Capstone Proposal Overview

PetFinder.my Adoption Prediction – Kaggle competition

Domain background

PetFinder is a platform that facilitates the adoption of pets in Malaysia. It enables an easy way for people to find a pet which needs care and shelter. As it is mention in the description of the competition, millions of animals around the world suffer by living on streets and unfortunately, in many cases, they are euthanized in shelters.

A recent initiative from PetFinder was to create a competition in Kaggle to motivate the development of machine learning tools that could enhance the adoption process. This is done by analyzing the profile information of pets in adoption and predicting how fast a pet could be adopted. Identifying weak profiles is a step to increase the chances of adoption by improving the weak features of the profile.

While doing some research in Kaggle, this competition caught my attention because I can empathize with the situation. I live in Mexico and as it happens in Malaysia it is common to observe animals living in streets in poor conditions and in some cases mistreated by people. I would like to apply the techniques that I learned in this course by building a model which has for cause the reduction of animal suffering.

Problem statement

According to the description in Kaggle, the problem is to "predict the speed at which a pet is adopted". This prediction would be done by modeling available features of the pet's online profile. In case the profile gives information about several pets, the speed is defined as the speed of adoption of all pets in the profile.

The speed of adoption is determined by a categorical variable in the range from 0 to 4. The values are defined as follows:

0: Pet was adopted in the same day.

1: Pet was adopted in the 1st week

2: Pet was adopted in the 1st month

3: Pet was adopted between 2nd and 3rd moth

4: No adoption after 100 days

This problem could be addressed as a supervised learning problem were a classification algorithm could be applied.

Datasets and inputs

Petfinder provided several datasets apart from the training and testing set to enhance the predictions. The main files are the next ones:

- train.csv – tabular data containing profile features
- test.csv – tabular data containing profile features
- breed_labels – id and breed of pet
- color_labels – color of pet
- state_labels – state location in Malaysia

Apart from these files Petfinder also provided images and its metadata. This metadata is represented by a json file which is the output of Google's Vision API. Similar to this, pet description was analyzed by Google's Natural Language API, so there is also metadata for the description of some pets. This data files could be used for supplementary analysis.

The data fields that are present in train.csv are:

- 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.

Note: The descriptions were taken from Kaggle data description section.

Solution statement

The problem of classifying the speed in categorical variables given some features enters in the domain of supervise learning. Specifically, the solution requires the application of a machine learning classification algorithm. From what I learned in this course there is no superior classification algorithm that is best suited for all problems. For this reason, my proposal is to try different algorithms like Random Forest, Boosted Trees and Neural Networks. In this sense, I could select the algorithm that best works with this dataset and tune the hyperparameters.

Benchmark model

As this problem follows the format of kaggle competitions a benchmark value useful for comparisons could be the scores of other competitors and the scores obtained by my own previous submissions. This is because all submissions are evaluated by the same metric independently of the algorithm used to solve this problem.

The model is evaluated by using the provided test set that doesn't contain the predicted variable. Kaggle requires competitors to submit a file with a certain format. By doing this the platform automatically evaluates the results and assigns a score for the model. Competitors are ranked according to the scores obtained.

Evaluation metrics

Submissions will be scored using the quadratic weighted kappa. As it is explained in Kaggle, this metric measures the agreement between two ratings. The range goes from 0 to 1, where 1 is a complete agreement.

$$k = 1 - \frac{p_o - p_e}{1 - p_e}$$

$p_o$is the accuracy among raters and p_e is the hypothetical probability of chance agreement. This metric is appropriate for this problem because it captures the agreement between two raters who each classify N items into C categories. As this is a classification task, the score obtained captures the objective of the problem.

Project design

Exploratory data analysis

In this step, I will do some high-level statistics to understand how many observations are in the dataset as well as the data type of the features and if there are null values. Knowing this is important for data preprocessing, and to split the data into balanced training and testing sets.

After this, I will proceed to do univariate and multivariate analysis by making plots like PDF, histograms and pair plots to understand the distribution of the dataset and observe features that are correlated. This analysis will help me determine which features are better indicators of early adoption.

Data preparation

In this section, I will transform all the categorical variables in the dataset by using one-hot encoding. I have observed that the training set has null values in the names column, I will change this value to a more representative string that later could be used for analysis. Also, the metadata of the descriptions has some missing observations because the API couldn't process this description, so I will separate the all the observations that also have available metadata.

Model selection and hyperparameters tuning

I will try different classification algorithms. During this section, I am planning to make predictions with Random Forest, Boosting algorithms and Neural Networks. I am planning to try different algorithms for me to have some benchmark model and compare the predictions by using the quadratic weighted kappa.

After the model selection step, I will select a final model for submission and tune the hyperparameters of this algorithm. For this I am planning to use grid-search or do some research to find the which values could improve the performance of the algorithm.

Present solution

At this final step, I will record the results of my model and document all the techniques I used to obtain the final version. During the model selection step, I will make some preliminary submissions to Kaggle, this will help me compare my score with the scores of other competitors and use this as an indicator of further improvement.

References

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

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

Cohen's kappa: https://en.wikipedia.org/wiki/Cohen%27s_kappa

Hands-On Machine Learning book: http://shop.oreilly.com/product/0636920052289.do