Modeling Pet Adoption Prediction

Lokeshwaran Ravi, Yuhan Xiao, Anita Chia

Carnegie Mellon University, PA 15213, USA
{lravi, yuhanx, sipeianc}@andrew.cmu.edu

Abstract

Real world data is often unclean, noisy, incomplete and imbalanced. These problems can be overcome by using specific mathematical techniques discussed in the paper. We train a dense neural network on the online profile data of pets in animal shelters to predict adoption speed. This allows us to gain some key insights into how to advertise pets as well as provide a numeric evaluation while developing online profiles for pets.

Introduction

An estimated 6.5 million pets are abandoned at shelters annually and more left to fend for themselves on the streets; about half of these animals are adopted while some find themselves at risk of euthanasia, from the statistics given by ASPCA. Given limited resources, these animals are often found in cramped conditions while awaiting adoption. For most animals, their livelihood depends on a brief description and some photographs that are typically found on the internet. By understanding the different factors that lead up to adoption, shelters can give every animal a fair chance of being adopted. This is similar to the adoption of children, where pictures accompanied by a description of their personality serve as the first introduction to potential adoptive parents. Caseworkers, while extremely good with children might need external help in the form of a tool that can evaluate their testimony. Since adoption records are sealed, we hope to address this problem by extension.

PetID, RescuerID	Unique hash ID of pet/rescuer
Adoption Speed	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
Texual Description	Profile write-up for this pet. The primary language used is English, with some in Malay or Chinese.

Figure 1: Composition of a typical animal profile

We have access to a vast amount of data pertaining to pet adoption. This includes photos and videos as well as a textual description in addition to medical and genetic records of the pets. Deep learning has the power to harness this data and learn from it. Since the data has wide modalities ranging from videos to breed types, we expect to develop multiple specific models that make a joint prediction. Our work tries to predict how fast a pet would get adopted given its online profile. In this particular work, we focus on the data fields and the textual description ignoring photos and videos. This would serve as a scoring tool for shelters creating these profiles. We use data from PetFinder.my posted on Kaggle, a description of which is shown in Figure 1.

Background and Related Work

PetFinder.my is Malaysia's leading animal welfare platform and shelter since 2008. It has a large database of over 150,000 homeless animals. PetFinder.my works closely with local and international organizations to improve the welfare of animals and find loving homes for these animals.

Since the dataset is sourced from Kaggle, there is an abundance of related work to borrow and compare from. The Kaggle leaderboard indicates 0.46613 to be the highest Quadratic Weighted Kappa to be achieved on the competition. The discussion boards on Kaggle provided an initial analysis of factors that might affect pet adoption speeds. The winners of the competition made publicly available the details of their methodology involving LightGBM Regression. There have also been similar competitions in the past hosted on Kaggle. Publications are available on the performance metric(Sophie 2014) as well as the contemporary methodology(Guolin et al. 2017) where the problem is modeled as a regression task. In this paper, we chose to model the problem as a classification task.

There are quite a few methods in existing literature to deal with data imbalance, de-noising and incomplete data. Nitesh et al worked on combating data imbalance through oversampling to create synthetic data. Tsung-Yi et al engineered the loss function to deal with the same problem. Salima et al compiled a list of supervised and unsupervised anomaly detection techniques. Pedro et al has worked on a list of techniques to aide with missing data in classification problems.

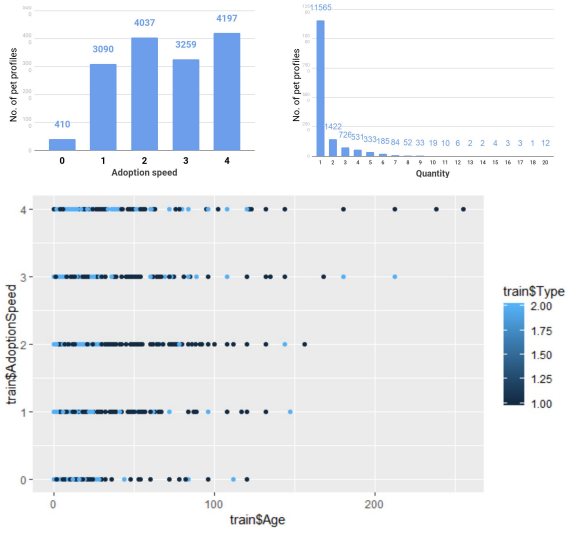

Figure 2, 3, 4 : Evidence of class imbalance in the dataset

Saya jumpa kucing ini yang datang di rumah. Mungkin ada orang yang buang. ...
他是被丢在工业区的宝宝，这里两条街大概只狗狗。...
LOCATION : Taman Desa, OFF Jalan Klang Lama ... *****We highly suggest the adopters to VACCINATE the cat(s) and to spay/neuter it ... Thank you! 😀😃😄

Figure 5: Evidence of noise

Preliminary inspection of the dataset revealed PetID and RescuerID were generated entirely at random and as such did not hold any information relevant to the prediction of the AdoptionSpeed. Hence, these fields were dropped from the dataset. This problem was modelled as a classification problem as opposed to contemporary approaches of modelling it as a regression problem. The dataset consists of 15000 datapoints. After a random shuffle, 1000 datapoints were set aside for testing. 3000 datapoints were set aside for hold-out validation. The rest 11000 datapoints were used for training.

After an initial data analysis using Weka, some class imbalances were discovered in the dataset. For example, only around 2.8% of the datapoints have class 0 adoption speed(pet adopted on the same day as it was listed), as shown in Figure 2, which could impede the accuracy for predicting same-day adoption. Similarly, less than 20% of the data points consist of animals adopted in a quantity of 2 or more(Figure 3), which can affect the prediction for multiple animals adopted together. Lastly, there is a limited number of datapoints on older dogs who get never get adopted(Figure 4). The textual description data of the dataset is noisy and uncleaned(Figure 5), due to the presence of multiple language use(Malay, Chinese and English), formatting elements and emojis, making it harder to make use of the information. Strategies to tackling all the dataset problems are discussed later in the paper.

Our Contributions

Data Representation Improvement

The representation of the data as given in the original dataset is not intuitive to learn. Hence, we explore various representations for the data. Continuous attributes with significance proportional to their magnitude such as Age, Size, FurLength, Health, Quantity, Fee, VideoAmt, PhotoAmt are represented as numeric attributes, whereas multi valued fields with choices of equal significance such as Breed, Color, State are represented as nominal attributes. The rest of the fields with only boolean choices such as Type, Gender, Vaccinated, Dewormed, Sterilized are represented as binary attributes.

Missing data in Binary fields: Investigation of the data revealed that nearly 12% of the data fields were incomplete/missing. Using Imputation would mean we deviate from the original distribution of the dataset in this case, since missing data is represented as missing data to the potential adopters. We are also unable to use ensemble methods to develop a separate model for incomplete data points since 12% of 150,000 is not nearly enough data to train another model without overfitting. We take a probabilistic approach to this issue since most of the missing information were binary or nominal attributes. For example, the attribute Vaccinated was represented as a nominal field with [0,1] representing Yes and [1,0] representing No and [0.5, 0.5] for missing data.

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 Search

In order to pick the optimal architecture for the network we carried out architecture search using 10 fold cross validation and hold out validation. Results are shown below.

No.	Hidden Layers
(1)	125, 250, 250, 125
(2)	**250, 500, 500, 250**
(3)	500, 1000, 1000, 500
(4)	125, 250, 500, 250, 125
(5)	250, 500, 1000, 500, 250
(6)	125, 250, 500, 500, 250, 125
(7)	250, 500, 1000, 1000, 500, 250

Figure 6: Different Architectures

No.	Validation					
	Metric (Quadratic Weighted Kappa)		Loss (Cross Entropy)		Accuracy %	
	10 fold	Hold-out	10 fold	Hold-out	10 fold	Hold-out
(1)	.8050	.9341	1.080	.5422	.8590	.9415
(2)	**.8088**	.9479	1.035	.4242	**.8616**	**.9528**
(3)	.8068	.9491	.9949	.3863	.8585	.9515
(4)	.8016	.9459	1.065	.4171	.8583	.9455
(5)	.8080	.9349	1.042	.4738	.8540	.9422
(6)	.8006	.9306	1.053	.4625	.8568	.9388
(7)	.8036	.9514	1.038	.3087	.8573	.9528

Figure 7: Performances under different architectures, for 10-fold cross validation and hold-out validation

For all the experiments, we used the ADAM optimizer with a learning rate of 0.0001, cross-entropy loss and trained to 600 epochs. We obtained the best results for each model using the training plots.

Data Augmentation

We wanted to find a more intuitive way of representing the StateID data field. With the help of wikipedia we were able to convert names of States in Malaysia into GDP of the state, area of the state and population of the state. We call this data augmentation. Another experiment we performed was converting numeric fields such as MaturitySize and FurLength into ordinal fields to see if we could get better performance. We call this Ordinal representation. The results of both experiments are given below.

[1]	[2]	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

*Figure 8: [1] Ordinal representation and [2] Data augmentation result*s

The results do not indicate any statistically significant variation. From this we infer that the StateID field does not hold enough variance to be converted to a numeric field as opposed to its current nominal representation. Also, representational change from numeric to ordinal did not provide any significant performance improvement.

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

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

and rewrite CE(p, y) = CE(p_t) = − log(pt)
We define Focal Loss as:

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

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

Figure 9: Results with focal loss for various focal intensities

Language model for Textual Description

We tried to leverage the textual descriptions of pets available to improve our performance. Unfortunately the textual descriptions were not cleaned for machine learning. The most common approach for a problem of this sort would be to build a Recurrent Neural Network. But the problem we are faced with is code-mixed text analysis. More specifically, we only have 15k data points to build a language model based in 3 languages. We know from PAC learning theorem that this is an inherently ill posed problem and hence we move away from Neural Networks. We report results on some weak classifiers typically used for text classification. The experiments in this subsection utilized tools such as LightSide, Weka and Stanford NLP ToolKit.

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

Figure 10: Comparison of different language models on the original dataset with Unigram word features

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: There are (1) presence of emojis and (2) presence of formatting elements.

Improved model:

We tried to fix some of the problems identified and improve our model. We cleaned the data to remove emojis, formatting elements as well as replaced numbers and converted all text to lowercase. We extracted all Chinese characters as separate words. 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

Figure 11: Improved language model

The improvement was minimal. We sought the guidance of Professor Carolyn Rose at CMU who is also the author of one of the tools we used, LightSide. After discussing the problems we faced with her and going over possible solutions at length, we understand that cleaning/pre-processing the data for code-mixed text analysis especially between languages such as English, Malay and Chinese has not been attempted as of yet. Such an endeavor would be worthy of a PhD in her opinion. Also we could find no readily available NLP toolkit for Malay. Considering any further effort towards building a language model would be futile, we dropped the textual descriptions.

Conclusion

Factors to Improve Adoption Speed	Factors to Avoid
Vaccination & deworming of pets	Excessive use of capitalization in description
Including videos in the pet's profile	Low-quality photos of the pet
Profile description in various languages	Photos of pets in cages

We developed some key insights while working on the data trying to interpret our models. The cost of pets does not seem to matter. Adopters are willing to pay any amount for a suitable pet for their family. The medical history of the

animal also doesn't seem to be a big issue to adopters. People are more than happy to nurse them back to health. The color of the pet also has little significance to the adopters.

Acknowledgments

This work was part of the class 17-537/737 Artificial Intelligence for Social Good. We would like to thank Professor Fei Fang and our Teaching Assistant Shurui Zhou for their guidance and feedback along the way. We would like to thank Professor Carolyn Rose for her insight into the language model.

References

"Pet Statistics." *ASPCA*, www.aspca.org/animal-homelessness/shelter-intake-and-surrender/pet-statistics.

"Cute Dogs & Cats For Adoption." *PetFinder.my*, www.petfinder.my/.

"PetFinder.my Adoption Prediction." *Kaggle*, www.kaggle.com/c/petfinder-adoption-prediction/leaderboard.

"Quadratic Kappa Metric Explained in 5 Simple Steps." *Kaggle*, www.kaggle.com/aroraaman/quadratic-kappa-metric-explained-in-5-simple-steps.

"Research Papers on Adoptability of Dogs & Cats." *Kaggle*, www.kaggle.com/c/petfinder-adoption-prediction/discussion/76031.

"The Hitchhiker's Guide to the PetFinder Competition." *Kaggle*, www.kaggle.com/c/petfinder-adoption-prediction/discussion/81597.

Ke, Guolin, et al. "LightGBM: A Highly Efficient Gradient Boosting Decision Tree." *31st Conference on Neural Information Processing Systems (NIPS 2017)*, 1 Jan. 1970, papers.nips.cc/paper/6907-lightgbm-a-highly-efficient-gradient-boosting-decision-tree.

"Are You Ready? Lets Learn Some Related Competitions." *Kaggle*, www.kaggle.com/c/petfinder-adoption-prediction/discussion/75969.

Vanbelle, Sophie. "A New Interpretation of the Weighted Kappa Coefficients." *Psychometrika*, vol. 81, no. 2, 2014, pp. 399–410., doi:10.1007/s11336-014-9439-4.

Chawla, Nitesh V, et al. *SMOTE: Synthetic Minority Over-Sampling Technique*. Journal of Artificial Intelligence Research 16 (2002): 321-357

Lin, Tsung-Yi, et al. "Focal Loss for Dense Object Detection." *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, doi:10.1109/iccv.2017.324.

Omar, Salima, et al. "Machine Learning Techniques for Anomaly Detection: An Overview." *International Journal of Computer Applications*, vol. 79, no. 2, 2013, pp. 33–41., doi:10.5120/13715-1478.

García-Laencina, Pedro J., et al. "Pattern Classification with Missing Data: a Review." *Neural Computing and Applications*, vol. 19, no. 2, 2009, pp. 263–282., doi:10.1007/s00521-009-0295-6.

"LightSIDE." *Carolyn Penstein Rose*, www.cs.cmu.edu/~cprose/LightSIDE.html.

"Weka 3: Machine Learning Software in Java." *Weka 3 - Data Mining with Open Source Machine Learning Software in Java*, www.cs.waikato.ac.nz/ml/weka/.

"The Stanford NLP Group." *The Stanford Natural Language Processing Group*, nlp.stanford.edu/software/.

Baltrusaitis, Tadas, et al. "Multimodal Machine Learning: A Survey and Taxonomy." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, 2019, pp. 423–443., doi:10.1109/tpami.2018.2798607.

DeLeeuw, and Jamie L. "Animal Shelter Dogs: Factors Predicting Adoption versus Euthanasia." *SOAR Home*, Wichita State University, 1 Dec. 2010, soar.wichita.edu/handle/10057/3647.