# Syracuse University

#### School of Information Studies

# IST652 – Scripting for Data Analysis

## Final Project Report

### Instructor: Professor Gregory Block

### Brandon Croarkin

#### Introduction

Every year, approximately 7.6 million animals end up in US shelters. Many animals are given up as unwanted by their owners, while others are picked up after getting lost or taken out of cruelty situations. Many of these animals find forever families to take them home, but many are not so lucky. 2.7 million dogs and cats are euthanized in the US every year.

Using data sourced from Kaggle, this report will detail some of the factors that are predictive of shelter outcomes and attempt to predict these outcomes using the data. With this knowledge, shelters can help optimize how they allocate their resources to help promote animals that are less likely to be adopted and can know where to focus their attention to maximize the number of animals that are able to find a home. Doing so could help make sure that animal shelters are able to find homes for as many animals as possible.

#### Data

The data for this report was found on [Kaggle](https://www.kaggle.com/c/shelter-animal-outcomes) but is originally sourced from Austin Animal Center. It contains data from October 2013 to March 2016. It contains intake information on the breed, color, sex, and age of the animals as well as their shelter outcomes. These outcomes include: Adoption, Died, Euthanasia, Return to owner, and Transfer. The data is split between a training and test set. The training set contains 26729 rows and 10 columns. These columns are shown below along with what the 5 outcome types are the animals types that are in the data.

* AnimalID
* Name
* DateTime
* OutcomeType
  + Adoption
  + Euthanasia
  + Return\_to\_owner
  + Died
  + Transfer
* OutcomeSubtype
* AnimalType
  + Dog
  + Cat
* SexuponOutcome
* AgeuponOutcome
* Breed
* Color

To improve the predictive abilities of any models created from this data, employment data from Austin was merged with the shelter outcomes dataset. It contains data on the total nonfarm employees in the Austin-Round Rock, TX area from 2008 to 2018. This data was included to test a theory that more employment will lead to more adoptions as more employed individuals are able to afford a pet.

#### Methods of Analysis

My method of analysis attempted to follow the rough machine learning pipeline detailed below:

1. Obtaining the data
2. Cleaning the data
3. Exploratory data analysis (EDA)
4. Feature Engineering
5. Model Building
6. Interpretation

After finding the topic I was interested in, the first step was finding the data sources and reading the data into Python where it can be analyzed. Kaggle is a great source for machine learning data sets and thus was one of the first locations I used to try and find the data. The data they had available fit my research topic and was readily available for download in a CSV format. To expand on this dataset, I wanted to find a secondary dataset that could complement this. The Austin employment data seemed like a good choice as it went with my hypothesis that more adoptions would happen when employment is strong, as there would be more people willing to adopt. BLS has this employment data available and I was able to download it as an Excel file to match the years covered by the Austin Shelter Outcome data.

Once these data sources were obtained and read into Python, the next step was cleaning and preparing the data so that it is ready for analysis. The Austin Shelter Outcome dataset came with missing data that had to be dealt with. The Austin employment data did not have any missing data but had to be re-formatted from a wide to a long structure and then further manipulated to be able to merge it with the Austin Shelter data.

After the data is cleaned and prepped, it can be analyzed to find interesting insight. This process of exploratory data analysis (EDA) mainly consisted of following basic intuition on where a pattern may exist and then visualizing these questions to test if the hypotheses I made were valid. The visualizations helped reinforce the findings of the data and make them easier to interpret.

From exploring the data, an understanding is made that assists with feature engineering, which is the process of creating additional features from the data that help fit a predictive model. Since a model is only as good as the data fed to it, this is an integral part of the pipeline. Without insight from a domain expert, this can sometimes be challenging though, which makes the EDA very integral to this pipeline.

Only after all these steps are completed did I attempt any model to create a prediction. I used a training and test set to evaluate the performance of the model. I tested a couple different models and went with the best performing model and adjusted a couple parameters to tune the best-performing model.

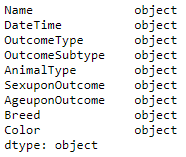
It is important to note that this is not a perfect linear pipeline though as many of the discoveries found during EDA and feature engineering lead to additional data cleaning and preparation.

#### Pre-Processing

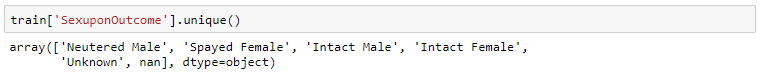
Before running any analysis on the data, there was pre-processing needed to clean the data, extract important features, and format the data so it is ready for analysis.

The first step was cleaning the data to remove unnecessary data, missing values, or any faulty values. The columns ‘AnimalID’ and ‘Name’ were removed for all analysis as they do not give any important information. Additionally, the column ‘OutcomeSubtype’ was removed for the predictions since its information would confound the results. The only outcome types that were kept for prediction were ‘Euthanasia’ and ‘Adoption’ as the other three categories are more influenced by outside factors and would likely result in worse results. Since the focus of this analysis is what leads to an animal being adopted or not being adopted, it was easiest to remove these extraneous factors for the predictions. They are included for the rest of the analysis though.

Additionally, the data types of many of the fields had to be updated for proper analysis. When the data was read in, all columns had a data type of “object” (see image below). Some of these had to be updated to properly perform the analyses. The ‘DateTime’ field was converted to a datetime type. ‘OutcomeType’, ‘OutcomeSubtype’, ‘AnimalType’, and ‘Breed’ were all converted to categorical variables. The ‘AgeuponOutcome’ was converted to a floating point number, however, due to how the data came in this had to be converted via feature engineering.



Next, I went through feature engineering to extract more features from the data. First, the ‘SexuponOutcome’ had to be split to extract additional information. The original fields (shown below), had the neutered and gender information combined into one field. To get more variables that will provide additional information for data analysis, this was split into a new ‘Neutered’ and ‘Gender’ field by splitting the text. All fields labeled ‘Spayed’ were converted to ‘Neutered’ so that these would be comparable.

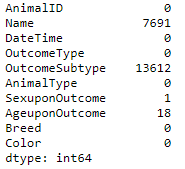


The ‘AgeuponOutome’ also had to be updated as it was all listed as strings (shown below). Since this information can’t be compared, they had to be converted to a day format. By splitting the string and multiplying the number using the ratios of 365 days in a year, 30 days in a month, and 7 days in a week they were all changed to an integer format for easy comparison.



Additional work to condense ‘Breed’ and ‘Color’ likely would have provided a boost in information that could have helped the model, but I lacked a good dictionary to condense the large range of values in both categories into a smaller list that could have provided more information. There were 1320 unique dog breeds listed just for dogs! To condense some of this information, breeds and colors that had less than 50 occurrences were converted to an ‘Other’ category to remove some of the information for more focused analysis.

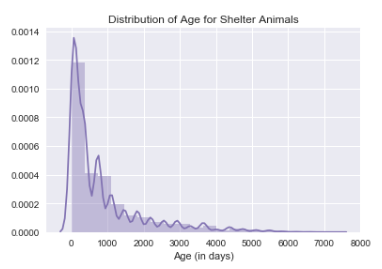
Finally, missing data had to be handled. After the feature engineering some missing fields emerged. ‘Name’ and ‘AnimalID’ were removed for analysis purposes. Since ‘OutcomeSubtype’ was only used for some EDA, the missing values were changed to ‘Unknown’ rather than deleting them. The rows with the missing values in ‘SexuponOutcome’ and ‘AgeuponOutcome’ were dropped so they did not affect any of the analysis. They represented very small amounts of data so I did not decide to impute their values.



#### Questions Answered

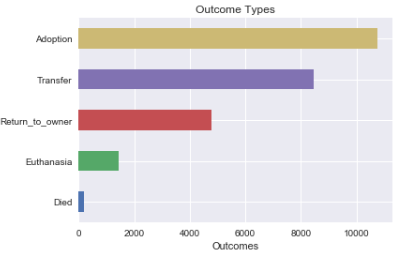
Once the data was cleaned and the additional features were created, I proceeded by making a list of questions that I had of the data and created data visuals that helped answer these questions to gain insight from the data.

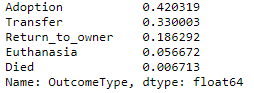
1. What is the distribution of ages for shelter animals?



There is a wide distribution of ages for the shelter animals. The distribution has right-skewed as there is a very long tail consisting of outlier ages. However, all these ages do seem within the realm of possibility and do not seem like bad data. The average age in the dataset was 819 days, which is equivalent to around 2.2 years old.

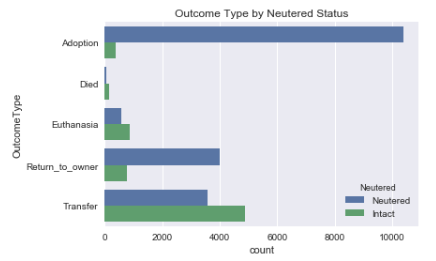
1. What are the different outcome types for shelter animals and how often do they occur?





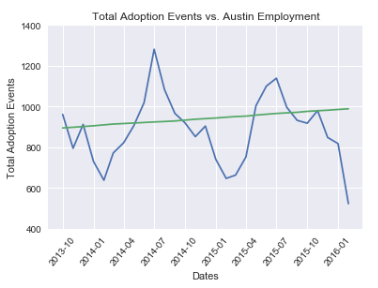
Adoption is the most common result for animals, so it is the plurality result but it does not constitute a majority as it is less than 50%. It is also interesting that transfer is the second most common result. Euthanasia only consists of around 6% of the total, so it is not a very common event.

1. Is there a difference in outcome type based on the neutered status of the animal?



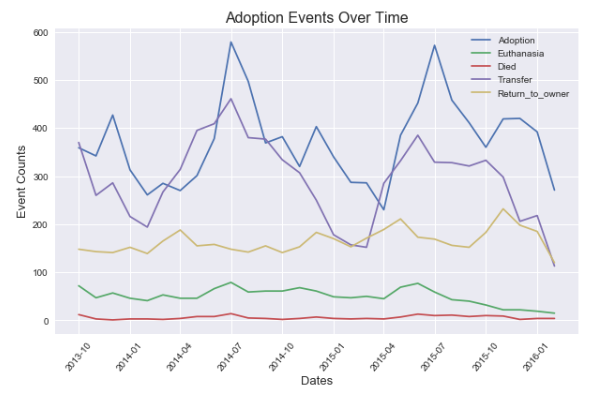
There is a very stark difference in adoption status for animals that are neutered vs. those that are not. This is a very interesting finding and could be strong evidence for neutering animals that do end up in shelters to improve their likelihood of being adopted.

1. Is there a relationship between adoption events and employment in Austin?



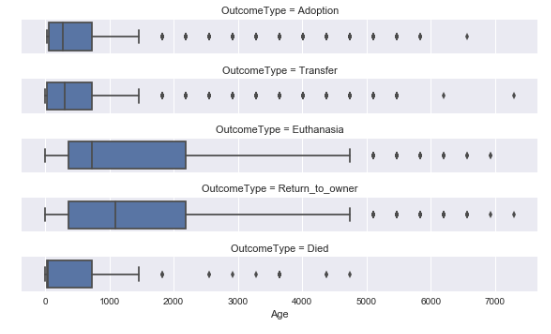
One of the reasons that I included Austin employment numbers is because I had a hypothesis that more employment would lead to more adoptions. While it is not clearly evident, there does appear to be a slight correlation to the two numbers, although the adoptions number have a much more cyclical nature.

1. What is the trend in adoption events over time?



This chart again highlights the cyclical nature of adoption events, with adoptions consistently spiking in July of each year. Knowing this is important and shows us that month can be an influential factor in predicting adoptions. There also appears to be a downward trend in euthanasia that begins around June 2016 and continues through the end of the data.

1. How do outcomes differ across animal ages?

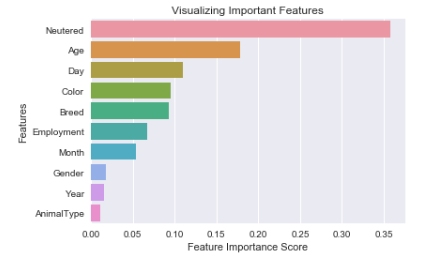


An initial hypothesis coming into this was that older animals would be more likely to not be adopted, as people tend to prefer younger animals. The plot shown above emphasizes this point. The age distribution of euthanized animals has a much higher mean than those that are adopted. This plot confirms that younger animals are more likely to be adopted.

1. How accurately can an animal’s adoption status be predicted?

Using a random forest model, I was able to predict animal outcomes (after filtering down to Adoption and Euthanasia) with 93% accuracy.

1. What are the most important features in determining an animal’s adoption status?



The plot above helps confirm some of the findings reach earlier. Neutered status and age are the two best variables for predicting whether an animal gets adopted or not. Knowing this can help shelters promote neutering for their animals and to use methods to help older animals get adopted.

#### Conclusion

This project has provided several key insights on Shelter Animal Outcomes in Austin, TX. While, it cannot be proven without additional data whether this data would apply similarly to all shelters across the United States, these insights could likely help others shelters form their own hypotheses to further explore to help ensure that as many animals as possible are able to find a home.

Knowledge such as the large impact of age and neutered status of animals can help drive actions to promote shelters across the US. However, there is still a lot of additional analysis that can be explored from this data. A good data dictionary of dog breeds could help condense this information into smaller dog and cat breed categories that would give insight on what breeds are the most/least likely to get adopted.

There was also a trend that appeared to be emerged towards less euthanasia. A longer time series of data would help discover whether this is a longer forming trend, or if the small trend spotted was merely an anomaly. Additionally, the time of the year and day of the week also played a large impact on adoptions. This can be used to help shelters allocate resources towards these peak adoption times or allow them to promote more during the less busy times of the year/week/month.

The predictive analysis conducted also provided good accuracy that could be used to focus at a smaller scale on specific animals that are a higher risk of not settling into a home. The 93% accuracy recorded was very high, although this accuracy would be skewed when additional outside factors like transfers, returning to owners, and natural deaths are included.

Some of the potential next steps to this project would involve collecting more data across more years and shelters in the US to follow up on these insights. Additional data dictionaries on dog and cat breeds could also be used to find additional insights. While this project has shed several important insights, it can still be improved to further improve the status and outcomes of shelter animals across the US.