Data Science Programming

Robinson College of Business
2017 MSA

Shelter Animal Outcome Prediction

--Hanchen Huang, Xiaotong Li, Zichao Wang, Qiyao Wu, Xuan Han

Objective

Approximately millions of companion animals would be captured and kept at U.S. animal shelters for various reasons. Those shelter animals would eventually face following outcomes: adopted, transferred between shelters and foster homes, returned to their owners and euthanized/died. Our objective of this project is to create several machine learning models, such as decision tree, random forest, and logistic regression, etc. to predict the outcome of shelter animals. The reason of creating various models is to select the best model with most accuracy. We deeply expect that our model could make shelter and volunteer to spend more time on those animals that tend to be euthanized.

Data Source

The dataset "Shelter Animal Outcome Dataset" we used for this project was downloaded from Official City of Austin open data portal. This dataset contains data that collected by Austin Shelter in the past 3 years. The latest update of this dataset is in November, 2016. The dataset not only contains animal outcomes, but also contains useful information such as sex, age, breed and color. In this dataset, the dependent variable for this project are the animal's outcomes (adopted, transferred or euthanized) and the independent variables are animal types, sex, age, breed and color.

Data Cleaning

We started our project with cleaning data in Excel. The original dataset contains more than 57,000 data. Some of the data are null. We deleted all the rolls that contain null value. The dataset also contains shelter information of other animal types such as bat, bird, and livestock. Since we only want to focus on the outcome of cat and dog, we used Excel filter feature filtered out all rolls contains animal type that other than dog and cat. After filtering, the dataset contains total number of 54608 data. The next step we took is to convert all categorical data (breed, color, etc.) into numerical data. We assigned 0 to cat and 1 to dog. We also divided all shelter outcomes into 3 types, 0 (Adoption, return to Owner); 1 (Died, Disposal, Euthanasia); 2 (Transfer). After cleaning data, we uploaded the dataset on Jupyter notebook for later use.

Data Analysis and Interpretation

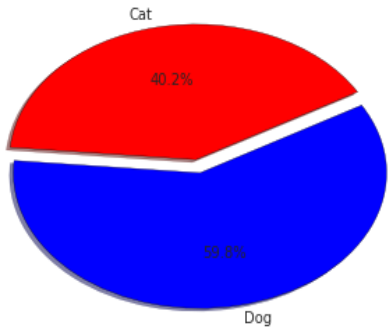


Figure 1

This dataset has many animal types, but we only consider two animal types for this project: cat (0) and dog (1).

Figure 1 displays the distribution of dogs and cats. The amount of cat and dog are 21959 and 32648 respectively.

The output graphs of the dataset presents the distribution of age upon outcome, sex upon outcome, animal color, breed type and outcome types. We categorized all outcomes into three classes which are 0: adoption or return to owner; 1: died, disposal or euthanasia; and 2: transfer to other shelters or foster homes.

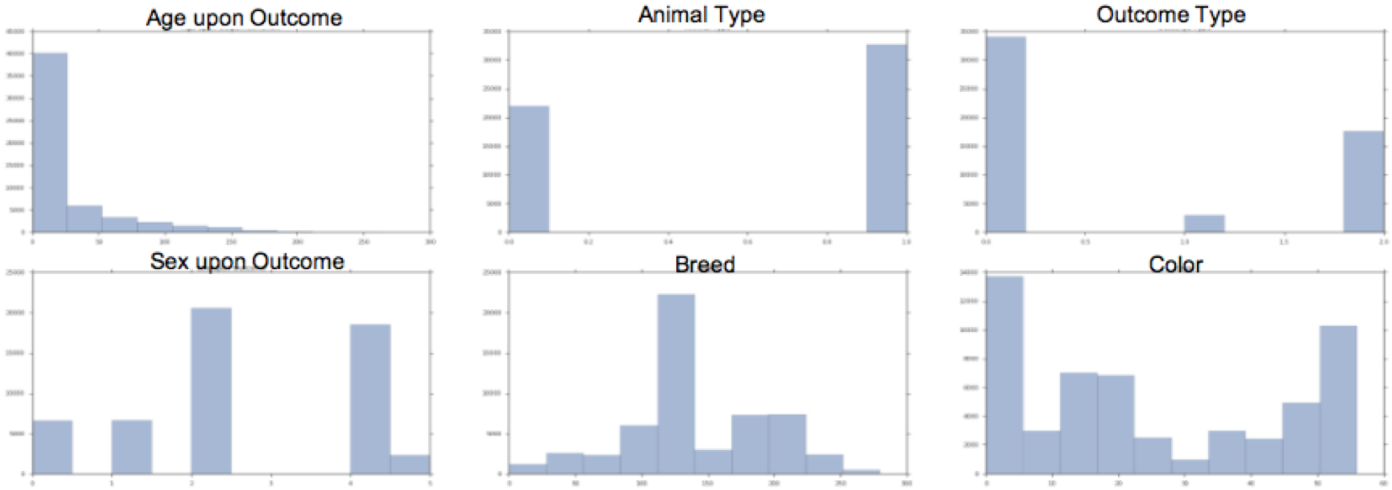


Figure 2

The first graph of Figure 2 is the age distribution of all cats/dogs in the shelter. From the graph, most animals that kept at Austin Shelter were at young age. The second graph shows the distribution of dogs and cats in Austin shelter. The third graph reveals the outcomes of shelter animals were adopted, disposed or euthanized/died. The fourth graph indicates that most animals in this shelter had been fixed (neutered/spayed). The last two graphs shows the distribution of animals breed type and color.

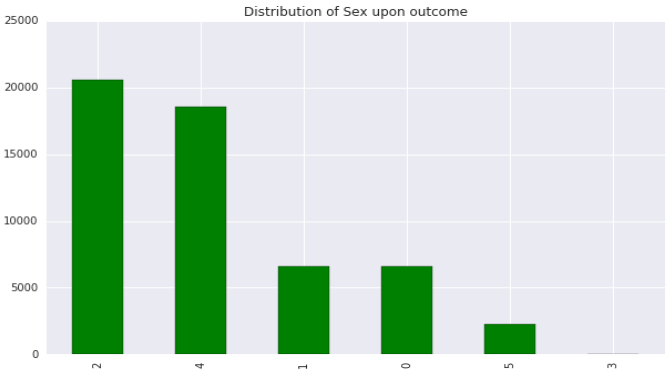


Figure 3

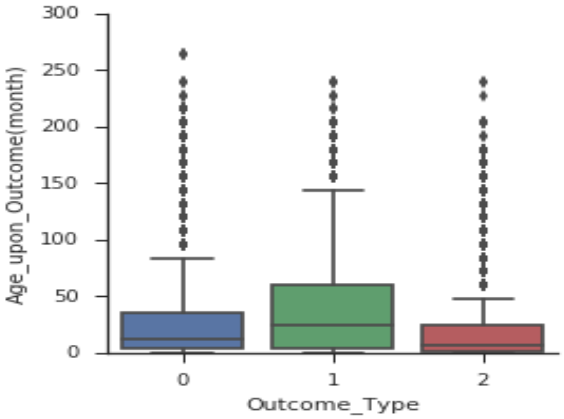


Figure 4

Figure 3 coincides with the fourth graph of Figure 2, the most animals in this shelter were fixed animals.
Figure 4 also coincides with the first graph of Figure 2, the most animals in this shelter were at young age.

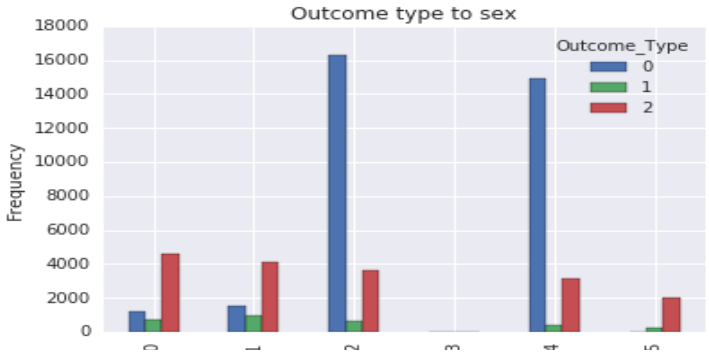


Figure 5 shows that the major reason that animal left the shelter was adoption or transferred back to owner.

Figure 5

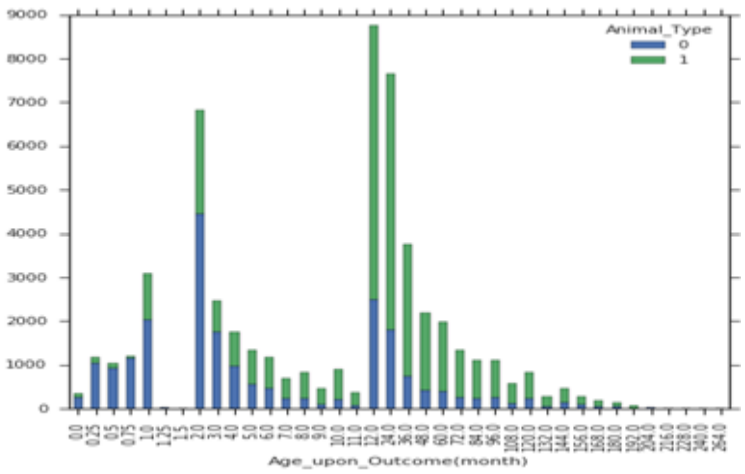


Figure 6

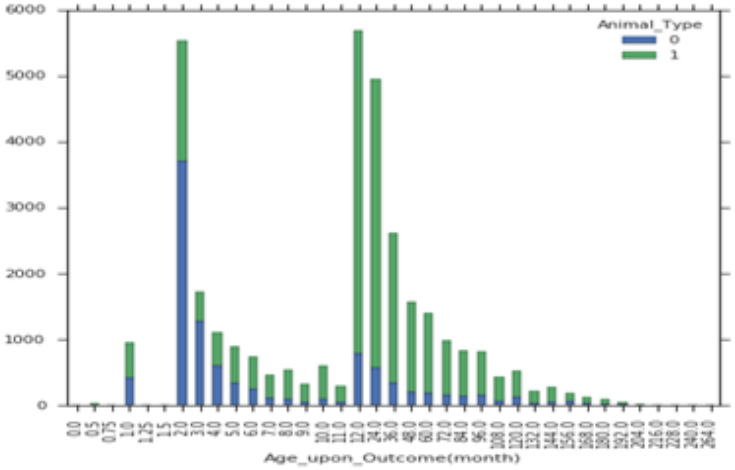


Figure 7

Figure 6 is the age distribution of cat and dog in the shelter and figure 7 shows the age distribution of cat and dog that were adopted.

Machine Learning Models

We did the following three models in both 3 class types and 2 class types, and wanted to compare the outcome between these 2 types.

Logistic Regression: measures the relationship between the categorical dependent variable and one or more independent variables by estimating probabilities using a logistic function, which is the cumulative logistic distribution. Because we have 3 different values in our Y, so we use multinomial logistic regression. And for better performance, we did robust scaling before running the regression, and use "sag" as the solver for faster convergence.

Table below are the confusion matrices:

Multinomial Logistic Regression,3 Class Type		Actual		
		0	1	2
Predict	0	5813	2	935
	1	404	11	204
	2	2119	3	1431

Binary Logistic Regression,2 Class Type		Actual	
		0	1
Predict	0	6804	12
	1	560	28

Coefficient for Two Class Types Logistic Regression:

	Animal_Type	Sex_upon_Outcome	Age_upon_Outcome_month	Breed	Color
Coefficient	-1.1489275	-0.71967593	0.01079331	0.00135937	-0.00197791

Random Forest: is an ensemble learning method for classification that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set.

Table below are the confusion matrices:

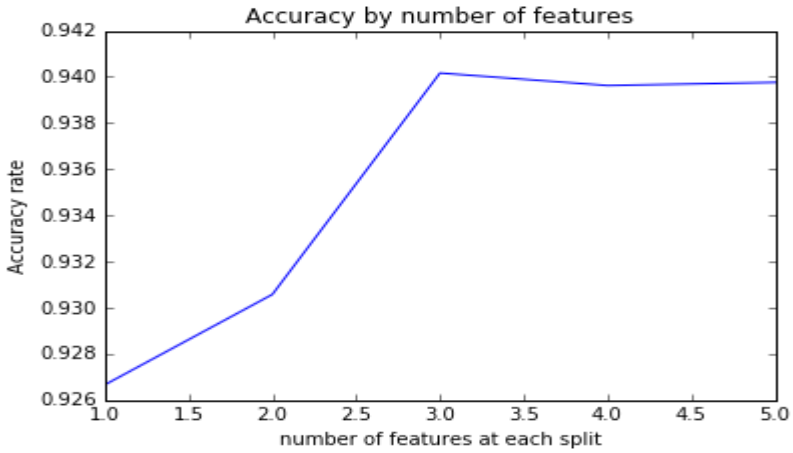
Random Forest, 3 Class Type		Actual		
		0	1	2
Predict	0	6090	72	588
	1	278	50	291
	2	1394	77	2082

Random Forest, 2 Class Type		Actual	
		0	1
Predict	0	6670	146
	1	362	226

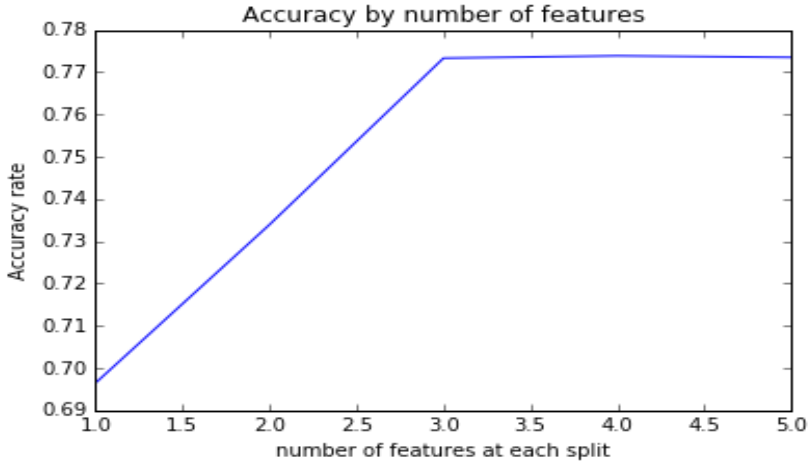
Decision Tree: is a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features.

First we want to determine the optimal number of features at each split.

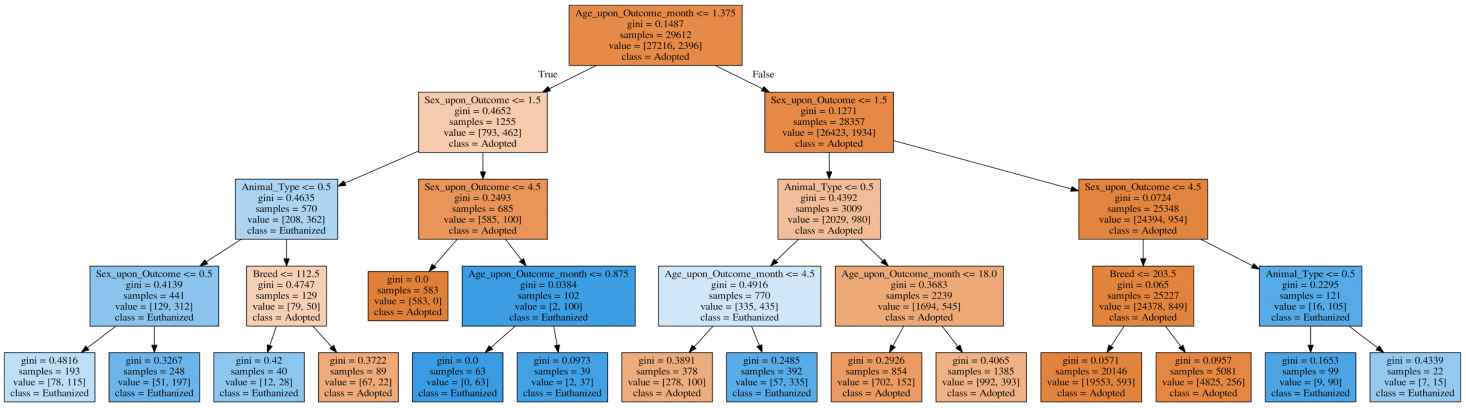
Two class types:



Three class types:



Tree for Two Class Types:



Tree for Three Class Types:

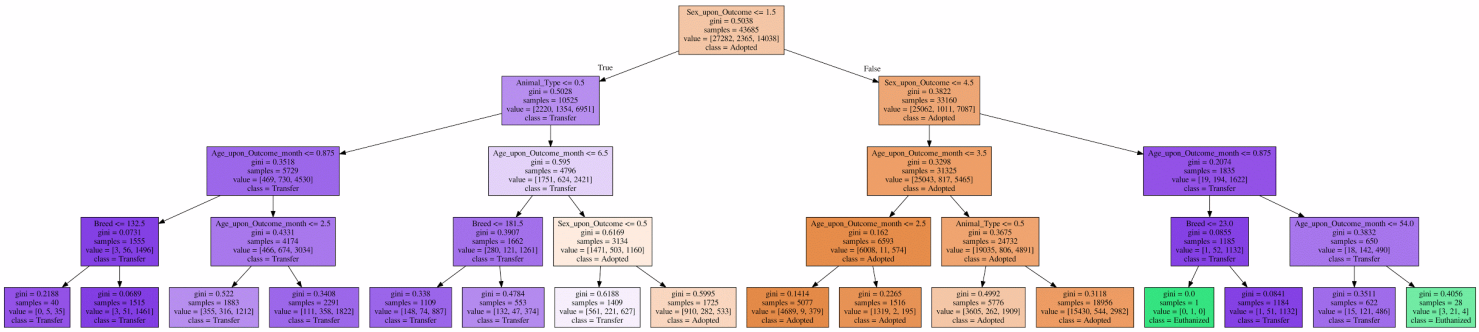


Table below are the confusion matrices:

Decision Tree, 3 Class Type		Actual		
		0	1	2
Predict	0	6425	0	325
	1	290	6	323
	2	1537	2	2014

Decision Tree, 2 Class Type		Actual	
		0	1
Predict	0	6761	55
	1	388	200

<u>K-Nearest Neighbor (KNN):</u> We imported the cleaned and prepared data set and let the Y be the 'Outcome_Type' and let X be ['Animal_Type','Sex_upon_Outcome','Breed','Age_upon_Outcome','Color']. After that, we use preprocessing.scale() to standardize the dataset along X and center to the mean and component wise scale to unit variance. We use cross-validation strategies and expose a train_test_split method which accepts the input dataset to be split and yields the train/test set indices for each iteration of the chosen cross-validation strategy. We define clf as neighbors.KNeighborsClassifier() and let the x_train,y_train fit into the classifier. We get an accuracy as 0.75. Besides that we also do a classification report for the precision and recall.

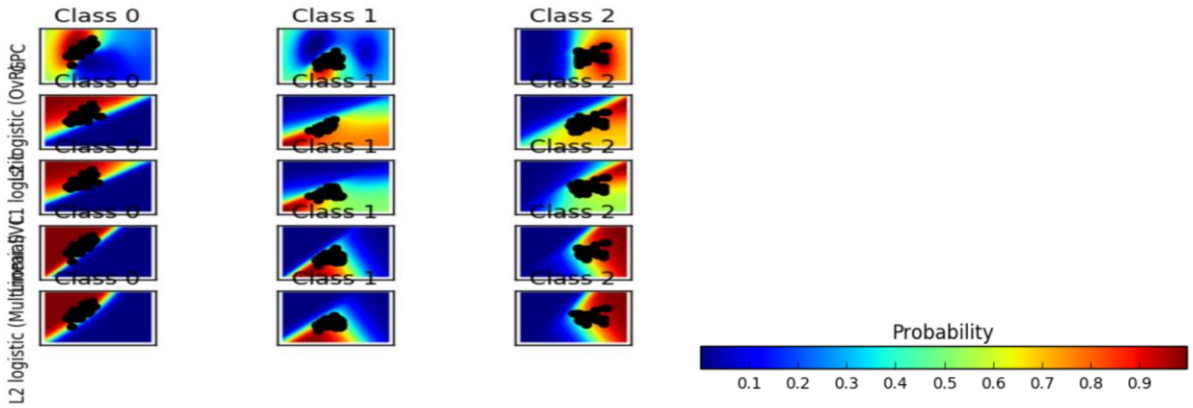
	precision	recall	f1-score	support
0	0.78	0.91	0.84	6776
1	0.22	0.06	0.10	614
2	0.71	0.58	0.64	3532
avg / total	0.73	0.75	0.73	10922

0.754165903681

We find that for type 0 we get the most accuracy and for type 1 we get the lowest. Type 0 is defined as Adopted and type 1 is defined as euthanasia.

Support Vector Classification (SVC): The last model we tried is SVC, we tried this model for self learning purpose. SVC is a supervised learning model with associated learning algorithms that analyze data used for classification and regression analysis (plot classification probability). We choose this model is not only because it has strong ability to process the dataset that contains more than 10,000(our dataset has more than 54,000 data) but also because it will retur 'best fit' hyperplane that divide and organize the data. Our model is able to handle multiple class classification according to One Vs. Rest. The graph below is our SVC model based on X(Animal_Type, Age_Upon_outcomes) and Y(Animal_Outcomes)

```
classif_rate for GPC : 82.666667
classif_rate for L2 logistic (OvR) : 76.666667
classif_rate for L1 logistic : 79.333333
classif_rate for Linear SVC : 82.000000
classif_rate for L2 logistic (Multinomial) : 82.000000
```



The model indicates that among all five classifiers, GPC classifier has the highest probability of success, which is 82.666667%. Each subplot represent all possible permutation and combination of X, Y under the particular class and classification.

Limitation of the model:
After we got the result from comparing Animal_Type, Age_Upon_outcomes with Animal_Outcomes, we also tried to compare other features(breed, sex) with outcomes. We tried to run the model for several times, but resulted the memory error every time. After realizing this issue, we tried to run our original model again, but also got memory error issue. We tried to fix the issue by sliding code into segments and run code one by one, but the issue persisted. Unfortunately, we are still in the search of solution to this issue.

Results and Conclusion:

After we got the results from Logistic Regression, Random Forest and Decision Tree in both 3 class types and 2 class types and the results from SVC and KNN, we want compare the precision and recall for each model. The detail classification reports are shown below:

Multinomial Logistic Regression:

	precision	recall	f1-score	support
0	0.70	0.86	0.77	6750
1	0.69	0.02	0.03	619
2	0.56	0.40	0.47	3553
avg / total	0.65	0.66	0.63	10922

Binary Logistic Regression:

	precision	recall	f1-score	support
0	0.92	1.00	0.96	6816
1	0.70	0.05	0.09	588
avg / total	0.91	0.92	0.89	7404

Decision Tree:

	precision	recall	f1-score	support
0	0.78	0.95	0.86	6750
1	0.75	0.01	0.02	619
2	0.76	0.57	0.65	3553
avg / total	0.77	0.77	0.74	10922

Decision Tree:

	precision	recall	f1-score	support
0	0.94	0.99	0.97	6816
1	0.79	0.33	0.46	588
avg / total	0.93	0.94	0.93	7404

Random Forest:

	precision	recall	f1-score	support
0	0.78	0.90	0.84	6750
1	0.25	0.08	0.12	619
2	0.70	0.59	0.64	3553

Random Forest:

	precision	recall	f1-score	support
0	0.95	0.98	0.96	6816
1	0.61	0.38	0.47	588

SVC (3 CLASS):

	precision	recall	f1-score	support
0	0.78	0.94	0.86	6776
1	0.54	0.01	0.02	614
2	0.75	0.58	0.66	3532
avg / total	0.76	0.77	0.74	10922

0.773850943051

KNN(3 CLASS):

	precision	recall	f1-score	support
0	0.78	0.91	0.84	6776
1	0.22	0.06	0.10	614
2	0.71	0.58	0.64	3532
avg / total	0.73	0.75	0.73	10922

0.754165903681

We can see that in general Two Class Types outperformed Three Class Types. One possible explanation for this fact is that every pet in class 2, which are pets being transferred to other shelters, will eventually end up in class 0 or class1(adopted or euthanized). When we perform a 3 class classification, we actually dealing with a class that a composed by the other two classes, so the accuracy was negatively influenced. Among the 5 Machine Learning Method, Decision Tree model yielded the best performance in prediction precision and recall, especially in predicting whether an animal will be euthanized. Therefore, we recommend Decision Tree model in predicting the outcome of pet adoption in animal shelters.

In addition to Machine Learning Model, we also discovered some interesting findings from our result.

1. According to the coefficients we got from logistic regression, it turns out that Animal Type and Sex_upon_Outcome have more influence than other variables on whether an animal will be adopted from shelter.
2. In general, Dogs have a better chance to be adopted than Cats.
3. Neutered or spayed animals are more likely to be adopted, and if also take sex into account the likelihood of adoption is as follows: Spayed Female > Neutered Male> Intact Male> Intact Female.
4. Age, Breed and Color have a very small effect on adoption, which is a little contradicted to our common sense.

Jupyter Notebook:

https://131.96.197.204:8000/user/xhan3/tree/DataScienceProgramming/Final_Datascienceprogramming

Reference:

http://scikit-learn.org/stable/auto_examples/classification/plot_classification_probability.html
https://data.austintexas.gov/Health/Austin-Animal-Center-Outcomes/9t4d-g238
http://scikit-learn.org/stable/auto_examples/model_selection/plot_precision_recall.html
http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html
http://scikit-learn.org/stable/modules/generated/sklearn.metrics.classification_report.html