Shelter Animal Outcome Classification Using Decision Tree, Random Forest, Logistic Regression and Naive Bayes

Group 4 Summer DATS 6202: Lydia, Yuke and Dadian

Introduction

The project aims to examine the life expectancy of an Austin shelter animals by using data collected in 2016 as evidence to investigate the cause of unwantedness in animal shelter, thereby to increase the awareness of inhumane treatment to shelter/abandoned animals by promoting protection to those animals.

Individual work

1. Data preparation: data cleaning, data encoding (label encode, one-hot-encode), create new features (*sex, fertility, MixColor, color, colorC, age, ageC, MixBreed, animal, HaveName, outcome)*.
2. Decision Tree Implement: Used GridSearchCV to gain the best parameters, calculated the confusion matrix, Classification Report, output decision tree graph, plot the ROC curve and calculate AUC value.
3. Wrote group proposal
4. Final report:
    a. Wrote decision tree method introduction
    b. Described the process of preprocessing and one-hot-encoding
    c. Sorted out variables' name and their description
    d. Finished decision tree analysis part
    e. Found the best machine learning method for dataset
    f. Wrote conclusion part
    g. Updated appendix

Method

<u>Decision Tree:</u>

Decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility.

There are several different types of node split criteria. In our report, we use Entropy to calculate the homogeneity of a sample, If the sample is completely homogeneous the entropy is zero and if the sample is an equally divided it has entropy of one.

$$\mathrm{E(S)} = \sum_{i=1}^{C} -p_i \log_2 p_i$$

where c is the number of classes in feature.
We also calculate the Information Gain to decide how to split the nodes of the decision tree.

$$\mathrm{IG(T,a)} = \mathrm{E(T)} - \sum_{V \in val(a)} \frac{|\{x \in T | x_a = v\}|}{|T|} E(\{x \in T | x_a = v\})$$

where $x_a = v$. is the value of the ath attribute of input and y is the corresponding target label.

Evaluation

Confusion Matrix

	Actual		**Measure**
Predicted	TP	FP	Positive Predictive: TP/(TP+FP)
	FN	TN	Negative Predictive: TN/(FN+TN)
Measure	Sensitivity/Recall TP/(TP + FN)	Specificity: TN/(FP + TN)	Accuracy: TP+TN/(TP+FN+FP+TN)

Classification Report

Metrics	Definition	Calculation
precision	Among positive samples, how many of them that we actually predict correct.	True Positive/(True Positive + False Positive)
recall	Among all the samples, how many positive of them that we actually predict correct.	True Positive/(True Positive + False Negative)
F1 - score	The F1 score is the harmonic average of the precision and recall.	2/F1 = 1/P + 1/R
ROC - AUC	AUC is the probability a randomly-chosen positive example is ranked more highly than a randomly-chosen negative example.	False Positive Rate= Number of False Positive / Number of real negative True Positive Rate= Number of True Positive / Number of real positive

Preprocessing

Original Training Dataset

Format	csv
size	716 KB
Observation on training dataset	26730
Number of variables	10
Number of feature	9

	AnimalID	Name	DateTime	OutcomeType	OutcomeSubtype	AnimalType	SexuponOutcome	AgeuponOutcome	Breed	Color
count	26729	19038	26729	26729	13117	26729	26728	26711	26729	26729
unique	26729	6374	22918	5	16	2	5	44	1380	366
top	A719262	Max	2015-08-11 00:00:00	Adoption	Partner	Dog	Neutered Male	1 year	Domestic Shorthair Mix	Black/White
freq	1	136	19	10769	7816	15595	9779	3969	8810	2824

Our dataset comes from Kaggle, there is no target value in test dataset, we cannot calculate the accuracy without target value. So, for this project, we will be splitting by training dataset to new training dataset and test dataset for several machine learning methods, and will encode categorical features to numerical for model processing.

The original features, *Name* and *OutcomeSubtype*, had most missings, it had to do the missing data imputation, so we deleted these features and created a new feature *HaveName*, which stands for whether this animal has name. We also deleted DateTime, according to the description provided by Kaggle, *Datetime* just recorded the time point that shelter staff updated the animal information, it was meaningless for classifying outcome. *SexuponOutcome* feature contained two information, one was gender, another was that whether the animals had fertility ability, so we separated this feature to two features, *sex* and *fertility*. We created new feature *MixBreed* to identify which animals had mix breed. *MixColor* were used to classify which animals had pattern (two colors). *Main Breed* and *colorC* were defined for main breed and color. For our purpose, we also created a new target, *Target*, which generalize categorization of outcome variable. After that, we deleted 20 rows for missing data (19 for *ageC*, 1 for *sex*).

It is noted that the most frequent name in the shelter is "Max", which coincides with the movie's main character, "Max". This implies the shelter's hope to return the animals back to the owner jus as it was in the movie plot.

Table: variables name and their descriptions

Variables Name	**Descriptions**	**Encoding to Numeric**
Main Breed	Ecoded main breed for animal	Assumed main breed for each animal. Defined as the first breed before "/" and "Mix".
sex	Gender	Male: 0, female: 1, unknown: 2
fertility	Fertility ability	Intact: 0, spayed: 1, unknown:2
MixColor	Whether the color of animal is mix	Mix: 1 , Pure: 0
colorC	The classifier code for main color	There are 57 different color, we encode color to the numerical type data. For example, 0 stands for Black, 1 stands for White. More information in is Appendix i.
outcome	Adoption, Died, Euthanasia, Return to owner and Transfer	Adoption: 0 Transfer: 1 Return_to_owner: 2 Euthanasia: 3 Died: 4
age	Age by day	For example: 1, 60, 365
ageC	The classifier code for age	ageC = i when i years, i are in [1,9] ageC=0 when year <1
MixBreed	Whether the breed of animal is mix	1: mix, 0: pure
animal	The type of animal	1: dog, 2: cat
HaveName	Whether this animal has name	1: yes, 0: no
Target	Generalized categorization of outcome variable	1: Survived 0: Died

One Hot Encoding

To further expand the dimensionality of our data, we applied One Hot Encoding method categorical variable. The resulting number of feature reached to 336.

Label design for modeling

We will be investigating based on two types of label variables:

Original outcome variable levels (Outcome):

Level	Label	Count
Adoption	0	10769
Transfer	1	9406
Return_to_owner	2	4785
Euthanasia	3	1553
Died	4	197

Aggregated outcome variable (Target):

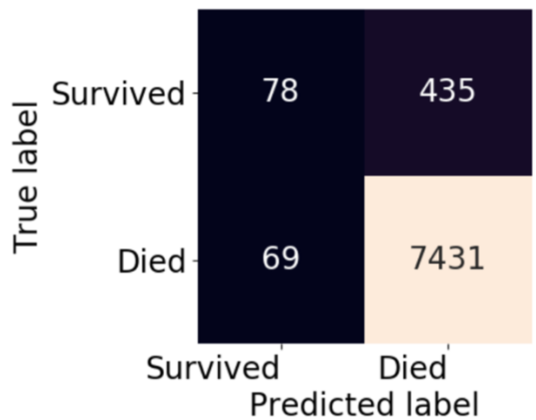
Target	Level	Label	Count
Survived	Adoption + Transfer + Return_to_owner	1	24960
Died	Euthanasia + Died	0	1750

Model Analysis

Decision Tree

When *Target* was target, which only contained 2 kinds of outcomes (survived and died), we used GridSearchCV to find the best parameters for decision tree. As the result, we chose the 15 be he maximum depth of the tree(max_depth=15), 11 be the minimum number of samples required to be at a leaf node(min_samples_leaf=11). We selected top 10 important features to present. The most important feature in this model was fertility ability. After that, the second most important feature was *ageC_0*, which meant that less-than-one-year old animals. *HaveName_1*(the indicator for the animals which have name) was the third most important feature. More detail is in the Appendix ii.

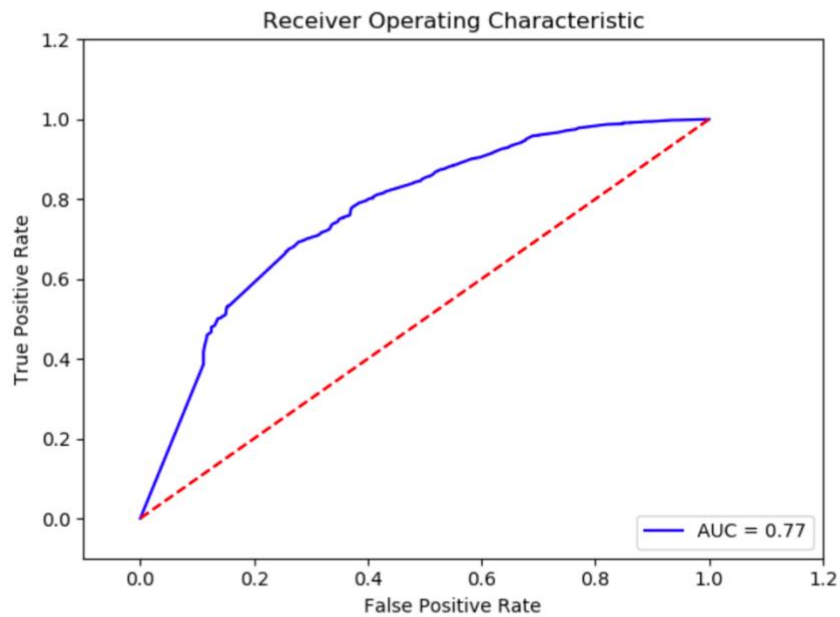
The accuracy was 93.71%. However, using *Target* be the target was not very appropriate. There were 17460 targets were equal to 1 in training dataset, which accounted for 93.38%. In testing dataset, 7500 (93.60%) targets were 1. So we could not determine that whether this high accuracy was caused by successful decision tree, or unbalanced dataset. To figure out this question, we used ROC curve and AUC score to visualize the performance of this classifier, From the plot, we saw a good "hump shape" curve, AUC=0.77, which meant that the probability that a randomly chosen Survived example was ranked higher than a randomly chosen died example is 77%. So, this classifier was not so "bad", it "learnt" something from training data.



Confusion matrix for Entropy Decision Tree when *Target* was target

Table: Classification Report when *Target* was target

target	Precision	Recall	F1-score	support
0	0.53	0.15	0.24	513
1	0.94	0.99	0.97	7500
avg/total	0.92	0.94	0.92	8013
Accuracy	93.71%			



When *outcome* was target, we also used GridSearchCV to gain the best parameters for decision tree. Final max_depth and min_samples_leaf were 7 and 19, respectively. The top 3 important features were *fertility_1*, *ageC_0* and *HaveName_0* (the indicator for the animals which have no name), respectively. The accuracy was 63.98%.

Confusion matrix for Entropy Decision Tree when *outcome* was target

Table: Classification Report when *outcome* was target

target	Precision	Recall	F1-score	support
0	0.66	0.80	0.72	3271
1	0.74	0.62	0.67	2828
2	0.45	0.49	0.47	1401
3	0.49	0.17	0.25	464
4	0.00	0.00	0.00	49
avg/total	0.64	0.64	0.63	8013
Accuracy	63.98%			

According to the performance comparison table, Logistic regression performed best with the highest AUC, precision, recall, accuracy.

Table: Performance comparison for 4 methods when *Target* is target

Model	AUC	Precision	Recall	Accuracy
Naive Bayes		0.90	0.71	0.71
Decision tree	0.77	0.92	0.94	0.94
Random Forest	0.77	0.94	0.98	0.93
Logistic Regression	0.80	0.94	1.00	0.94

Conclusion

For our report, we chose logistic regression method be the best machine learning method when variable *Target* was target. The most important feature for shelter animals survival were fertility ability, whether having name or not, younger animals or older animals, “pit bull” breed, “domestic longhair” breed, and gender. Using logistic regression method, we classified almost all survived animals correctly (Recall=1.00).

Limitation

Mainly showed as the following aspects:

1. Our dataset was unbalanced, which impact the determination of performance of classifier.
2. We encoded variable outcome to binary variables as our target, which let the situation be simpler than the reality.

Appendix

Appendix i: Color Code

color	code	color	code	color	code	color	code
Black	0	White	1	Brown Tabby	2	Tan	4
Orange Tabby	5	Blue	6	Tricolor	7	Red	8
Brown Brindle	9	Blue Tabby	10	Tortie	11	Calico	12
Chocolate	13	Torbie	14	Sable	15	Cream Tabby	16
Buff	17	Yellow	18	Gray	19	Cream	20
Fawn	21	Lynx Point	22	Blue Merle	23	Seal Point	24
Black Brindle	25	Flame Point	26	Gold	27	Brown Merle	28
Black Smoke	29	Black Tabby	30	Silver	31	Red Merle	32
Gray Tabby	33	Blue Tick	34	Orange	35	Silver Tabby	36
Red Tick	37	Lilac Point	38	Tortie Point	39	Yellow Brindle	40
Blue Point	41	Calico Point	42	Apricot	43	Chocolate Point	44
Blue Cream	45	Liver	46	Blue Tiger	47	Blue Smoke	48
Liver Tick	49	Brown Tiger	50	Black Tiger	51	Agouti	52
Silver Lynx Point	53	Orange Tiger	54	Ruddy	55	Pink	56
Brown	3						

Appendix ii: Top 10 important features, their rates and descriptions for entropy decision tree when *Target* was target

Top	Feature	Rate	Description
Top 1	fertility_1	0.25013	The animals are spayed.
Top 2	ageC_0	0.15456	The animals which is younger than 1 year.
Top 3	HaveName_1	0.06069	The animals which has name.
Top 4	Main Breed_159	0.06027	The animals’ breed is “pit bull”.
Top 5	ageC_1	0.05419	1-year-old animals
Top 6	animal_1	0.0503	Dog
Top 7	ageC_2	0.04157	2-year-old animals
Top 8	MixColor_1	0.02396	The animals which colors are mixed
Top 9	sex_0	0.0235	Male animals
Top 10	ageC_3	0.01981	3-year-old animals

Appendix iii: Top 10 important features, their rates and descriptions for entropy decision tree when *outcome* was target

Top	Feature	Rate	Description
Top 1	fertility_1	0.50466	The animals are spayed.
Top 2	ageC_0	0.18482	The animals which is younger than 1 year.
Top 3	HaveName_0	0.09453	The animals which has no name.
Top 4	animal_2	0.06628	Cat
Top 5	HaveName_1	0.02948	The animals which has name.
Top 6	ageC_1	0.02628	1-year-old animals
Top 7	animal_1	0.02133	Dog
Top 8	ageC_2	0.01926	2-year-old animals
Top 9	Main Breed_159	0.01197	The animals’ breed is “pit bull”.
Top 10	sex_1	0.00563	Male animals