github repository.

Solution Statement

It is proposed that we use Supervised Learning techniques, as learned during Udacity's Machine Learning Nanodegree to help us provide a solution. Specifically, this proposal includes utilizing 2 different approaches to build the prediction models:

1. Decision trees
2. SVMs

The models will be trained using the data in the *train.csv* file, and tested by the data in the *test.csv* file.

Benchmark Model

The results for the models produced by each proposed technique will be evaluated against the data provided in the *test.csv* file, and run through Kaggle's evaluation engine. Per Kaggle's forum entries, a reasonable benchmark goal to reach would be a Coefficient of Variance (CV) of 0.70. Please refer to the following forum entries for reference:

- https://www.kaggle.com/c/shelter-animal-outcomes/forums/t/19791/scored-73025-my-features
- https://www.kaggle.com/c/shelter-animal-outcomes/forums/t/20325/why-are-my-scores-so-bad
- https://www.kaggle.com/c/shelter-animal-outcomes/forums/t/19791/scored-73025-my-features

Side note: Unfortunately, the data issue explain in Note 2 in the Datasets and Inputs section of this proposal allows for an exploit in the competition's leaderboard, as explained here:

https://www.kaggle.com/c/shelter-animal-outcomes/forums/t/22124/leaderboard-submissions-claiming-exploit.

This discourages from looking into the competition's leaderboard for benchmarking.

Evaluation Metrics

As discussed in the project's evaluation web page, the solutions will be evaluated using the multi-class logarithmic loss formula as such:

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

where *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*.

Project Design

The following steps are proposed for this project:

1. The first step will require to perform some data exploration to get a better perspective of the domain.
2. After that, the input data will need to be processed to appropriately convert it into measurable features that could be fed into the proposed Machine Learning algorithms.
3. Once the training data is ready, 2 models will be generated for the predictions, one using Decision Trees, and another one using SVMs.
4. The models generated will be evaluated using the data in the *test.csv* file.
5. Iteratively, and depending on the outcome of the tests, further fine tuning of the models and meta-params will be done while running the tests again for further evaluation.
6. Once testing reaches satisfactory levels (presumably > 90% success rate), the models will be submitted to Kaggle's evaluation engine for verification, and reported to the final project's report. Further fine-tuning of the models, and resubmissions to Kaggle may be needed accordingly.
7. The solutions will be provided using python 2.7, scikit-learn. Similarly, along with the final project report, the corresponding supporting material will be provided (graphs, tables, etc.) in a Jupyter notebook file.