CS630-02 Foundation of Intelligent System
Team Project Final Report
Team: Shih-Ting Huang (sh3964@rit.edu)
Wei-Yao Ku (wxk6489@rit.edu)

I. TITLE

Prediction of Outcomes in Animal Shelter

II. INTRODUCTION/MOTIVE

Annually, there are 6.5 million homeless animals impounded in US animal shelter but not all of them had a happy ending. Sadly, near one-third of the companion animal still being euthanized with humanely destroyed. We desire to help the animal shelter to decrease the unnecessary killing and finally reach the No kill goal of companion animals. The aim of this project is to build a system that can accurately predict outcomes of dogs and cats in the shelter. The outcomes include returned, adopted or euthanized. We compared the prediction rates using different classification algorithms. The prediction was based on the relation between outcomes and animal features such as gender, age, breed, color and staying time in the shelter. Our model could provide knowledge for the animal shelter to enhance the adoption rate.

III. DATA

We used data from "Animal Service Intake and Outcome-Louisville Metro Open Data" (https://data.louisvilleky.gov/dataset/animal-service-intake-and-outcome). Compared to other candidate datasets, this dataset included more comprehensive information such as intake date, outcome date, and date of birth. There were around 140,000 records with 23 attributes in the original dataset. We worked on data matching and data cleaning to streamline the data frame for our analyses.

The data cleaning steps were as follows. (1) We focused on dogs and cats and removed other animal types like rabbits, birds, etc. (2) New variables of Age and Stay Time were calculated from Date Of Birth, Intake Date, and Outcome Date. Age and Stay Time could provide meaningful information than the raw variables. (3) If there was a secondary label for color and breed types, we created an indicator of marked or mixed. (4) The outcomes were collapsed as adopted versus others because the adoption is our primary outcome in this project. (5) Some less meaningful attributes such as Intake Reason and Outcome Subtype were removed.

Next, we divided data into two groups as dogs (73,488 records) and cats (65,115 records) for model building with 11 feature attributes: Stay Time, Intake Type, Primary Color, Color Mark, Primary Breed, Secondary Breed, Gender, Secondary Color, Age, Reproductive Status, Asilomar Status. Since attribute values were not numeric, we generated dummy variables for each attribute. Therefore, there were 74 and 66 attributes in dogs' and cats' data sets, respectively. Please see the attachment one for the variable listing.

IV. PROPOSED METHOD and IDEA BEHIND

Since the raw data included labels for the outcomes, supervised learning will be fitted to build up our prediction model. There were two primary forms of supervised learning: regression and classification. Because we only focus on the binary outcomes (adopted as 1 versus others as 0), we chose three classification methods: (1) logistic regression: a linear classifier, (2) decision tree: one-stage, and non-linear classifier, and (3) random forests: multi-stage classifier, to build the prediction models.

1. Logistic Regression:
   The concept behind logistic regression is similar to linear regression, assuming there exists a linear relationship between input and output. The main difference is that the output of logistic regression is always between 0 and 1 (binary classification). The learning method will find the best coefficients of each attribute to build the prediction model.

2. Decision Tree:
   We choose decision tree learning as one of our methods because decision tree learning is one of the simple methods to understand and interpret learning algorithm. The decision tree learning algorithm adopts a greedy divide-and-conquer strategy: always test the most important attribute first. This test divides the problem up into smaller sub-problems that can then be solved recursively.

3. Random Forest:
   The random forest is the extension of the decision tree, a collection of decision trees. The idea is that instead of making a decision by one tree, it assigns the label based on the majority vote from trees. Each tree in forests is trained with a subset of training examples, and the classification criteria are different at each level. This mechanism can overcome the over-fitting problem. However, if most of the variables have weak interactions, random forests will be less efficient.

V. EXPERIMENTAL RESULTS and ANALYSIS

1. Accuracy and running time of different methods
   In general, the predicted correction rate was lower and unstable with less training data sets, but it increased and became stable after feeding with more data. Overall, the predicting correction rate of cat dataset performed better than dog dataset for all four methods. The best method to build the model in dog dataset is the logistic regression (Figure 1 and Table 1) and the decision tree with 5-depth limitation has better performance in cat dataset (Figure 3 and Table 2). The decision tree without depth limitation performs the worst in both datasets.

   Both random forest and depth-limited decision tree had mechanisms to avoid overfitting and enhance the correction rate. As we mentioned before, random forest was a collection of decision trees. Therefore, performance of running time was the worst when using random forest (Figure 2 and Figure 4). On the other hand, the decision tree with 5-depth limitation was the fastest algorithm to generate the model prediction in both datasets. Thus, taking both accuracy and running time into account, the best method to train our prediction model was depth-limited decision tree.

Table 1: Dog adoption outcome prediction accuracy rate

Logistic Regression	Decision Tree	Decision Tree (d=5)	Random Forest
80.69%	78.47%	79.18%	79.23%

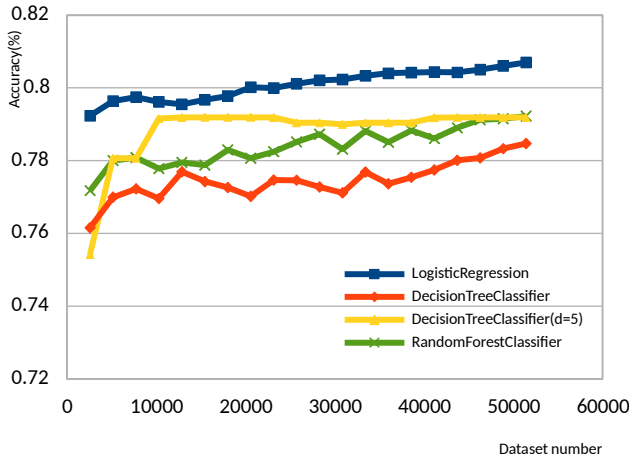

Fig. 1: Dog dataset number v.s. accuracy

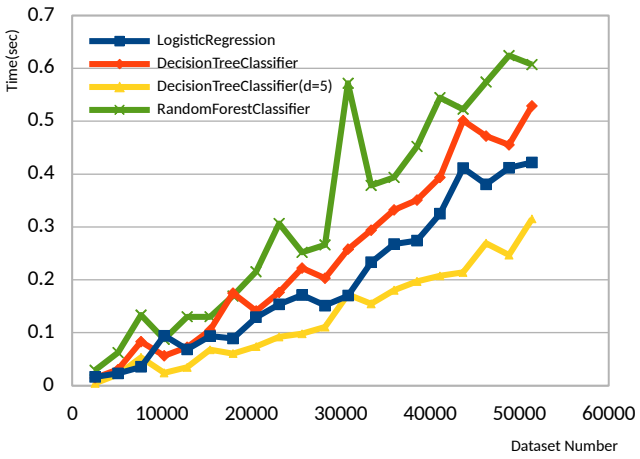

Fig. 2: Dog dataset number v.s. running time

Table 2: Cat adoption outcome prediction accuracy rate

Logistic Regression	Decision Tree	Decision Tree (d=5)	Random Forest
89.73%	88.34%	90.00%	88.74%

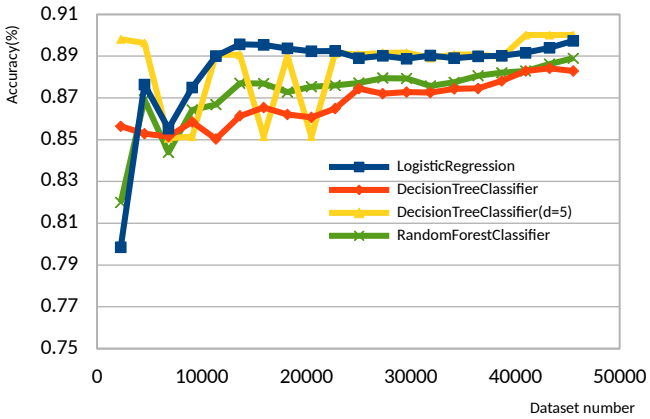

Fig. 3: Cat dataset number v.s. accuracy

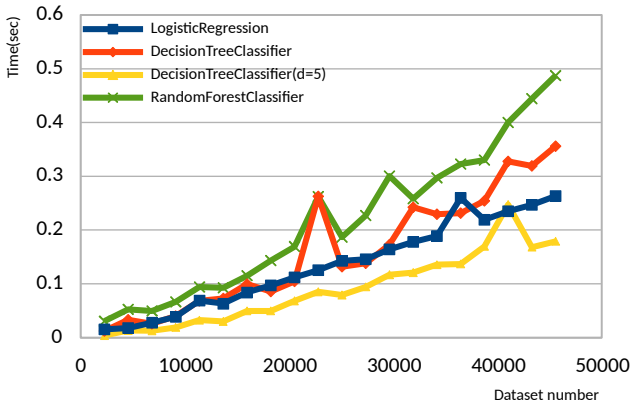

Fig. 4: Cat dataset number v.s. running time

2. Decision tree model result analysis
   The decision tree (d = 5) model of dog dataset and cat dataset were shown in Figure 5 and Figure 6. The node's color with darker blue indicated a higher probability of being adopted; darker orange indicated a lower probability of being adopted. We found that most blue nodes (high probability of being adopted) were altered, stay time longer than one day, and puppy/young.

- Top 2 adopted condition in cats' model:
  - Altered, stay time not <1 day, puppy, intake type not medical, breed not Manx cat.
  - Altered, stay time not <1 day, not puppy, young, intake type not medical.

- Top 2 adopted condition in dog' model:
  - Altered, stay time not <1 day, puppy, breed not Rottweiler dog, intake type not medical.
  - Altered, stay time not <1 day, not puppy, young, intake type not medical.

VI. CONCLUSIONS

In our analyses, four methods provided the prediction with more than 80% of accuracy rate for both dogs' and cats' outcome. This information can help shelter to quickly identify which animal needs more attention because it has less chance to be adopted. We also found in both cats and dogs the "Reproductive Status", "Stay Time", and "Age" are the most discriminatory features when building the decision tree. The altered animals would have higher chance to be adopted. In the future, we can analyze each feature related to the final outcomes to provide more information for shelter to improve and increase the adoption rate.

Attachment: attribute table

Attribute Table

Attribute	DOG		CAT	
x(0)	Stay Time	1-0.5y	Stay Time	1-0.5y
x(1)		<1 day		<1 day
x(2)		>2y		0.5-2y
x(3)		0.5-2y		>2y
x(4)		unknown		unknown
x(5)	In Take Type	STREET	In Take Type	STREET
x(6)		OTHER		ABANDON
x(7)		ABANDON		OTHER
x(8)		MEDICAL		DEAD
x(9)		DEAD		TEMPORARY
x(10)		TAKEOVER		MEDICAL
x(11)		TEMPORARY		TAKEOVER
x(12)	1st Color	BLACK	1st Color	BLACK
x(13)		TRICOLOR		WHITE
x(14)		BROWN		TORTIE
x(15)		BLUE		GRAY
x(16)		TAN		ORANGE
x(17)		CHOCOLATE		CHOCOLATE
x(18)		WHITE		BUFF
x(19)		RED		BROWN
x(20)		GOLD		YELLOW
x(21)		GRAY		CALICO
x(22)		SABLE		SEAL
x(23)		YELLOW		TAN
x(24)		APRICOT		CREAM
x(25)		FAWN		TRICOLOR
x(26)		OTHER		UNKNOWN
x(27)		CREAM		OTHER
x(28)		SILVER		FLAME
x(29)		BUFF		SILVER
x(30)		UNKNOWN		BLUE
x(31)	Color Mark	no_mark		LYNX
x(32)		mark	Color Mark	no_mark
x(33)	1st Breed	UNKNOWN		mark
x(34)		BEAGLE	1st Breed	UNKNOWN
x(35)		PIT BULL TERRIER		DOMESTIC SHORTHAIR
x(36)		OTHER		PERSIAN
x(37)		AMERICAN PIT BULL TERRIER		DOMESTIC LONGHAIR
x(38)		LABRADOR RETRIEVER		DOMESTIC MEDIUMHAIR
x(39)		ROTTWEILER		SIAMESE
x(40)		GERMAN SHEPHERD DOG		AMERICAN SHORTHAIR
x(41)		CHOW CHOW		OTHER
x(42)		SHIH TZU		MAINE COON
x(43)		SIBERIAN HUSKY		MANX
x(44)		BORDER COLLIE		HIMALAYAN
x(45)		COCKER SPANIEL		SNOWSHOE
x(46)		JACK RUSS TER		BENGAL
x(47)		POODLE - MINIATURE		RUSSIAN BLUE
x(48)		GOLDEN RETRIEVER	2nd Breed	no
x(49)		BOXER		MIX
x(50)		POMERANIAN	Gender	FEMALE
x(51)		CHIHUAHUA - SMOOTH COATED		MALE
x(52)		AUSTRALIAN SHEPHERD		unknown
x(53)		DACHSHUND	2nd Color	no
x(54)		YORKSHIRE TERRIER		MIX
x(55)	2nd Breed	MIX	Age	UNKNOWN
x(56)		no		ADULT
x(57)	Gender	FEMALE		OLD
x(58)		MALE		YOUNG
x(59)		unknown		PUPPY
x(60)	2nd Color	MIX	Reproductive status	ALTERED
x(61)		no		FERTILE
x(62)		ADULT		unknown

Attribute Table

x(63)	Age	UNKNOWN	Asilomar Status	HEALTHY
x(64)		PUPPY		UNHEALTHY/UNTREATABLE
x(65)		OLD		TREATABLE/MANAGEABLE
x(66)		YOUNG		unknown
x(67)	Reproductive status	FERTILE		
x(68)		ALTERED		
x(69)		unknown		
x(70)	Asilomar Status	HEALTHY		
x(71)		UNHEALTHY/UNTREATABLE		
x(72)		TREATABLE/MANAGEABLE		
x(73)		unknown		

Fig5: DOG decision tree(d=5)

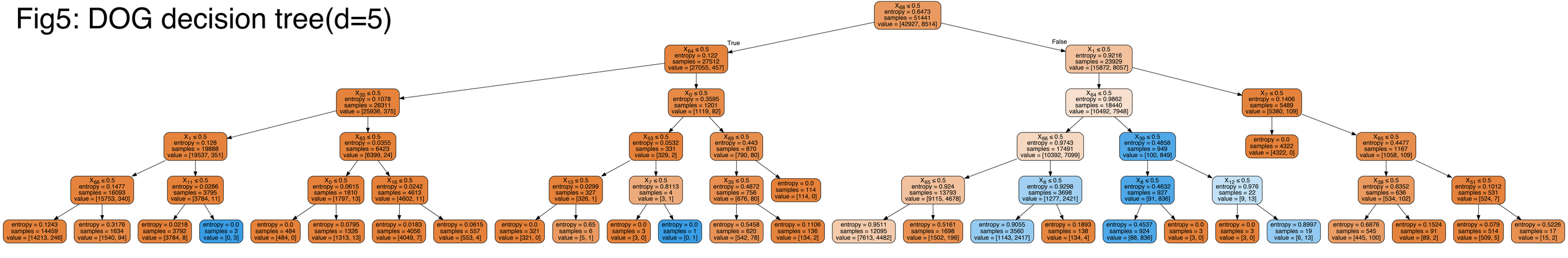

Fig6: CAT decision tree(d=5)