Shelter Animal Outcome Predicting

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

Project Description

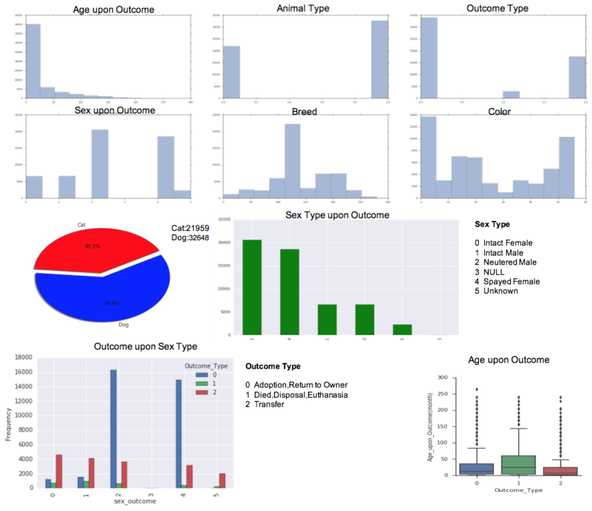
Project Goals

- The purpose of this project is to select the best model from various machine learning models such as decision tree, random forest,etc. to predict the outcome of shelter animals (Dog & Cat) based on information from Austin Animal Center Public Adoption Information.
- The dependent (target) variable for this project is the animal's outcome(Adopted, Transferring,Return to Owners, Died, Disposal or Euthanasia)
- The independent variables are animal types, sex, age, breed and color.

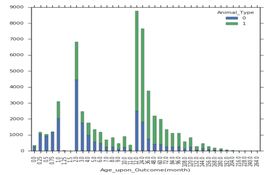
Data Exploration

- This dataset named "Shelter Animal Outcome Dataset" can be downloaded from Official City of Austin open data portal:
  https://data.austintexas.gov/Health/Austin-Animal-Center-Outcomes/9t4d-g238
- This dataset has 54076 instances with three outcomes and 37016 instances with two outcomes and 5 independent variables which were extracted from Official City of Austin dataset.

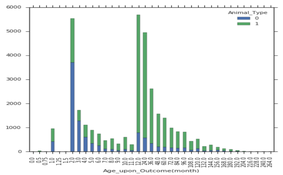
Data Distribution



Data Distribution



According to the graph Age_upon_outcome, most dogs in shelters are 12 months, 24 months or 2 months old. Most cats in animal shelters are 2 months, 12 months or 1 month old.



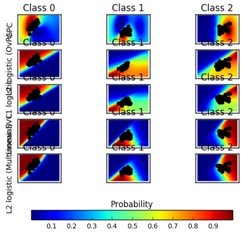
The graph above is the age distribution of animals adopted, most adopted cats are 2 months, 3 months or 12 months. Most adopted Dogs are adopted at age of 1 month, 12 months and 24 months.

Machine Learning Models

- **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.
- **Random forests** 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.
- **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.
- **k-Nearest Neighbor(KNN) is** K nearest neighbors is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure.
- **SVC** is supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis (plot classification probability).

- This dataset is a multi-class case, so we use "sag" to regularize the logistic regression.
- **Keycode: LogisticRegression(c=5,solver=' ', max_iter=1000, multi_class='multinomial', random_state=42)**

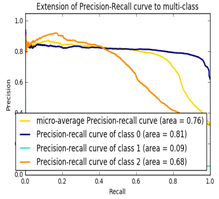
```
classif_rate for GPC : 82.666667
classif_rate for L2 logistic (OvR) : 76.666667
classif_rate for L1 logistic : 79.333333
classif_rate for Linear SVC : 82.000000
classif_rate for L2 logistic (Multinomial) : 82.000000
```



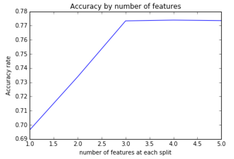
SVC report

	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

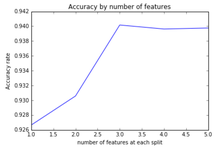


Random Forest: Three Class Types



VS

Random Forest: Two Class Types

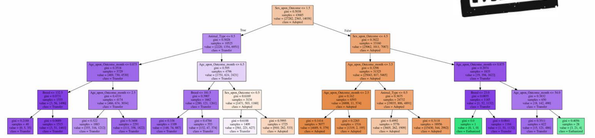


KNN report

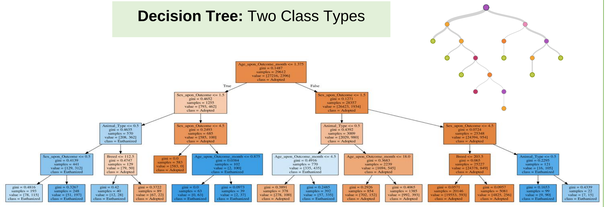
	precision	recall	f1-score	support
0	0.78	0.91	0.84	6776
1	0.22	0.06	0.10	614
2	0.71	0.58	0.64	3532
avg / total	0.73	0.75	0.73	10922

0.754169903681

Decision Tree: Three Class Types



Decision Tree: Two Class Types



Logistic Regression: Two Types Class

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

Comparison Among Different Methods: Three Class Type

VS

Comparison Among Different Methods: Two Class Type

```
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

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

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

avg / total       0.73      0.75      0.73     10922
```

$$\text{Precision} = \frac{tp}{tp + fp}$$

$$\text{Recall} = \frac{tp}{tp + fn}$$

$$F1 = \frac{2P}{2TP + FP + FN}$$

```
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       0.95      0.99      0.97      6816
          1       0.78      0.34      0.47       588

avg / total       0.93      0.94      0.93      7404

Random Forest:
             precision    recall  f1-score   support

          0       0.95      0.98      0.96      6816
          1       0.61      0.38      0.47       588

avg / total       0.92      0.93      0.92      7404
```

Conclusions

- After comparison the result between different Machine Learning Method, 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.
- In general, Dogs have a better chance to be adopted than cat.
- 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.
- Age, Breed and Color have a very small effect on adoption, which is a little contradicted to our common sense.

Georgia State University | J. MACK ROBINSON COLLEGE OF BUSINESS