Machine Learning - Capstone Proposal

Miguel Ángel Nieto - **Tuesday, 28th February 2017**

Domain Background

Every year there are around 3.9 million dogs and 3.4 million cats that enter shelters. That number take a new significance with the percentage of euthanized, 26% of dogs and 41% of cats. Those are national estimates only for the US, but the problem happens worldwide.

Shelters are most of the time at full capacity and still the number of animals that are under their protection is a tiny fraction of the ones that live on the streets. Only taking cats in account, it is estimated that there are more than 70 million abandoned in the US. Companion animals without food or clean water resources, no access to vaccines and not spayed, just making the problem to grow exponentially. To understand the difficulty for these animals to survive, we only need to compare the life expectancy. 4 years on the streets, 16 in a house. Statistics from: aspca

Problem Statement

There are some animals that get adopted in days. Some others need months or are never adopted. There are different reasons, like age, color or breed that make some of them more prone to adoption. In the other hand, those with problems would need extra work and attention in order to give them a better life. An animal could be categorized as "in danger", based on those parameters and possible others. That would help to automate the process so people working in the shelter can focus their energy on them to increase the chances or being adopted. For example, with more social media publications, better spot on the shelter, more advertising in local neighbourhoods.

Datasets and Inputs

The dataset is from ASPCA, American Society for the Prevention of Cruelty to Animals, and has been published in kaggle.com as part of Machine Learning competition, Shelter Animal Outcomes. The dataset contain different characteristics of each animal, like the ones mentioned in the previous section. A part from those, it also includes "Name", "Sex" and the final Outcome, that could be Adoption, Transfer, Euthanasia, Return to owner and Died.

Solution Statement

With all the different details that describe each animal, and the outcome of each of them, it is possible to create a model that could predict the outcome of the animals that enter the shelter. The person working on the shelter would just need to introduce the data, and the model will give a prediction. This prediction could help to prioritise some cases, for example if it predicts that euthanasia or dead are the most probable outcomes.

Benchmark Model

There are 5 different outcomes. A person just choosing randomly would predict the outcome 20% of the time. The Machine Learning model that created for this capstone should be able to predict with much better accuracy.

Evaluation Metrics

To measure if the model works as expected, accuracy will be used as a metric. The general term "accuracy" is used to describe the closeness of a measurement to the true value. The model will check the results of all predictions, compare them with real outcomes, and calculate how close we are from 100% success.

Project Design

- **Data Visualization**

Different graphs will be used to understand the data better and extract the possible correlations between different features.

- **Features Engineering**

Data inside the features will be cleaned up and fixed. Some of the tasks will be splitting the DateTime, fill cells without values, adjust the different "colour" descriptions. After the data is clean and organized, new possible features will be investigated. Some will be created by splitting the existing ones, or by merging them.

- **Data Split**

Data will be separated in two. 80% for training and 20% for testing. Different classification methods will be used and compared. Decision Trees, Boosting, Linear Regression, SVM and Deep Neural Networks will be tested and the accuracy compared between them.

- **Hyperparameter Tuning**

The one with better accuracy as baseline will be tuned to get the best possible numbers and will be used as final model.