Pet Adoption Prediction

17-537/737 AI For Social Good (Spring 2019)

Loki Ravi, Yuhan Xiao, Anita Chia

Fig 1. Marley & Kiara

Fig 2. Darlie

Fig 3. [No name provided]

Fig 1. Adopted within 1 day - Marley(male) & Kiara(female), 30 month old, Golden Retriever, vaccinated, dewormed, healthy, 1 video, 3 pictures

Description: "We are looking for a new home for our lovely Golden Retrivers, Marley (Male) and Kiara (Female) ... We hope to find them a new home as a pair as they have strong bonding with each other. You can also view them on the video below:[video]"

Fig 2. Adopted within 1 month - Darlie, 48 month old, male, not sure if vaccinated, dewormed, or sterilized, healthy, 1 picture

Description: "PLS HELP!! He was found on the stray with a pair of deaf ear … he is very optimistic and kind. Any details and explanation, pls call me, Wendy."

Fig 3. Not adopted after 100 days - [No name provided], 2 month old, male, minor Injury, 1 picture

Description: "He came to my house himself seeking for new home ... I was unable to keep him so he temporary stay at back yard which don't have any protection"

Related Work

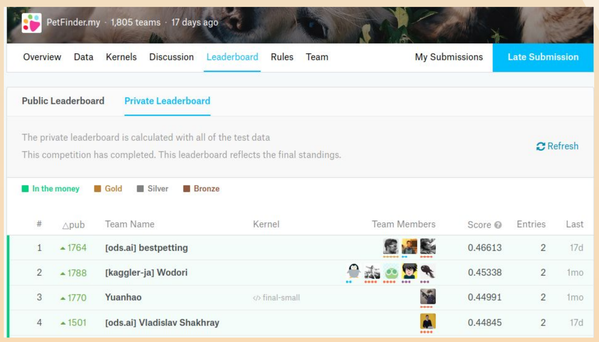


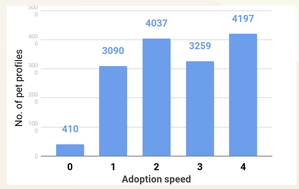
Fig 4. PetFinder.my Kaggle competition

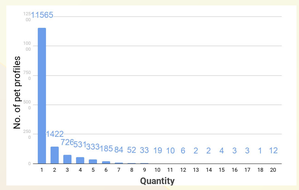
Dataset

PetID, RescuerID	Unique hash ID of pet/rescuer
AdoptionSpeed	*0 = Pet adopted on the same day as it was listed, 1 = Pet adopted within the 1st week after being listed, 2 = Pet adopted within 1st month after being listed, 3 = Pet adopted between 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).*
Type	*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
Breed2	Secondary breed of pet, if pet is of mixed breed
Gender	*1 = Male, 2 = Female, 3 = Mixed, if profile represents group of pets*
Color1, Color2, Color3	Color(s) of pet

MaturitySize	*1 = Small, 2 = Medium, 3 = Large, 4 = Extra Large, 0 = Not Specified*
FurLength	*1 = Short, 2 = Medium, 3 = Long, 0 = Not Specified*
Vaccinated, Dewormed, Sterilized	*1 = Yes, 2 = No, 3 = Not Sure*
Health	*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
VideoAmt, PhotoAmt	Total uploaded videos/photos for this pet
Description	Profile write-up for this pet. The primary language used is English, with some in Malay or Chinese.

Table 1. Descriptions of data fields





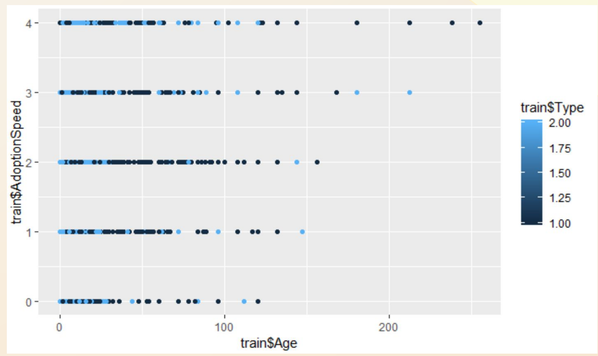


Fig 5, 6. 7. Proof of Class Imbalance

1	Saya jumpa kucing ini yang datang di rumah. Mungkin ada orang yang buang. ...
2	他是被丢在工业区的宝宝，这里两条街大概只狗狗。...
3	Always have ideas and tricks for his playfulness … LOCATION : Taman Desa, OFF Jalan Klang Lama ... *****We highly suggest the adopters to VACCINATE the cat(s) and to spay/neuter it once it reaches its maturity. Thank you! 😀😃😄
4	Introducing strong-willed LuckyPaws! ❤️ Male kitten, 2 months old, white with orange patches and hazel eyes. … ***PLEASE READ***: - LuckyPaws must be kept INDOORS at all times. ...

Table 2. Evidence of Noise

Model

Gradient Boosting Regression:

LightGBM was chosen to conduct regression. It seems to be the most efficient as compared to other boosting algorithms. It is able to learn faster than a typical neural network.

Dataset representation after feature engineering:

- Numeric field: Age, Size, FurLength, Health, Quantity, Fee, VideoAmt, PhotoAmt
- Nominal fields: Breed, Color, State
- Binary fields: Type, Gender, Vaccinated, Dewormed, Sterilized

Missing data in Binary fields:

Usually represented as [1,0] or [0,1]

- Taking a probabilistic approach to fix missing fields => [0.5 , 0.5]

Normalization of Numeric fields:

Nominal and Binary fields always lie between [0,1]. Numeric fields having a much higher upper bound would mean when multiplied by a small weight in a Neuron, they would have an undue impact on the classification. Hence we normalize the numeric fields to lie within [0,1] or [-1,1] according to whether there is a negative effect for the attribute.

Dense Neural Network:

Architecture Selection:

Hyperparameters	Validation					
Hidden Layers	**Metric (Quadratic Weighted Kappa)**		**Loss (Cross Entropy)**		**Accuracy %**	
	10 fold CV	**Hold-out validation**	**10 fold CV**	**Hold-out validation**	**10 fold CV**	**Hold-out validation**
125, 250, 250, 125	.8050	.9341	1.080	.5422	.8590	.9415
250, 500, 500, 250	**.8088**	.9479	1.035	.4242	**.8616**	**.9528**
500, 1000, 1000, 500	.8068	.9491	.9949	.3863	.8585	.9515
125, 250, 500, 250, 125	.8016	.9459	1.065	.4171	.8583	.9455
250, 500, 1000, 500, 250	.8080	.9349	1.042	.4738	.8540	.9422
125, 250, 500, 500, 250, 125	.8006	.9306	1.053	.4625	.8568	.9388
250, 500, 1000, 1000, 500, 250	.8036	.9514	1.038	.3087	.8573	.9528

Table 3. Architecture freeze

Optimizer: ADAM default settings

Loss: Xentropy

Experiments:

Data augmentation with extra data about the state:

ONE_HOT	DATA_AUG	Metric (Quadratic Weighted Kappa)	Loss (Cross Entropy)	Accuracy %
0	0	.8098	1.183	.8621
0	1	.8016	1.209	.8537
1	0	.8026	1.173	.8620
1	1	.7820	1.195	.8502

Table 4. One-hot representation and data augmentation results

Dealing with Class Imbalance:

Focal Loss:

To combat the class imbalance we decided to use the inverse class frequency to do weighted Xentropy. We use a Focus term(f) to control the effect of the weights.

Regular Xentropy is defined as:

$$\mathrm{CE}(p, y) = \begin{cases} -\log(p) & \text{if } y = 1 \\ -\log(1-p) & \text{otherwise.} \end{cases} \quad (1)$$

in the above y ∈ {±1} specifies the ground-truth class and p ∈ [0, 1] is the model's estimated probability for the class with label y = 1. For notational convenience, we define p_t:

$$p_t = \begin{cases} p & \text{if } y = 1 \\ 1-p & \text{otherwise,} \end{cases} \quad (2)$$

and rewrite CE(p, y) = CE(p_t) = − log(p_t)

We define Focal Loss as:

$$FL(p_t) = \text{ClassFrequency}^{-f} \times -\log(p_t)$$

Focal Intensity	Metric (Quadratic Weighted Kappa)	Accuracy %
0	.8624	.8687
1	.8766	.8624
2	.8558	.8703

Table 5. Focal intensity results

Language model for Textual Description:

Model	Accuracy %	Kappa
Naive Bayes with Kernel Estimator	33.70	.1184
Logistic Regression with L1+L2 regularization	36.63	.1552
Support Vector Machine	35.54	.1465
Adaboost M1	29.20	.0194

Table 6. Language models

Limitations:

- Problems in feature representation:
  - A bag-of-words embedding only indicates to the model the presence/absence of a word and hence all information regarding the cohesive placement of words in a sentence is lost.
- Problems in the feature space:
  - These problems arise from having three very different languages in the text. Each character in Chinese represents a word as opposed to English where words are white-space separated. Hence the unigram extractor extracted entire sentences in Chinese as a single unigram.
- Lack of pre-processing:
  - Proper nouns such as names of pets/states etc should be pre-processed into a common representation such as <Name> and <State>
- Lack of cleaning:
  - Emojis
  - Presence of formatting elements

We tried to fix some of the problems identified and improve our model. We have reported our results below. All of these experiments were conducted with Logistic Regression as the model.

Model	Accuracy %	Kappa
Cleaned emojis/numbers + Separated Chinese characters	36.42	.1529
+ Unigrams + Bigrams + Trigrams (Words)	38.75	.1859
+ Description Length	38.70	.185
+ Character N-grams	37.45	.1701

Table 7. Improved language model

Future Work

Utilize the 70k images to build a Image model.