Predicting Dog Outcomes in Austin Animal Center using regression and tree-based methods

Hoang Nguyen
Management Engineering
Ulsan National Institute of Science and Technology
Email: hoangnguyen3892@unist.ac.kr

Abstract

There are more than three hundred different breeds of dog in the world. Aside from how different they look, they have been considered as the best friend of human in many aspects such as pet, work and accompany because of their personality. The aim of this study is to identify potential outcomes of the dogs in Austin Shelter Center. We access three different classification techniques to explore not only the characteristics associated with the outcomes but also the performance of each technique which can benefit the shelter and dog-lover community. The ability of model to predict new instances is evaluated based on the logarithmic loss which is computed by using predicted probabilities of each class and its true outcome label. The lower logarithmic loss indicates the better model performance. All of the models considered in this study do classification well for the dataset. However, each model is recommended due to different evaluation criterias.

I. OBJECTIVES

It is estimated that there are 70-80 million dogs are living with humans in the United States and approximately 4 million of dogs enter homeless center annually. Of these dogs, approximately one third are euthanized, 35% are adopted, and 26% of dogs who got lost are returned to their owner [1].

[1]http://www.aspca.org/animal-homelessness/shelter-intake-and-surrender/pet-statistics

Several studies have been conducted to investigate which factors led to dog adoptions. In 2005, [Diesel et al., 2010] examined the success of dog adoptions at 14 re-homing centers in the UK by getting the caregivers answers on a questionnaire six months after adoption. They also carried out a multivariate logistic regression analysis using the variables from their data. Compared with small dogs, they pointed out that medium and large-sized dogs were almost twice as likely to have unsuccessful adoptions. More remarkably, in the study, behavioral issues primarily drove caregivers decision on whether or not to accept a dog. They declared that the less aggressive toward humans a dog was, the more likelihood he got adopted [Diesel et al., 2008].

In another work by [Siettou et al., 2014], the researchers showed that dogs characteristics, namely age, size, appearance, and behavior, played quintessential roles on a predicting model whether a dog would be adopted. Recently, [Hill and Murphy, 2016] emphasized behavior and size were the most important among those characteristics.

Our objective of this study is to build a predicting model based on the Austin Animal Center dataset. Our model uses some features, such as breed, color, sex, time of adoption, and age, to forecast the outcome of homeless dogs. Three methods are exploited to build this model: *Multinomial Logistic Regression (MLR)*, *Random Forest (RF)* and *Extreme Gradient Boosting (XGBoost)*. The results from three methods, and the evaluation criteria in terms of the accuracy and computing time are compared carefully to give an insight of their demerits and merits.

II. Methods

a) Multinomial Logistic Regression: This is a traditional and widely-used approach for predicting dependent variable which is nominal and has more than two levels. In comparison to the other techniques, MLR has been applied in many fields for decades; and it is quite simple and computationally efficient. One critical purpose in choosing MLR is that it gives the regression coefficients, confidence intervals and p-values which are easy to be interpreted. Therefore, the first model we build is MLR to evaluate the interactions between different features and identify the significant variables corresponding to five different outcomes. From this, we can also understand trends in the dog's outcomes.

b) Random Forest: The random forest is a tree-based classification method was first proposed by [Ho, 1995] and further developed by [Breiman, 1996] with the aim of improving the poor prediction of the decision tree. In this method, the tree is constructed by using a bootstrap sample of the data and then split the node into two children nodes based on the best fit among the random subset of the variables. The algorithm yields an ensemble that robust against overfitting and gives a better result than bagging [Breiman, 2001]. Another strength of RF that makes it standing among one of the most popular classification methods is the model outputs *Feature Importances* which is very handy for feature selection. As far as we aware, this technique seems to be very efficiently applicable in genetics and the neuroscience.

c) Extreme Gradient Boosting (XGBoost): XGBoost gives very competitive results to RF and currently is a state-of-the-art method on many problems [Chen and Guestrin, 2016]. It is a gradient boosting technique based on tree ensembles. Since the resurgence of boosting algorithm, it has been applied into many problems such as Protein Classification [Cai et al., 2006] using LogitBoost [Friedman et al., 2000] or Music Classification [Bergstra et al., 2006] using Adaboost [Freund and Schapire, 1996]. XGBoost is standing over the other boosting technique by using regularization to avoid overfitting. XGBoost allows the flexibility in building the model best fit to user's optimization objective and scoring criteria, however, this means that XGBoost requires hyper-parameter setting. Otherwise, XGBoost is also expensive in terms of model complexity and computational time, which requires much more care during training time, compared to MLR and RF. XGBoost also outputs the Feature importances corresponding to weight, gain, and cover which will be discussed in Section VI.

III. Sample

For this study, the dataset is collected from an open source website of Austin Animal Center in Texas, US (see: AAC). This dataset contains information on over 15,000 instances covering the time period of three years from October 1, 2013 to February 28, 2016. After pre-processing, there is a total of 15,499 cases left for running through the model.

The dogs that are rescued and fostered in this center end with five different outcomes, returned to their owner, adopted, transferred, euthanized or died. The visual representation of outcomes of our sample is given on Figure 1.

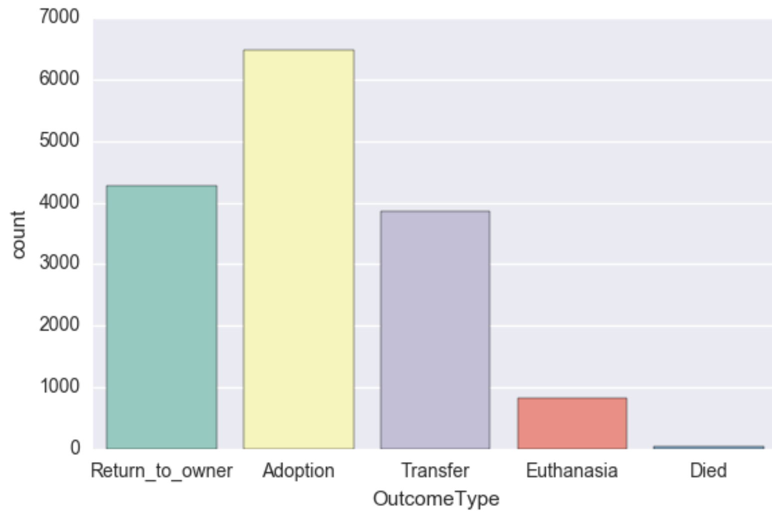

Fig. 1: Count of each level of dependence variables

At the time the dog entered the center, it is given a unique identification number and all of its physical characteristics such as age, sex and breed were tracked by staffs during the intake. Outcome represents the status of animals as they leave the Animal Center. This dataset comprises information about time, name, sex, age and breed which are used in this study.

In order to obtain more features, we categorize dogs into nine different groups including Sporting, NonSporting, Hound, Working, Terrier, Toy, Herding, Pit bull and Unknown. The first seven groups are official dog groups which are fully described by American Kennel Club (AKC), the largest purebred dog registry in the world (see: AKC). Pit bull is one of the most common breeds, however, it does not belong to any official group. We believe Pit bull will give some interesting findings therefore, Pit bull is considered as a separated group in this study. And the last features, the size of the dog, is derived based on its breed and group. The detail of how to obtain these variables is discussed further in Section IV.

Table I shows the summary of the independent variables which is obtained from the dataset. We treat age and size as ordinal variables while the other variables are nominal.

TABLE I: Variables and descriptions

Variables	Levels
Size	Small, Medium, Large, Giant
Age	Juvenile (younger than 2 years old)
	Adult (2-6 years old)
	Old (older than 6 years old)
Sex	Female/Male
Time of outcome	Morning, Midday, Lateday, Night
Day of outcome	Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, Sunday
Group of dog	Sporting, Non Sporting, Hound, Working, Terrier, Toy, Herding, Pitbull, Unknown
Breed of dog	Pure/Mix/Hybrid
Dog has name	Yes/No
Dog is intact	Yes/No
Dog has mixed color	Yes/No

IV. Measures

We could decide whether the dog belongs to *pure*, *mixed* or *hybrid* breed using breed information. To make it simple and feasible for analyzing, we only keep the primary breed of the dogs and then categorizes them into different groups.

Regarding to the age, we only consider the age in years and group into the stages of canine life which is *Juvenile* (younger than 2 years old), *Adult* (from 2 to six years old) and *Old* (older than six years old). Otherwise, the dataset also provides information about gender and whether the dog is neutered/spayed or not. Sex gives information on their original sex for those who were neutered or spayed.

The size of the dog is discovered not based on not its current size but the group it belongs to. This variable reflects the size that dog can reach when it is fully mature. It should be noted here that size and age are treated as the ordinal variables and then encoded, Size (0: Small, 1 = Medium, 2 = Large, 3 = Giant) and Age (0: Juvenile, 1: Adult, 2: Old). The other variables are treated as the categorical variables.

After selecting and processing these features, we predict the fate of the dogs including *Adoption*, *Died*, *Euthanasia*, *Return to owner* and *Transfer*.

Table II gives an inside into the meaning of some terms and Table III illustrates the frequency count for all variables mentioned in this paper.

TABLE II: Glossary of terms

Term	Explanation
Pure breed	A pure breed dog is one that has been registered and proved that the dogs mother and the dogs father are both of the same breed.
Mixed breed	A mixed breed dog is the offspring of two or more different dog breeds where neither the mother nor the father is a registered as pure breed dog. Usually the mixed breed dogs ancestry is unknown
Hybrid breed	A hybrid breed is the offspring of two or more known, but different, dog breeds.

TABLE III: Descriptive statistics of all categorical variables used in the analyses

Variables	Levels	Frequency	Percent	Cumulative frequency	Cumulative percent
Size	Small	5,308	34.25	5,308	34.25
	Medium	8,463	54.60	13,771	88.85
	Large	1,520	9.81	15,291	98.66
	Giant	208	1.34	15,499	100.00
Age	Juvenile	7,206	46.49	7,206	46.49
	Adult	6420	41.42	13,626	87.92
	Old	1873	12.08	15,499	100.00
Sex	Female	7,200	46.45	7,200	46.45
	Male	8299	53.55	15,499	100.00
Time of outcome	Morning	471	3.04	471	3.04
	Midday	7,231	46.65	7,702	98.19
	Lateday	7,516	48.49	15,218	98.19
	Night	281	1.81	15,499	100.00
Day of outcome	Monday	2,220	14.32	2,220	14.32
	Tuesday	2,098	13.54	4,318	27.86
	Wednesday	1,984	12.80	6,302	40.66
	Thursday	2,044	13.19	8,346	53.85
	Friday	2,084	13.45	10,430	67.29
	Saturday	2,469	15.93	12,899	83.22
	Sunday	2,600	16.78	15,499	100.00
Group of dog	Sporting	2,388	15.41	2,388	15.41
	Non Sporting	913	5.89	3,301	21.30
	Hound	1,244	8.03	4,545	29.32
	Working	1,468	9.47	6,013	38.80
	Terrier	1,305	8.42	7,318	47.22
	Toy	3,331	21.49	10,649	68.71
	Herding	2,539	16.38	13,188	85.09
	Pitbull	2,190	14.13	15,378	99.22
	Unknown	121	0.78	15,499	100.00
Breed of dog	Pure	1,131	7.30	1,131	7.30
	Mix	11,363	1973.31	12,494	80.61
	Hybrid	3005	19.39	15,499	100.00
Dog has name	Yes	12,896	83.21	12,896	83.21
	No	2,603	16.79	15,499	100.00
Dog is intact	Yes	2,991	19.30	2,991	19.30
	No	12,508	80.70	15,499	100.00
Dog has mixed color	Yes	9,955	64.23	9,955	64.23
	No	5,544	35.77	15,499	100.00

N = 15,400: Sample size

V. Analysis procedure

A. *Multinomial Logistic Regression*

First, we perform the logistic regression model built in VGAM 1.0-0 [Yee and Wild, 1996], [Wild and Yee, 1996], [Yee and Mackenzie, 2002] to explore the trends and relationships between the independent variables when the outcomes occurred. The results of this model are illustrated in Table IV. In addition, all odds ratios from the logistic regression with *p-value* $\leq$ 0.05 are looked at in detail and *McFaddens R-squared* is also computed to evaluate the goodness of fit of the model.

B. Random Forest

We generate this model by using Scikit-learn [Pedregosa et al., 2011] implemented in Python. Furthermore, in order to optimize the model, parameters are selected within grid-search space from the k-fold cross-validation procedure in which the dataset is randomly divided into several k-fold subset with equal size. Each of these subsets is used to check the performance while the remaining data is used for training. After that, the best parameter giving the highest score is applied to the whole dataset. After that, we compute some other statistical measures, precision and recall, to see how successful the model can predict new instances.

The importance of all features is ranked and visualized to compare with the results derived from the other two models. The accuracy and stability of this method are evaluated based on the percent of the matched predicted labels over total predicted sample and the logarithmic loss calculated from the predicted probabilities.

C. Extreme Gradient Boost

XGBoost is performed using the same package as Random forest. Compare to the former, this method seems to be more sophisticated that requires more parameters and training time. The parameters are chosen by tuning approach which is also supported by cross-validation. We follow the steps from *Analytics Vidhya* [2] and adopt the code of *Aarshay Jain* [3]. We tune the parameter until the model stops improving our customized score function. Latter on, the sam procedure in RF is performed in which, we investigate the precision and recall; and then dig into the ”Feature importances”.

[2]https://analyticsvidhya.com/blog/2016/03/complete-guide-parameter-tuning-xgboost-with-codes-python/

[3]https://github.com/aarshayj/Analytics_Vidhya/tree/master/Articles/Parameter_Tuning_XGBoost_with_Example/

The importance data table after finishing the XGBoost process shows us some figures about the $Gain$, the $Cover$, and the $Weight$. First, the Gain indicates the relative contribution of the corresponding feature to the model which is computed by taking each features influence for each tree in the model. The higher values the metric has when compared to other features; the more important role it plays in generating a predicting model. Second, the Cover metric shows us about the number of observations according to this feature. Finally, the Weight is the percentage representing the relative number of times a particular feature occurs in the trees of the model. It is concluded that the Gain seems to be the most relevant attribute to explain the relative significance of each feature. The accuracy and stability of XGBoost model could be also derived by the similar manner with the Random Forest model.

VI. Results

A. Multinomial Logistic Regression

The results of the MLR are shown in Table IV. This model has the AIC of 30,298 and the McFadden's R-squared of 21.79%.

The feature influencing greatly on the outcome is canine age. Baby dogs have more chances to get adopted when compared with adult ones. They are also more likely get involved in pet transferring. Apart from the positive outcomes, not-fully-developed dogs have a greater tendency to die before reaching their new homes. When the dog is getting older, there is a higher distinct possibility that their previous owners want to bring them back, or they may get euthanized. This result associated with the bond between the owners and their pet.

Dog breed has a lot to do with the outcome. It is revealed that pure-breed dogs tend to be returned to previous owners or easily get transferred to other caregivers. Purebred dogs are in favor since people look for particular characteristics of breed and expect their pet to be healthier. More remarkably, Pit bull could not possibly get adopted and four times more involved in euthanasia than others. This might be caused by negative stereotypes about Pit bull making them being discriminated among other dogs. The sporting dogs who participate in hunting and other field activities are more often being transferred to new owner. Meanwhile, the working

TABLE IV: Summary of variables for Multinomial logistic regression model

Variable	Odds ratio/95%CI				p values			
	Died	Euthanasia	Return to owner	Transfer	Died	Euthanasia	Return to owner	Transfer
Intercept	0.01 ([-6.76, -2.25])	0.08 ([-3.19, -1.87])	0.01 ([-4.97, -3.98])	0.36 ([-1.44, -0.59])	0.00	0.00	0.00	0.00
Age	2.62 ([0.52, 1.40])	5.20 ([1.52, 1.77])	2.96 ([1.02, 1.15])	1.44 ([0.29, 0.44])	0.00	0.00	0.00	0.00
Size	1.74 ([-0.20, 1.31])	1.03 ([-0.20, 1.31])	1.00 ([-0.11, 0.10])	0.93 ([-0.19, 0.04])	0.15	0.78	0.94	0.23
Sex								
Female	1.00	1.00	1.00	1.00				
Male	2.62 ([-0.39, 0.79])	5.20 ([0.26, 0.59])	2.96 ([0.24, 0.41])	1.44 ([0.11, 0.29])	0.49	0.00	0.00	0.00
Time of Outcome								
Morning	1.00	1.00	1.00	1.00				
Midday	0.24 ([-2.25, -0.64])	0.38 ([-1.27, -0.65])	5.57 ([1.36, 2.08])	3.18 ([0.87, 1.47])	0.00	0.00	0.00	0.00
Lateday	0.12 ([-2.98, -1.27])	0.19 ([-1.98, -1.35])	4.20 ([1.08, 1.79])	0.95 ([-0.35, 0.24])	0.00	0.00	0.00	0.72
Night	0.61 ([-2.60, 1.62])	0.05 ([-4.92, -0.90])	2.88 ([0.44, 1.68])	13.26 ([2.16, 3.01])	0.65	0.01	0.00	0.00
Day of Outcome								
Monday	1.00	1.00	1.00	1.00				
Tuesday	1.14 ([-0.83, 1.09])	0.92 ([-0.38, 0.22])	1.22 ([0.04, 0.36])	1.06 ([-0.12, 0.24])	0.79	0.60	0.02	0.50
Wednesday	0.34 ([-2.42, 0.28])	1.02 ([-0.28, 0.32])	1.17 ([-0.01, 0.32])	1.07 ([-0.11, 0.25])	0.12	0.90	0.06	0.44
Thursday	0.63 ([-1.61, 0.68])	1.23 ([-0.08, 0.5])	1.10 ([-0.07, 0.26])	1.30 ([0.08, 0.43])	0.42	0.15	0.27	0.00
Friday	0.84 ([-1.22, 0.86])	0.98 ([-0.31, 0.28])	1.03 ([-0.14, 0.19])	1.19 ([0.00, 0.35])	0.74	0.92	0.73	0.05
Saturday	0.43 ([-1.89, 0.19])	0.32 ([-1.48, -0.83])	0.57 ([-0.71, -0.40])	0.50 ([-0.86, -0.52])	0.11	0.00	0.00	0.00
Sunday	0.70 ([-1.36, 0.64])	0.40 ([-1.24, -0.60])	0.63 ([-0.61, -0.30])	0.60 ([-0.68, -0.34])	0.48	0.00	0.00	0.00
Group of dog								
Herding	1.00	1.00	1.00	1.00				
Sporting	1.47 ([-0.83, 1.60])	1.02 ([-0.30, 0.35])	1.12 ([-0.04, 0.27])	1.19 ([0.01, 0.35])	0.53	0.90	0.16	0.04
NonSporting	1.24 ([-1.50, 1.94])	1.12 ([-0.31, 0.53])	1.29 ([0.05, 0.47])	1.26 ([-0.01, 0.47])	0.80	0.60	0.02	0.06
Hound	0.62 ([-2.67, 1.71])	0.62 ([-0.97, 0.00])	0.94 ([-0.27, 0.15])	1.21 ([-0.03, 0.42])	0.67	0.05	0.58	0.09
Working	1.91 ([-0.58, 1.87])	1.62 ([0.15, 0.82])	1.31 ([0.10, 0.44])	1.28 ([0.05, 0.44])	0.30	0.01	0.00	0.00
Terrier	1.22 ([-1.66, 2.06])	0.72 ([-0.79, 0.14])	0.97 ([-0.24, 0.18])	1.00 ([-0.24, 0.23])	0.83	0.17	0.76	0.97
Toy	4.42 ([-0.02, 2.99])	0.66 ([-0.84, 0.01])	0.90 ([-0.31, 0.09])	1.17 ([-0.06, 0.38])	0.05	0.06	0.06	0.30
Pitbull	3.05 ([-0.05, 2.29])	4.66 ([1.25, 1.83])	1.68 ([0.35, 0.68])	1.49 ([0.21, 0.58])	0.06	0.00	0.00	0.00
Unknown	0.00 ([-596.24, 594.82])	1.50 ([-0.49, 1.30])	0.65 ([-0.96, 0.11])	1.40 ([-0.16, 0.83])	0.97	0.38	0.12	0.18
Breed of dog								
Pure	1.00	1.00	1.00	1.00				
Mix	0.85 ([-1.39, 1.05])	1.08 ([-0.27, 0.42])	0.73 ([-0.49, -0.15])	0.62 ([-0.67, -0.30])	0.79	0.67	0.00	0.00
Hybrid	0.87 ([-1.51, 1.24])	1.04 ([-0.35, 0.43])	0.52 ([-0.84, -0.46])	0.57 ([-0.77, -0.36])	0.85	0.84	0.00	0.00
Dog has name								
No	1.00	1.00	1.00	1.00				
Yes	0.23 ([-2.21, -0.85])	0.46 ([-1.00, -0.57])	7.19 ([1.76, 2.19])	0.61 ([-0.62, -0.37])	0.00	0.00	0.00	0.00
Dog is intact								
No	1.00	1.00	1.00	1.00				
Yes	68.45 ([3.56, 4.90])	65.93 ([3.94, 4.43])	17.79 ([2.67, 3.09])	33.84 ([3.32, 3.72])	0.00	0.00	0.00	0.00
Dog has mixed color								
No	1.00	1.00	1.00	1.00				
Yes	1.00 ([-0.62, 0.61])	1.00 ([-0.18, 0.18])	0.89 ([-0.21, -0.02])	0.91 ([-0.19, 0.01])	0.99	0.97	0.01	0.08

Sample size (N) = 15,499, Akaikes Information Criterion (AIC) = 30,298, McFadden's R-squared = 21.79%.

dogs who are perceived as good companions due to their intelligence and capacity more involve in compassionate death or reunion with the owners.

The sex feature also has a quintessential role to play in anticipating the outcome. Twice as male dogs are reunited to their previous home compared to female ones, but there are four times as male dogs get euthanized as female ones.

Furthermore, if dogs are neutered or spayed, the possibility of their successful adoption is 20 times more while the probability of being euthanized and transfer are 65 and 17 times less respectively compared to intact dogs. This result supports the opinion that dogs should be neutered or spayed.

Some other features also contribute to predicting the outcome. It is observed that those dogs who have name easily to get back to their owners after strays. The ownership transfer more highly occurs on Thursday and Friday while at weekends, dogs have more chance to get adopted.

From MLR, top three most important features are *Age*, *Name* and *Intactness*.

B. Random Forest

The result from RF indicates that *Age*, *Size* and *Intactness* have the greatest impact on predicting the outcome of the dog while the group of dog and time of day only have minor influences. This points out the size is under-valuated in MLR but it is perceived to be important in RF. This result links to the other finding of [Diesel et al., 2008] that smaller dogs are more likely to be adopted than large dog due to the cheaper keeping cost and less damage potential. The day when the even occurs can also reveal the fate of the dog. The summary of important features in RF model is showed in Figure 2.

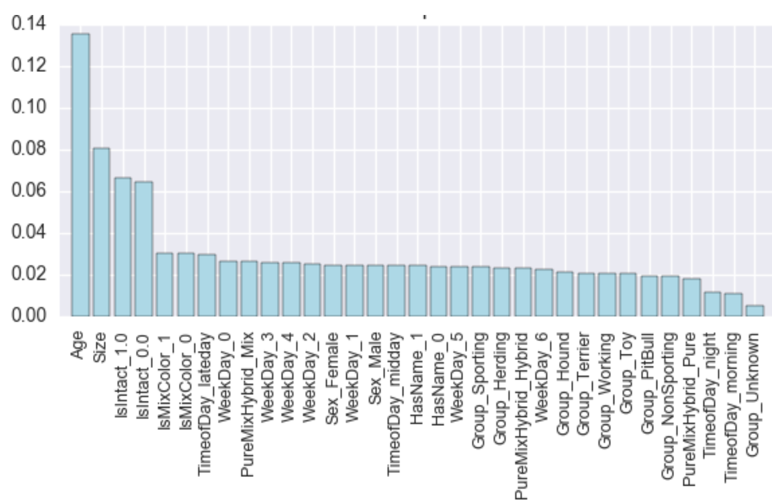

Fig. 2: Important features according to Random Forest

RF can predict 54.74% (+/-1.01%) instances correctly and the logarithmic loss is -1.58 (+/-0.25).

The precision and recall scores for different outcomes are shown in Table V. These values are interpreted together with those obtained from XGBoost in Section VI-C.

C. Extreme Gradient Boost

Among three types of feature importance, Gain reflects the extent that variables contribute to predicting model. It can be seen from Figure 3 that XGBoost yields similar results with MLR which indicates *Intactness*, *Age* and *Name* can reveal the fate of dog significantly.

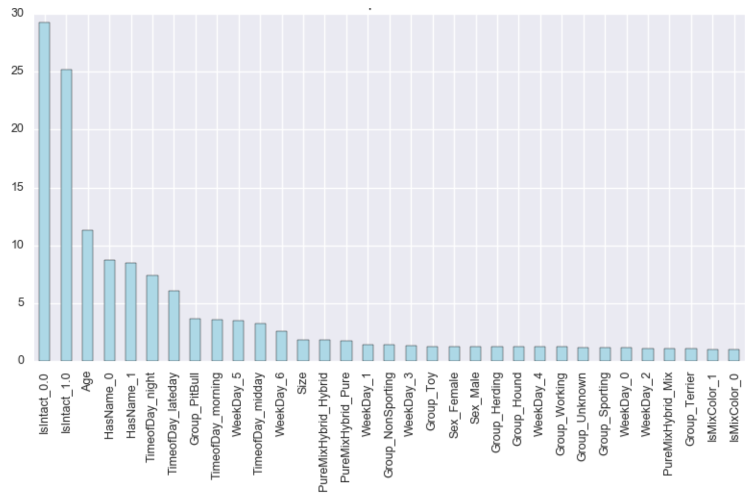

Fig. 3: Important features (Gain) according to XGBoost

Figure 4 shows the rank of feature based on its number of related observations when the model is doing classification. It seems like unknown-group dog, night time and dog's intactness are involving many times during training.

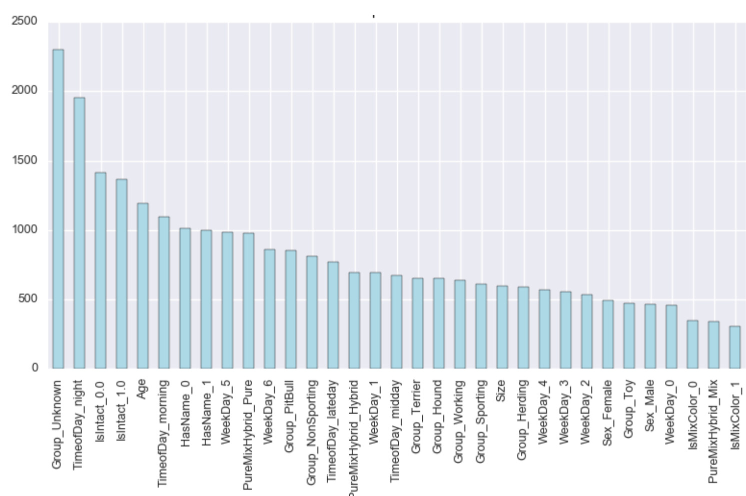

Fig. 4: Important features (Cover) according to XGBoost

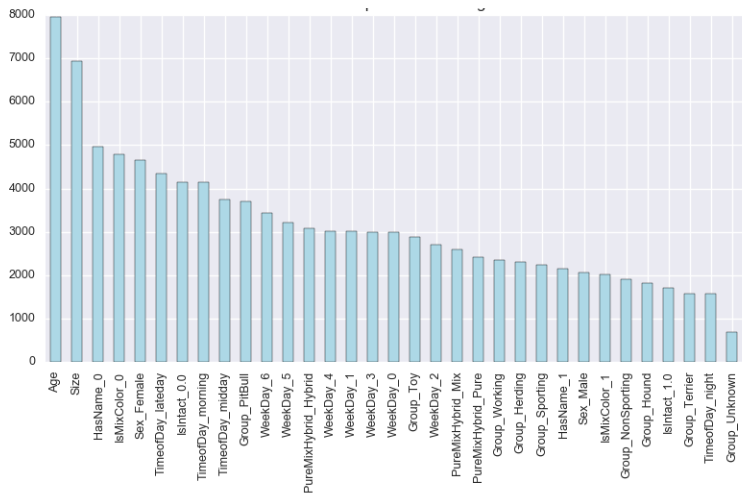

Fig. 5: Important features (Weight) according to XGBoost

Both ensemble classifiers seems to be able to predict new instances with acceptable accuracy. XGBoost appears to be more reliable and give more appropriate results compared to RF since the accuracy is 59.67% (+/-1.00%) and logarithmic loss is -0.96 (+/- 0.02), which makes it so far being the best model in this paper. But there is a trade-off for the higher accuracy. The most obvious disadvantage of this model is that it is computationally intensive because six hyper-parameters are tuned to build the model while in RF, there are only two hyper-parameters.

The precision and recall derived from both RF and XGBoost shown in Table V reveals that both RF and XGBoost give quite appropriate results in the instances regarding adoption, return to owner and transfer. In contrast, it performs the prediction quite poorly in any instances related to the death of dogs including mercy killing because of inadequate information. This brings us back to the fact that Austin Animal Center is a no-kill center where live of the dog is treasured. Assisted suicide is limited to animals that are medically suffering and to dogs that pose an immediate risk to public safety and cannot be safely rehabilitated[4] while animal deaths are unexpected. Therefore, the reason for being euthanasia is not only the age, which is a variable in our model, but also the temperament of the dogs, which is beyond the scope of this paper.

Table V summarizes the statistical measures for three models.

[4]http://www.austintexas.gov/blog/no-kill-austin

TABLE V: Statistical measures

Model	Measure
Multinomial Logistic Regression	McFadden's R-squared = 21.79%
	AIC = 30,298
Random Forest	Accuracy = 54.67% (+/-1.07%)
	Log-loss = -1.58 (+/-0.25)
	Precision = [0.61, 0.00, 0.27, 0.44, 0.52]
	Recall = [0.71, 0.00, 0.16, 0.42, 0.43]
Extreme Gradient Boosting	Accuracy = 59.67% (+/-1.00%)
	Log-loss = -0.96 (+/-0.02)
	Precision = [0.63, 0.00, 0.24, 0.46, 0.66]
	Recall = [0.79, 0.00, 0.15, 0.51, 0.43]

Labels for precision and recall are Adoption, Died, Euthanasia, Return to owner and Transfer respectively

REFERENCES

[Bergstra et al., 2006] Bergstra, J., Casagrande, N., Erhan, D., Eck, D., and Kégl, B. (2006). Aggregate features and adaboost for music classification. *Machine Learning*, 65(2):473–484.

[Breiman, 1996] Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24(2):123–140.

[Breiman, 2001] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

[Cai et al., 2006] Cai, Y.-D., Feng, K.-Y., Lu, W.-C., and Chou, K.-C. (2006). Using logitboost classifier to predict protein structural classes. *Journal of Theoretical Biology*, 238(1):172 – 176.

[Chen and Guestrin, 2016] Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *CoRR*, abs/1603.02754.

[Diesel et al., 2010] Diesel, G., Brodbelt, D., and Pfeiffer, D. U. (2010). Characteristics of relinquished dogs and their owners at 14 rehoming centers in the united kingdom. *Journal of Applied Animal Welfare Science*, 13(1):15–30.

[Diesel et al., 2008] Diesel, G., Pfeiffer, D., and Brodbelt, D. (2008). Factors affecting the success of rehoming dogs in the uk during 2005. *Preventive Veterinary Medicine*, 84(3):228–241.

[Freund and Schapire, 1996] Freund, Y. and Schapire, R. E. (1996). Game theory, on-line prediction and boosting. In *Proceedings of the Ninth Annual Conference on Computational Learning Theory*, COLT '96, pages 325–332, New York, NY, USA. ACM.

[Friedman et al., 2000] Friedman, J., Hastie, T., and Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *Ann. Statist.*, 28(2):337–407.

[Hill and Murphy, 2016] Hill, S. E. and Murphy, N. C. (2016). Analysis of dog adoption success and failure using surveys with vignettes. *Journal of Applied Animal Welfare Science*, 19(2):144–156. PMID: 26863942.

[Ho, 1995] Ho, T. K. (1995). Random decision forests. pages 278–.

[Pedregosa et al., 2011] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V.nand Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

[Siettou et al., 2014] Siettou, C., Fraser, I. M., and Fraser, R. W. (2014). Investigating some of the factors that influence consumer choice when adopting a shelter dog in the united kingdom. *Journal of Applied Animal Welfare Science*, 17(2):136–147. PMID: 24665953.

[Wild and Yee, 1996] Wild, C. J. and Yee, T. W. (1996). Additive extensions to generalized estimating equation methods. *J. Roy. Statist. Soc. Ser. B*, 58(4):711–725.

[Yee and Mackenzie, 2002] Yee, T. W. and Mackenzie, M. (2002). Vector generalized additive models in plant ecology. *Ecological Modelling*, 157:141–156.

[Yee and Wild, 1996] Yee, T. W. and Wild, C. J. (1996). Vector generalized additive models. *Journal of the Royal Statistical Society, Series B, Methodological*, 58:481–493.