Pet Adoption Prediction

Capstone Proposal

❖ Domain Background

Stray animals are animals that are born on the street or being abandoned by people. Most stray animals live their lives suffering from human abusing, hunger, disease and other health problems. They also may harm humans and other animals through aggressive conflicts and can spread diseases such as rabies.

Kaggle describe stray animals' problem as follows:

"Millions of stray animals suffer on the streets or are euthanized in shelters every day around the world. If homes can be found for them, many precious lives can be saved — and more happy families created.

PetFinder.my has been Malaysia's leading animal welfare platform since 2008, with a database of more than 150,000 animals. PetFinder collaborates closely with animal lovers, media, corporations, and global organizations to improve animal welfare.

Animal adoption rates are strongly correlated to the metadata associated with their online profiles, such as descriptive text and photo characteristics. As one example, PetFinder is currently experimenting with a simple AI tool called the Cuteness Meter, which ranks how cute a pet is based on qualities present in their photos."

❖ Problem Statement

The aim of this project is to develop an algorithm to predict the adoptability of pets.

If we can predict how quickly a pet is adopted, we would be able to guide shelters and rescuers around the world on improving their pet profiles' appeal, reducing animal suffering and getting more pets to be adopted.

The adoptability of pets can be classified into 5 categories representing how quickly the pet is adopted.

0 – Pet was adopted on the same day as it was listed .
1 - Pet was adopted between 1 and 7 days (1st week) after being listed .
2 - Pet was adopted between 8 and 30 days (1st month) after being listed .
3 - Pet was adopted between 31 and 90 days (2nd & 3rd month) after being listed .
4 - No adoption after 100 days of being listed. (There are no pets in this dataset that waited between 90 and 100 days).

Note: This problems is a competition on kaggle called 'PetFinder.my Adoption Prediction'
https://www.kaggle.com/c/petfinder-adoption-prediction

❖ Datasets and Inputs

In this project, I will use the same dataset of the competition on kaggle. The data included text, tabular, and image data.

Data contains:

- Tabular/text data contain these fields:
    - PetID - Unique hash ID of pet profile
    - AdoptionSpeed - Categorical speed of adoption. Lower is faster. This is the value to predict. See below section for more info.
    - Type - Type of animal (1 = Dog, 2 = Cat)
    - Name - Name of pet (Empty if not named)
    - Age - Age of pet when listed, in months
    - Breed1 - Primary breed of pet (Refer to BreedLabels dictionary)
    - Breed2 - Secondary breed of pet, if pet is of mixed breed (Refer to BreedLabels dictionary)
    - Gender - Gender of pet (1 = Male, 2 = Female, 3 = Mixed, if profile represents group of pets)
    - Color1 - Color 1 of pet (Refer to ColorLabels dictionary)
    - Color2 - Color 2 of pet (Refer to ColorLabels dictionary)
    - Color3 - Color 3 of pet (Refer to ColorLabels dictionary)
    - MaturitySize - Size at maturity (1 = Small, 2 = Medium, 3 = Large, 4 = Extra Large, 0 = Not Specified)
    - FurLength - Fur length (1 = Short, 2 = Medium, 3 = Long, 0 = Not Specified)
    - Vaccinated - Pet has been vaccinated (1 = Yes, 2 = No, 3 = Not Sure)
    - Dewormed - Pet has been dewormed (1 = Yes, 2 = No, 3 = Not Sure)
    - Sterilized - Pet has been spayed / neutered (1 = Yes, 2 = No, 3 = Not Sure)
    - Health - Health Condition (1 = Healthy, 2 = Minor Injury, 3 = Serious Injury, 0 = Not Specified)
    - Quantity - Number of pets represented in profile
    - Fee - Adoption fee (0 = Free)
    - State - State location in Malaysia (Refer to StateLabels dictionary)
    - RescuerID - Unique hash ID of rescuer
    - VideoAmt - Total uploaded videos for this pet
    - PhotoAmt - Total uploaded photos for this pet
    - Description - Profile write-up for this pet. The primary language used is English, with some in Malay or Chinese.
- Images

For pets that have photos, they will be named in the format of PetID-ImageNumber.jpg. Image 1 is the profile (default) photo set for the pet. For privacy purposes, faces, phone numbers and emails have been masked.

- Image Metadata
  It's analysis of images on Face Annotation, Label Annotation, Text Annotation and Image Properties using Google's Vision API.
  It may be useful in image analysis.

- Sentiment Data
  It's analysis of pet profile's description on sentiment and key entities using Google's Natural Language API.
  It may be useful in pet description analysis

This data is available on kaggle platform

https://www.kaggle.com/c/petfinder-adoption-prediction/data

I think that the data provided is very good and related to the problem we want to solve.

The data give much information about the pet that, for sure, will help in predicting the adoptability of pets. For example, pet's age, health and color affect its adoptability. Also, providing an image for the pet may increase its adoptability and pet's breed is a factor in the adoptability of the pet, some breeds may be more adoptable than other.

❖ Solution Statement

This problem is a classification problem, which classify a pet into 5 categories. We can use neural networks to perform this classification.

This network will be fed with tabular data, image analysis and text description analysis and the output should one of this categories.

Image can be analyzed using CNN network to extract important information.

❖ Benchmark Model

I am going to use a naïve classifier, as a benchmark model, that always classifies the pet to the middle category (Category 2: Pet was adopted between 8 and 30 days (1st month) after being listed)

❖ Evaluation Metrics

I will use as a metric the quadratic weighted kappa which is used by kaggle competition to calculate the scores.

This metric is described on kaggle as follows

"This metric typically varies from 0 (random agreement between raters) to 1 (complete agreement between raters). In the event that there is less agreement between the raters than expected by chance, the metric may go below 0. The quadratic weighted kappa is calculated between the scores which are expected/known and the predicted scores.

Results have 5 possible ratings, 0,1,2,3,4. The quadratic weighted kappa is calculated as follows.

First, an N x N histogram matrix O is constructed, such that $O_{i,j}$ corresponds to the number of adoption records that have a rating of i (actual) and received a predicted rating j. An N-by-N matrix of weights, w, is calculated based on the difference between actual and predicted rating scores:

$$W_{i,j} = \frac{(i-j)^2}{(N-1)^2}$$

An N-by-N histogram matrix of expected ratings, E, is calculated, assuming that there is no correlation between rating scores. This is calculated as the outer product between the actual rating's histogram vector of ratings and the predicted rating's histogram vector of ratings, normalized such that E and O have the same sum.

From these three matrices, the quadratic weighted kappa is calculated as :"

$$k = 1 - \frac{\sum_{i,j} W_{i,j} O_{i,j}}{\sum_{i,j} W_{i,j} E_{i,j}}$$

Link on kaggle: https://www.kaggle.com/c/petfinder-adoption-prediction#evaluation

❖ Project Design

First, I am going to analyze the data to see the most effective features to be used and which feature is not important. I can use some feature selection methods for this task.

Second, after deciding which features I will use, I will do some preprocessing on the data such as normalization.

Third, I am going to extract the important information from the pet image (if exists) using pre-trained model to be faster.

Finally, I am going to combine (important feature + image information + image metadata + text description analysis) and fed them to neural network. The output layer of this network consists of 5 nodes (5 categories).

The network should minimize categorical cross entropy loss function.
After training the network, I will use the metric described above to evaluate the network's performance.