Predicting the likelihood of shelter animal adoption

Final Report

Ben Ying
University of Pittsburgh
bey8@pitt.ed

Kaixiang Li
University of Pittsburgh
kal212@pitt.edu

ABSTRACT

How to place the stray dogs and cats generally become a social problem. Actually, most shelters are overwhelmed today. We try to give a prediction of outcomes for the animals that entered shelter through data mining and machine learning function. This paper mainly talked about the usage of data, what features are important for adoption, comparison of different models and also describe some potential future work based on results we get now. In this way, help stuff manages their work more reasonably like giving extra help for the animals with low adoption probability and increase utilization of shelters.

KEYWORDS

Adoption likelihood prediction, data mining

ACM Reference Format:
Ben Ying and Kaixiang Li. 2018. Predicting the likelihood of shelter animal adoption: Final Report. In *Proceedings of Final Project (Data Mining 18Fall),* Ben Ying and Kaixiang Li (Eds.). ACM, New York, NY, USA, Article 4, 5 pages. https://doi.org/10.475/123_4

1 INTRODUCTION

Every year, approximately 7.6 million companion animals end up in US shelters. Many animals are given up as unwanted by their owners, while others are picked up after getting lost or taken out of cruelty situations. Many of these animals find forever families to take them home, but just as many are not so lucky. 2.7 million dogs and cats are euthanized in the US every year[3].[1]

By analyzing the historical adoption data, we can leverage the data the predict the animal adoption outcomes and understand how features of animal would impact the final outcome of these animals. We need figure out which features would impact most, and for each feature, we can get the most popular features. Volunteers at small animal shelters would normally get a rough analyze on these features because their familiarity on the business experiences, they may know which breed, which color would be easy to get adopted quickly. But it would be difficult for them to help promote those less attracted animals. Animal features such as breeds, parent breeds, color, size, even name and adoption days could lead to different outcome of the animals.

By knowing what really matters for animal adoption, we can help help us improve the animal management and appropriate promotion for some animal breeds lacking attention, which could help enhance the overall adoption rate of the animal shelters. In the current work, we will analyze the data from Austin Animal Center from October 1st, 2013 to March 2016 to analyze animal adopters' preference and predict the shelter animals' final outcome. Data exploration is the major concern when analyzing the data. The original data set has a lot of interesting variables that needs prepossessing before they could be used as prediction variables.

2 RELATED WORK AND MOTIVATION

Basically, related work all focus on the model selection, then compare the accuracy of the models. The main difference is about features selection, some make use of all of them but someone chose to have a function of only animal age, gender, and breed. We agree to extract as much as potential variables firstly. The article from markborg[2] provides some inspiration of how to preprocess the raw data and what else data we can take advantage of.

After getting the complete data set, data cleaning like dealing with the null value or categorical feature transformation is necessary, then we visualize the distribution of the features, the relationship with outcomes, try to take a simple selection firstly. We find that some features are difficult to quantify like color, and some data is imbalance such as breed with many possible values. How to process these data could be a challenge while predicting.

According to the discussion and research, we add some other predictors to explain some features but also keep the original ones and regularize further. In this process, some data are oversimplified depend on our ability, at the end of this paper, we point out some future research work except for the summary and conclusion.

3 DATA

This section will introduce the whole process of expanding data, mainly contains data transformation and the relationship between predictor and outcomes.

3.1 Data Source

Austin, Texas is the largest No Kill community in the nation, and home to the Austin Animal Center. Austin Animal Center provides shelter to more than 16,000 animals each year and animal protection and pet resource services to all of Austin and Travis County. This data describes the information of each animal when they left the shelter according to the following features[5]:

[1] Statistics data is from ASPCA: https://www.aspca.org/animal-homelessness/shelter-intake-and-surrender/pet-statistics

[2] https://mark-borg.github.io/blog/2016/shelter-animal-competition/

Data Mining 18Fall, Nov 2018, Pittsburgh, PA USA
© 2018 Association for Computing Machinery.
ACM ISBN 123-4567-24-567/08/06…$15.00
https://doi.org/10.475/123_4

- Animal ID - unique numeric identifier for each animal
- Name - name identifier for each animal, some animals are unnamed.
- Date Time - exact time when the animals left shelter.
- Outcome Type - includes Adoption, Died, Euthanasia, Return to owner, and Transfer.
- Animal Type - includes Dog, Cat.
- SexuponOutcome - castration situation when the animals left shelter
- Age upon Outcome - age text data formatted as "years or months".
- Breed - Animal breed with mixed or not information.
- Color - Animal fur color. One may have multiple colors.

3.2 Data Preprocessing Approaches

Animal type is a primary factor here such as cats and dogs has a huge difference, this would impact the final outcome a lot and overlays on other features. So we will partition the data into different categories first and get prediction results based on different animal types.

Name is converted into a new binary column called named. In this paper, we did a simple processing firstly, but the original data could provide more information such as the popularity of the names can influence the outcome on some level, the first impression or the sex tendency that name brings. But this information needs much other data support, so we don't consider that yet.

Breed originally is just a text column, we separate it into three new features include MixOrNot which represent it is purebred or not, parent1 and parent2

Color is similar to the breed, presently we can know that the animal is solid-colored, bicolor or tricolor as well as the main color.

Other features, we did preprocess such as reformatting, regularization.

3.3 Data Exploration

This section talked about the exploration of data after preprocessing, which include each features distribution and the relationship with adoption. Information visualization can express the relationship more intuitively and be helpful to give a figure out the important predictors, reduce the number of features that we are going to import into models.

We wanted to explain some predictors from some other angles, for example, breed. The original data is only the name of the breed, but breed means a lot like the size of the animal, the price, or even the prevalence rate of disease. Considering the size data is easy to access, we tried to import it into the model training in this paper and compare the accuracy with former models. [3]

Current data simply represents the corresponding information, not the weight we give or encode in the models. In addition, statistical methods depend on the different features specific situation, we leveraged python, R and Excel.

Age

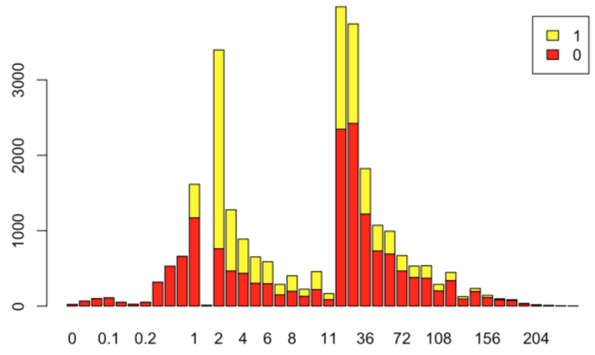


Figure 1: relationship between age and adoption.

The outcome type is transformed into adoption or not, we use 1 represent the animal is adopted successfully. As figure shows, various age range has a different rate of adoption clearly.

Name

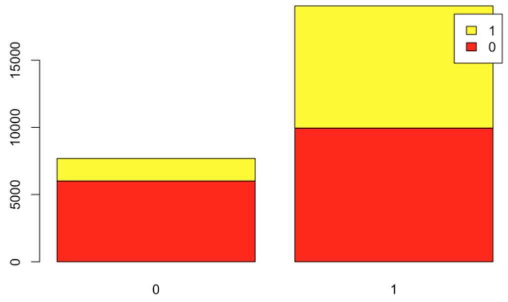


Figure 2: relationship between named and adoption.

As shown, name the shelter animals could be a good idea for lightening the shelter's load.

Breed

As previously mentioned, in the new feature MixOrNot, 0 means purebred, 1 means the animal shows as the main breed but impure. 2 means hybrids and breed of parents is clear.

[3] Data source: https://www.bil-jac.com/dog-breed-library.php

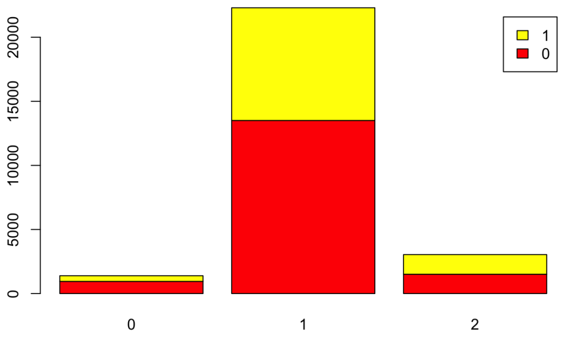
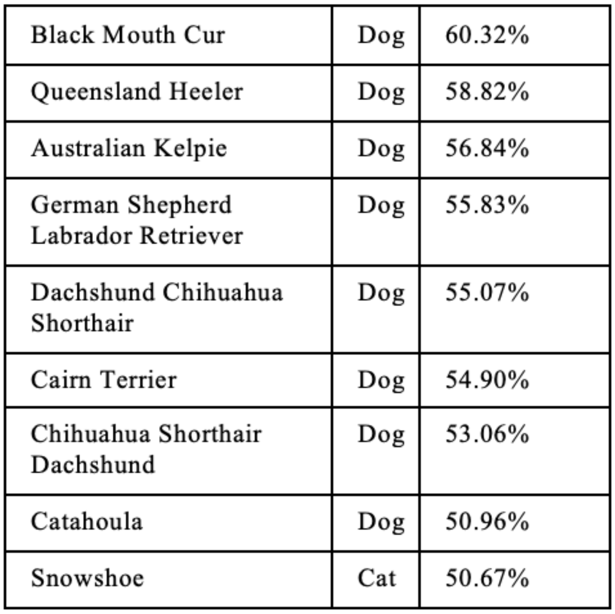


Figure 3: relationship between BreedMixed and adoption

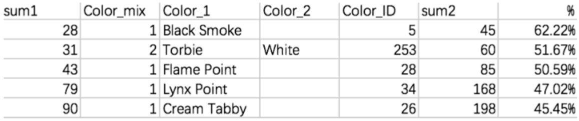


Black Mouth Cur	Dog	60.32%
Queensland Heeler	Dog	58.82%
Australian Kelpie	Dog	56.84%
German Shepherd Labrador Retriever	Dog	55.83%
Dachshund Chihuahua Shorthair	Dog	55.07%
Cairn Terrier	Dog	54.90%
Chihuahua Shorthair Dachshund	Dog	53.06%
Catahoula	Dog	50.96%
Snowshoe	Cat	50.67%

Figure 4: Adopted breed top10

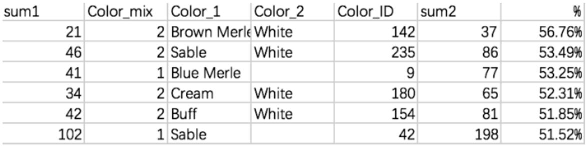
We did statistics for the breed with a large number of samples. Here is a bias: some really valuable and popular breeds may have less possibility to live in this shelter. As shown, mixed breed animals have slightly higher adoption rate. The purity of breed is not an important factor.

Color



sum1	Color_mix	Color_1	Color_2	Color_ID	sum2	%
28	1	Black Smoke		5	45	62.22%
31	2	Torbie	White	253	60	51.67%
43	1	Flame Point		28	85	50.59%
79	1	Lynx Point		34	168	47.02%
90	1	Cream Tabby		26	198	45.45%

Figure 5: relationship between Color and cats' adoption



sum1	Color_mix	Color_1	Color_2	Color_ID	sum2	%
21	2	Brown Merle	White	142	37	56.76%
46	2	Sable	White	235	86	53.49%
41	1	Blue Merle		9	77	53.25%
34	2	Cream	White	180	65	52.31%
42	2	Buff	White	154	81	51.85%
102	1	Sable		42	198	51.52%

Figure 6: relationship between Color and dogs' adoption

The result tells us the prediction should be considered separately depending on the animal type. For cats and dogs, color as predictor shows a clear difference.

SexuponOutcome

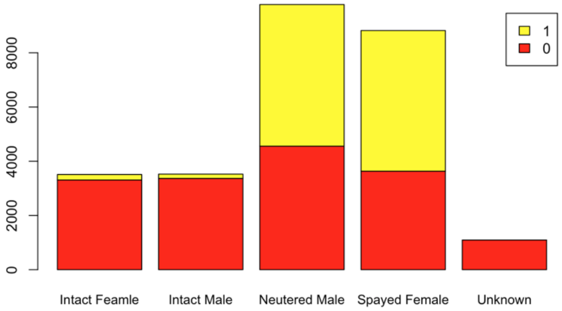


Figure 7: relationship between SexuponOutcome and adoption

Compared with intact animals, castration can increase the chances of adoption. But from the perspective of shelter, there is a consideration about the cost of surgery.

4 METHODS

This model proposes to model historical shelter animal adoption data from Austin Animal Center and evaluate the application of these models in predicting the outcome of shelter animals.
After preprocessing the data, we could use features hasName, breeds, breedMixed, colors, colorMixed, gender, age, animal type and left-shelter time to predict animal outcome.
Our main objective is to predict with logistic regression to work as the baseline and predict with another classifier methods such as SVM, KNN, Decision Tree, Ada and Random Forest to compare with the baseline results. By comparing the accurate rate using different groups of features, we can find the important features determine dogs/cats are adopted or not.
Our second objective is to give extra help for the animals with low adoption probability and increase utilization of shelters based on the prediction result. This could help guide the volunteers in the animal shelter to choose their promotion strategies[2].

4.1 Baseline Model

We applied Logistic Regression as the baseline.
The table below shows the model training results. It shows some important features.

Features	Estimate	Pr
SexuponOutcomeNeutered.Male	3.118504	< 2e-16 ***
SexuponOutcomeSpayed.Female	3.283913	< 2e-16 ***
age in month	-0.019860	< 2e-16 ***
named	1.038565	< 2e-16 ***

The error rate is 18.3% according to the statistics. We'll try different models to see if we could get better results.

	btest	
ytest	0	1
0	2132	612
1	201	1502

4.2 Decision Tree

Decision Tree is easy to interpret and easy to visualize. It can also handle different types of data in the same model. But it doesn't build the same way every time and prone to over-fitting.

4.3 Random Forest

Our next machine learning algorithm is Random Forest. This is an ensemble-type classifier that makes use of bagging on 500 decision trees.

One limitation with the random Forest package of R is that it only allows up to 53 levels for categorical predictor variables. For Breed1 alone, we have 220 distinct breed categories. Thus we need to reduce these category levels in order for random forests to work[1].

Random Forest is reasonably accurate. It can also handle different types of data in the same model. And it guards against over-fitting. But it doesn't predict outside sample and it is a black box for us.

4.4 SVM

We also tried an SVM classifier on this data set, from the e1071 R package. The SVM fails if the input data has NA's in it (missing values) , thus we must ensure that all missing values are set to Unknown or Not Available. SVM is very flexible and accurate. But it is computationally demanding and offers low explanatory power. Also it needs tuning. The tune() method can be used for tuning the parameters of the SVM. This method performs a grid search with 10-fold cross-validation, which is computationally quite heavy.

5 RESULTS AND EVALUATION

The following table shows the error rate for different models and its used parameters. For KNN, the error rate for raw data is 0.2665938. After data preprocessing, the error rate decreased a lot.

Model	Error rate	Parameters
KNN	< 0.1544464	Neighbor = 5
Decision Tree	< 0.1571481	Default setting

SVM	0.1599341	gamma=$10^{(} - 3 : -1)$
		cost=$10^{(} - 1 : 1))$

Ada	< 0.1537998	

Too many breed features may lead to crash when using Decision tree and SVM. So we ranked the features and chose the top 52 features.

6 DISCUSSION

In the future work, we will reduce the number of breeds and colors by using PCA & lasso. Currently we experienced reducing the number of breeds and colors by ranking these features by the count, and then leave the top 52 features based on the restriction of R.

We could also implement multiclass prediction. Currently we only support predicting the outcome whether the cat or the dog is adopted or not. But we could predict more to predict the outcome to be adoption, died, euthanasia, return_to_owner or transfer.

We could also import more data related to the breed. Currently we only get the name of the breed, but we can get more animal info from the breed name from some public data like animal size, animal activity, animal life length. By digesting the breed info into more accurate info, we could predict more accurate result.

7 CONCLUSION

In the final project, we built models to predict animal shelter outcomes in cats and dogs for the Austin Animal Shelter. We cleaned and preprocessed our dataset and engineered more precise features from the original dataset, including more detailed information about sex, coat color and breed. With those finalized features, we created logistic regression, decision tree, random forest, SVM, tuned SVM, KNN and Ada boost models that predicted shelter outcomes and whether or not the animal was adopted for cats and dogs. We found that intact and outcome age are the most important two features. The coast color or pattern, breed played less significant roles. However, for dogs, five features ranked top in the model: breed, intact, outcome age, coat color and outcome weekday. Also having a name or not is a more important feature for determining adopting outcomes in cats than in dogs. On the other hand, the breed of the animal is a more important feature in dogs than cats.

Both Random Forest and SVM models have high accuracies in predicting adoption. As both models provide accurate predictions on adoption outcomes, We would prefer the Random Forest model over the SVM model. With Random Forest, we are better able to gain insights into the features that are most important in determining adoption outcomes, which we can better translate into action[4].

8 RESPONSIBILITY

Ben Ying: Data preprocessing, model training, report writing.
Kaixiang Li: Feature engineering, report writing.

9 ACKNOWLEDGMENTS

The authors would like to thank professor Yuru Lin and Teaching assistants for helpful comments and suggestions.

REFERENCES

[1] Mark Borg. 2016. Kaggle Shelter Animal Outcome Competition. Retrieved December 10, 2019 from https://mark-borg.github.io/blog/2016/shelter-animal-competition/

[2] Joanne. 2016. Predicting animal shelter outcomes. Retrieved December 10, 2019 from https://github.com/jlinGG/Thinkful-DS-Bootcamp/blob/master/Capstone-3.ipynb

[3] Kaggle. 2017. Shelter Animal Outcomes. Retrieved December 10, 2019 from https://www.kaggle.com/c/shelter-animal-outcomes

[4] Adam Levenson. 2016. How Data Science Can Help Us Save Animals. Retrieved December 10, 2019 from https://www.thinkful.com/blog/how-data-science-can-help-us-save-animals/

[5] Kaggle Team. 2016. Predicting Shelter Animal Outcomes. Retrieved December 10, 2019 from http://blog.kaggle.com/2016/08/05/predicting-shelter-animal-outcomes-team-kaggle-for-the-paws-andras-zsom/