CS3244-2010-0014

PAWBABILITY OF ADOGTION

DETERMINING THE ADOPTABILITY OF PETS IN ANIMAL SHELTERS

Tan Yoong Kang Colin Lim York Tee Gorden Luk Chi Heng Lim Xi Chen Terry Tan Jie Yi Chua Xin Xuan

ABSTRACT

In this project, we aim to predict the adoptability of a pet in an animal shelter given its portfolio. Additionally, we want to identify which traits are desirable and important when people choose to adopt a pet. To do so, we applied Decision Trees and a Convolutional Neural Network to deal with the complex mix of variable types in our data. From our prediction models, we would examine the significant variables and derive insights on how to improve the adoptability of a pet.

MOTIVATION

Every year, approximately 1.5 million shelter animals are being euthanized. This is due to the large numbers of animals being abandoned, yet there are not enough animals being successfully adopted. Hence, we hope that through this project, we can successfully derive an algorithm that can produce features for the profile of said animals that increases the chances of these animals being adopted and thus decrease the euthanization rate in shelters.

METHODOLOGY

Our dataset comprises a plethora of variable types, notably photos and tabular data regarding characteristics of the pet for adoption. Hence, we decided to have two separate approaches:

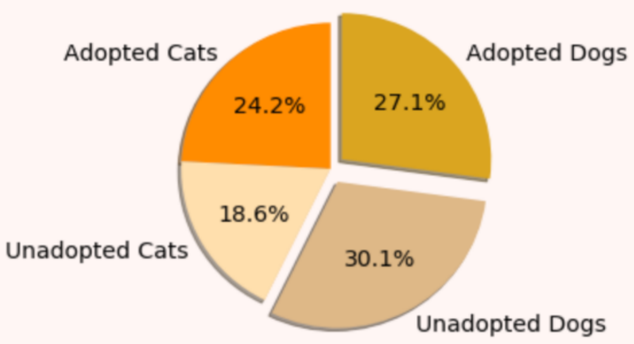
1. Computer Vision (CV) to handle the important aspects in the images of the pets. Convolution Neural Network is used due to its high accuracy in image classification. Images have high dimensionality, thus CNN will serve to reduce the number of parameters required, while preserving important features. This can decrease the problem of overfitting which will lead to high test cost. Hence, resulting in a high accuracy for image classification.
2. Decision Trees (DT) to track the mix of categorical, continuous and ordinal features. Given the structure of our data mutation and mix of variables, we have tried implementing different classification methods such as Multinomial Naive Bayes, Logistic Regression and K-Nearest Neighbors, but chose Decision Trees as not only does it give a high enough accuracy score, but we are able to determine important variables by dissecting the nodes and branches.

FRAMEWORK

DECISION TREE

DATA EXPLORATION & ANALYSIS

- Converted some continuous variables to ordinal data based on their respective quantiles as it's skewed right.
- Columns on "Breed1" and "Breed2" has high dimensions as they can take around 300 different values. Thus, we collapsed this into a single binary variable: 1 indicates the animal is a pure breed while 0 indicates its a mixed breed.
- Dealt with "Age" specially, as the difference in months at earlier stages bear more significance than during old age. Hence, a more intricate interval is used to bin "Age" into ordinal data, where we split between [0-1], (1-2], (2-3], (3-6], (6-12], and (12 and above), to account for younger animals in greater detail.
- The data set only consist of cats and dogs and the distribution is shown as below:



CAT

- Criterion = Entropy
- Max_depth = 8
- Rationale: Optimal accuracy score was achieved at max_depth = 8. We went with it as its score is significantly larger than other max_depth values (by more than 0.005) and has enough representation power.

Evaluation Matrix	Score (to 3.s.f)
Accuracy Score	0.638
Specificity	0.713
sensitivity	0.552
F1 score	0.588
AUC	0.633

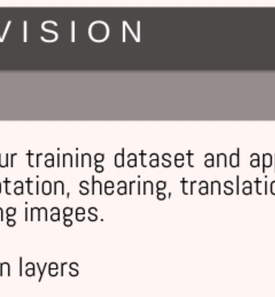
1. Age, Amount of Photos, Fur Length, Sterilized, Maturity Size.
2. For cats, Age is the most pivotal feature with the largest Gini importance of 0.579 compared to PhotoAmt at 0.152.
3. The relationship between MaturitySize and the adoptability of cats is in contrast with dogs, which will be specified later.
4. Maturity Size & Cats: (1 = Small, 2 = Medium, 3 = Large, 4 = Extra Large)
   Cats of the smallest size are heavily favoured, but sizes larger than small have similar adoption rate.

VS

Decision Tree Specification

Evaluation Scores

Important Features (in descending order)

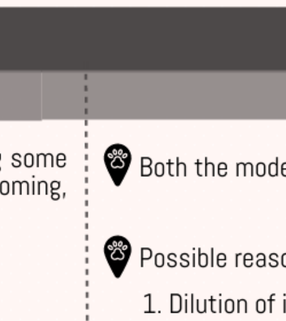


DOG

- Criterion = Entropy
- Max_depth = 5
- Rationale: Optimal accuracy score was achieved at max_depth = 2, but we choose max_depth = 5 (a decrease of 0.003), as it takes more features into account for better specialization, as a max_depth = 2 may generalize too much.

Evaluation Matrix	Score (to 3.s.f)
Accuracy Score	0.638
Specificity	0.713
sensitivity	0.552
F1 score	0.588
AUC	0.633

1. Age, Amount of Photos, Maturity Size, Fur Length, Sterilized
2. Despite having the same attributes as cats, the importance across these features are rather equal, having similar values for Gini importance where Age is only 0.07 higher than Amount of Photos.
3. Previously mentioned, the relationship for Maturity Size and adoptability of dogs differs in distribution from cats, hence we will showcase the contrast.
4. Maturity Size & Dogs: (1 = Small, 2 = Medium, 3 = Large, 4 = Extra Large)
   Dogs of the largest size are heavily favoured, but sizes larger than that have similar adoption rate.



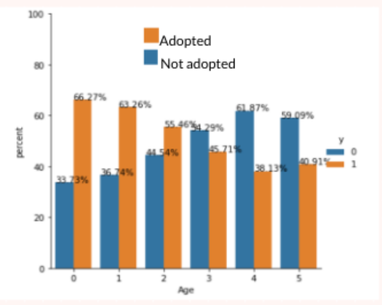
GENERAL FLOW

- Try two criterions: Gini Index and Entropy → Iterate and plot accuracy scores across an interval of lengths for maximum tree depth (max_depth) → Identify best max_depth and review accuracy scores near that depth for pruning → Finalize choice of max_depth and criterion, then implement Decision Tree for the dataset.
- Output Accuracy scores: Comparison across [Accuracy, Specificity, Sensitivity, F1-Score and AUC]. Tuned with a validation set and using a test set to showcase final scores.
- Output Important features: Calculated via "model.feature_importances_", which is based on Gini importance, where how much "impurity" this feature reduces is computed and normalized. The greater the reduction, the better.

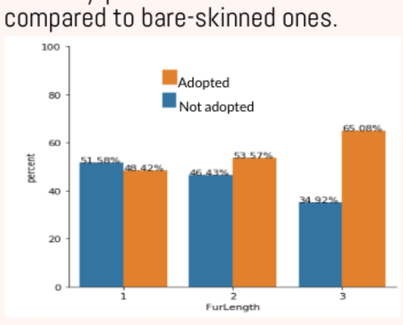
RESULTS & FINDINGS

Besides Maturity Size, all other important attributes were common in both Cats and Dogs. We analyzed their distributions separately and realized that they are very identical. Hence, a combined evaluation is depicted below, to reduce repetition. Blue bar depicts "not adopted" while orange depicts "adopted".

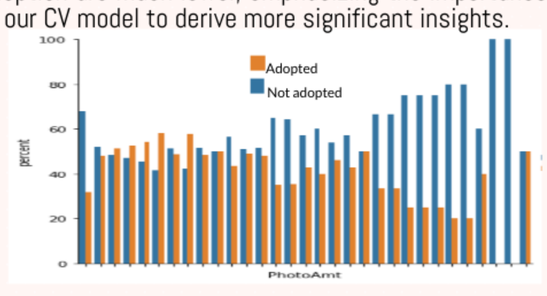
Age
As the pet grows older, its chances of adoption severely decreases, especially in the earlier months.



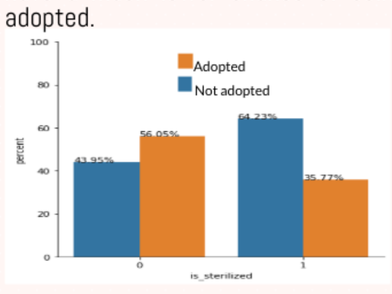
Fur Length
Longer Fur Length seems to be more appealing. This could imply that fluffy and furry pets are more attractive as compared to bare-skinned ones.



Amount of Photos:
Surprisingly, pets with a smaller amount of photos are more likely to be adopted. However, we note that when pets have no photos at all, their chances of adoption are much lower, emphasizing the importance of our CV model to derive more significant insights.



Sterilized
(Binary, 1 = Has been sterilized, 0 = Has not been sterilized):
Surprisingly, pets who are sterilized have a much lower chance of being adopted.



COMPUTER VISION

DATA EXPLORATION & ANALYSIS

- Dropped rows containing no images.
- Used the first image for pets that have more than one photos as it is the profile picture of the pet in the Adoption page.
- Converted AdoptionSpeed Column to an Adoption column that contains only binary values.
  Adoption speed of 0 - 3 is classified as Adopted, while adoption speed of 4 is classified as not Adopted.
- Bottom 10% of the data to be used as testing. Out of the remaining 90%, 80% as training and 20% for validation.

GENERAL FLOW

- Use image augmentation on existing images from our training dataset and applying some image transformation operations to them, such as rotation, shearing, translation, zooming, and so on, to produce new, altered versions of existing images.
- Model: Pre-trained Xception model + Additional hidden layers
  Activation function: Sigmoid
  Optimizer: Stochastic Gradient Descent
  Loss function: Binary Crossentropy
  Metric: Accuracy

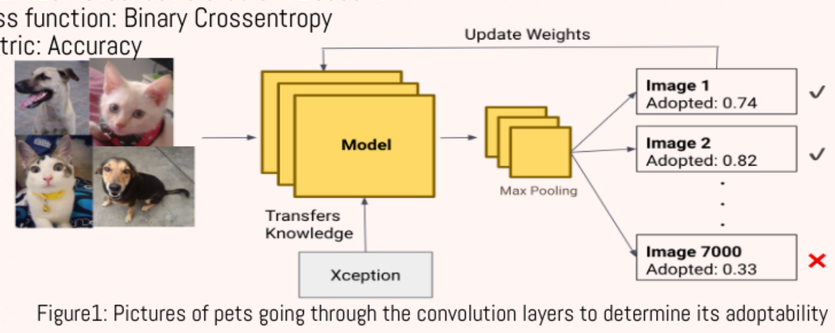


Figure1: Pictures of pets going through the convolution layers to determine its adoptability

RESULTS & FINDINGS

- Both the models for cat and dog yield an accuracy of more than 70%.
- Possible reasons for pets not being adopted include
  1. Dilution of images with text and other images
  2. Collaged images
  3. Background noise (ie presence of human)
  4. Irrelevant images
  5. More than one pet in the photo

CONCLUSION

For the approach under computer vision, separating the data into cats and dogs has increased our accuracy. The general findings for CV is that images which have the pet as the main focus without any additional texts, humans and other animals tend to yield higher adoptability.

On the other hand, DT derived that the Age, Amount of Photos, Fur Length, Maturity Size and whether the pet is sterilized or not are the top five significant factors, where the importance and relationship across these attributes differs to an extent between cats and dogs.

Overall, our general accuracy score is approximately 60%. We note that external factors such as the location of the shelter, number of volunteers etc. were not accounted for in our project due to missing data and these may actually play a vital role in adoptability. All in all, with these features identified, shelters may focus on these features to improve the profile of the animals in their shelter, generally helping to increase their adoptability.

REFERENCE

1. A Comprehensive Hands-on Guide to Transfer Learning with Real-World Applications in Deep Learning, Dipanjan(DJ) Sarkar, https://towardsdatascience.com/a-comprehensive-hands-on-guide-to-transfer-learning-with-real-world-applications-in-deep-learning-212bf3b2f27a
2. Transfer learning and fine-tuning, Billy Lamberta, Bastian Eichenberger and others, https://www.tensorflow.org/tutorials/images/transfer_learning
3. Stanford CS229, Andrew Ng, http://cs229.stanford.edu/syllabus.html