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

Hoang Nguyen
Ulsan National Institutes of Science and Technology

Abstract

It is reported that millions of dogs come to the homeless center in the US annually. Of these dogs, approximately one third are euthanized, 35% are adopted, and 26% of dogs who got lost are returned to their owner.

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 center and dog-lover community.

The ability of the model to predict new instances is evaluated based on the logarithmic loss which is computed by using predicted probabilities of each class and true outcome label. The lower logarithmic loss indicates the better model performance.

Objectives

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 and Extreme Gradient Boosting (XGBoost). The results from three methods are compared carefully to give an insight of their demerits and merits.

Methods and Sample

The dataset is collected from an open source website of Austin Animal Center in Texas, US. This dataset contains information of over 15,000 cases covering the time period from October 1, 2013 to February 28, 2016. After pre-processing, there are a total of 15,499 cases left for running through the model.

Methods are used:

- Multinomial Logistic Regression (MLR): a traditional and widely-used approach tha gives the regression coefficients, confidence intervals and p-values used to see the interactions between variables.
- Random Forest (RF): a tree-based classification method constructed by using a bootstrap sample of the data and then split the note into two children nodes. It outputs “Feature Importances”, precision and recall
- Extreme Gradient Boosting (XGBoost)): a gradient boosting technique based on tree ensembles. XGBoost also outputs “Feature importances”, precision and recall. Besides, it gives more accurate results.

Figure 1. Visualization of tree-based method.

Results

- Interactions between variables upon the outcomes:
  - Austin Shelter Animal is a no-kill center.
  - Young dogs have more chances to get adopted and transferred.
  - Older dogs have a higher distinct possibility that their previous owners want to bring them back, or they may get euthanized.
  - Pure-breed dogs tend to be returned to previous owners or easily get transferred.
  - Pit-bull could not possibly get adopted and four times more involved in euthanasia than others because of negative stereotypes.
  - Dogs who have name are easily to be get back to their owners.
- Feature importance:
  - Multinomial Logistic Regression: Age, Name and Intactness.
  - Random Forest: Age, Size and Intactness.
  - Extreme Gradient Boosting : Intactness, Age and Name.

Analysis procedure

- Multinomial Logistic Regression:
  - Regression model is performed to explore the trends and relationships between independent variables when the outcomes occurred.
  - All interactions with p values 0.005 are looked at in detail and McFadden’s R-squared is computed to evaluate the goodness of fit of the model.
- Random Forest:
  - Parameters, number of trees needed to build and maximum number of features in the tree, are selected within grid search space from k-fold cross-validation.
  - Outcome prediction is performed and the accuracy is computed.
  - Some statistical measures, precision and recall, are calculated to see how successful the model can predict new instances.
  - The feature importance is visualized.
- Extreme Gradient Boosting:
  - A set of booster parameters are tuned to build the model with minimum customized score (logarithmic loss is used in this case).
  - Check the performance of model by using cross-validation.
  - The last two steps in Random Forest are also applied in XGBoost.

Chart 1. Count of each level of dependence variables.

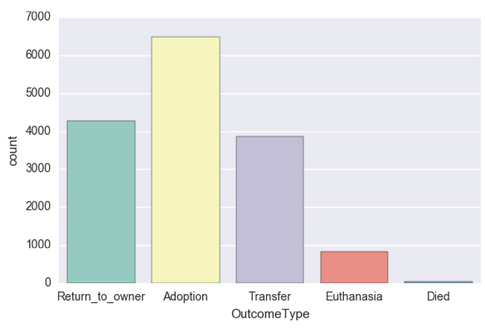

Chart 2. Important features according to Random Forest.

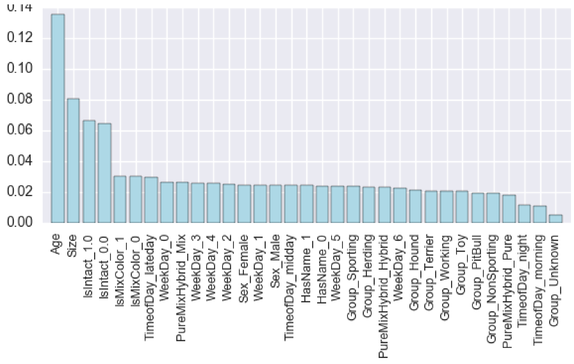

Chart 3. Important features according to XGBoost.

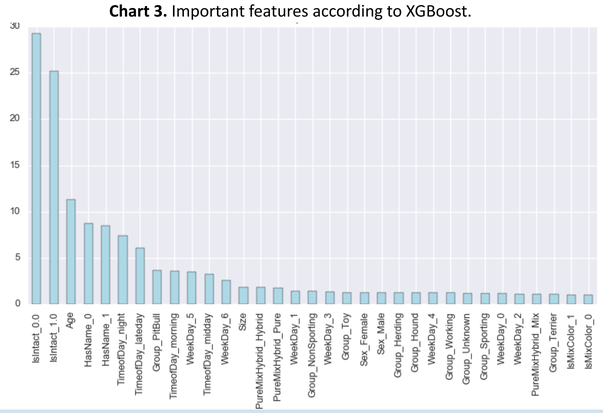

Table 1. Summary of three methods.

Model	Score	“Best-fit” (personal opinion)
Multinomial Logistic Regression	McFadden's R-squared = 21.79%	3
	AIC = 30,298	
Random Forest	Accuracy = 54.67% (+/-1.07%)	2
	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%)	1
	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]	

Contact Information

Hoang Thi Cam Nguyen
School of Management Engineering
Email: hoangnguyen3892@unist.ac.kr
Website: https://hnguyen.info
Phone: 010-7238-3892

References

1. **[Breiman, 2001]** Breiman, L. (2001). Random forests. Machine Learning, 45(1):5–32
2. **[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.
3. **[Breiman, 1996]** Breiman, L. (1996). Bagging predictors. Machine Learning, 24(2):123–140.
4. **[Chen and Guestrin, 2016]** Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. CoRR, abs/1603.02754
5. **[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.
6. **[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.
7. **[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