Using Predictive Analytics to Reduce Length of Stay and Identify Adoption Channels for Animals in a Shelter

# Introduction

## Background

According to research estimates from the American Society for the Prevention of Cruelty to Animals (ASPCA), over 6.5 million pets enter shelters each year [1]. Of those, approximately 1.5 million animals are euthanized [1]. Despite the extraordinary efforts of the animal welfare community, euthanasia is often unavoidable, especially in areas of the South where many pet owners have not adopted the spay and neuter practice for their pets.

## Problem Statement

Both municipal and privately owned shelters are always looking for and creating programs to move their adoptable animals faster. Depending on the program, additional resources may be needed to keep them running in an industry where resources, specifically financial, are already running low. Reducing the length of stay for animals in the shelter through adoption, as opposed to euthanasia, is a cost-effective solution that can make room for new animals coming in and optimize the shelter’s resources. Using characteristics about the animals to place them in the appropriate adoption program can increase their likelihood of adoption.

## Scope

The objective of this project is to determine if taking a closer look at shelter data and animal characteristics can not only predict the likelihood that an animal will be adopted out quickly, but also direct shelter managers to make subtle changes that will decrease the length of stay for animals and effectively route them through the best adoption channel.

## Document Overview

This project proposal will outline the key business objectives, preliminary requirements, expected results and plan for moving forward. As the data analysis progresses, it will serve as a reminder of project goals and keep the project on course.

# Preliminary Requirements

## Technical Approach

This project will follow the CRISP-DM recommended stages for a data science project including Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation and Deployment. Both R and Python will be used to manipulate the data, create the models and evaluate success.

## Resources

### Domain Experts

Alicia Vial, Director of Communications, and Ana Zorrilla, Chief Executive Officer, at the Louisiana SPCA will be consulted as needed with questions throughout the course of the project.

### Data Sources

The data used to conduct this analysis will come from the Louisiana SPCA located in Algiers, Louisiana. Input features will be taken from animals whose intake date was between January 1, 2018 and December 31, 2019. The evaluation time needed to determine the length of stay for animals that came in near with end of the input time range is between January 1, 2020 and February 29, 2020. This cutoff date was chosen to avoid any inconsistencies related to changes to operations made in response to the COVID pandemic.

Multiple data files were retrieved from their shelter operation software PetPoint. A unique animal identifier exists to join the following files:

* Animal by Intake – Intake data for animals that entered the shelter from 1/1/2018 through 12/31/2019.
* Animal by Outcome– Outcome data for animals with intake dates between 1/1/2018 through 12/31/2019. The target evaluation date for adoption will be 2/29/2020.
* Animal Location History – Location information for animals with intake dates between 1/1/2018 through 12/31/2019 at their time of release before the target evaluation date of 2/29/2020.
* Animal Medical History – Medical information for animals with intake dates between 1/1/2018 through 12/31/2019.
* Animal Memo History – Memos taken for animals between 1/1/2018 through 12/31/2019. Notes are contained in this file and can reveal information regarding training and observed behavior.

## Analysis

Analysis will begin by combining the datasets for animals in the intake range of 1/1/2018 through 12/31/2019 and filtering to include the outcome of adopted. Single variable summary statistics and visualizations will be used to describe the data and make determinations on the most relevant data. Multiple variable summary statistics and visualizations will be performed on the most interesting features. The data will be transformed as necessary and decisions will need to be made regarding the correction of outliers and missing data. Features will then be fed into preliminary algorithms to begin to reduce them to include the optimum features for prediction. A training and test dataset will be created from the original data to evaluate the effectiveness of the model.

Multiple target variables will be used: 1. A binary result of 1 for adopted within 7 days or 0 for adopted after 7 days. 2. A length of stay for each animal. The dataset will contain both continuous and categorical variables so linear regression will not likely be used as an algorithm unless the most important features are all continuous. Logistic regression is an ideal algorithm for quantitative and qualitative variables and will likely be the starting point for analysis.

## Requirements

The most important requirement related to securing the data is to receive permission from the CEO of the Louisiana SPCA, Ana Zorrilla, to use their animal data for the initial analysis which was granted. The approximate schedule required to complete the project on time is shown below based on start dates because many stages are iterative. A detailed project plan can be found in the execution and management section of this document.

|  |  |
| --- | --- |
| Stage | Start Date |
| Business Understanding | 6/8/2020 |
| Data Understanding | 6/8/2020 |
| Data Preparation | 6/15/2020 |
| Modeling | 6/15/2020 |
| Evaluation | 6/28/2020 |
| Deployment | 7/15/2020 |

## Assumptions and Risks

The analysis will proceed with the assumption that the data sources are accurate; of course, the risk of outliers and missing data is always possible and will be addressed during data preparation. Decisions will need to be made that involve making assumptions or taking a risk that can affect the model.

1. The data file contains all animals with an intake date in the input period of 1/1/2018 through 12/31/2019 regardless of the outcome. An algorithm could first be used including all outcomes to predict whether animals will be adopted to affect the routing of that animal through the system. However, because this study focuses on shortening the length of stay for animals available for adoption, the intake data will be filtered for animals with an outcome of adopted. If another study is completed to determine adoptability at intake, the entirety of the intake file can be used. For this same reason, other live release outcomes such as Return to Owner or Transfer will also be excluded.

2. Per the recommendation of Alicia Vial, puppies and kittens under six months of age will be excluded because often those animals are often routed through foster care for a long period of time until they are old enough and heavy enough to be spayed or neutered and placed for adoption.

3. Decisions made to correct missing values and outliers may affect the final model.

4. Depending on the intake type of the animal, a holding period is required before that animal is available for adoption. If the data does not contain a “date made available” for adoption separate from the intake date, then the domain experts will need to be consulted to apply standard holding periods for each intake type.

## Model Deployment

Deployment of the model would involve running the data for the currently available animals for adoption through the algorithm and evaluating the results over time to see if they are accurate. If controllable factors that reduce the length of stay are identified by the analysis, then modifications can be made at the time of placement to reduce the animal’s stay in the shelter. If trends are recognized based on animal characteristics and adoption channels, an animal can be routed to that program sooner.

## Testing and Evaluation

Initial testing will be completed using a testing dataset split from the initial data for the input period. Subsequent testing can be performed for animals with an intake period after 12/31/2019. The success of the model that predicts the likelihood that an animal is adopted within 7 days can be evaluated using accuracy if the target classes are evenly distributed, otherwise F1 score can be used. For the prediction of length of stay, continuous evaluation statistics such as mean-squared error and R2 can be used to determine success.

# Expected Results

Factors such as age, breed and size of the animal are expected to be large contributors to how quickly an animal is adopted. While identifying these benchmarks is helpful, they are uncontrollable by the animal organization and cannot be adjusted to increase the adoptability of the animal. I hope to uncover details that are controllable by the shelter that can be adjusted with little financial impact, such as:

* Kennel location
* Programs in which the animal participated such as foster, offsite adoption or commercial adoption partners
* Observations about animal behavior through notes made during assessment
* Medical details

I also hope to uncover trends that can lead the organization to route the animals appropriately throughout their adoption channels based on animal characteristics such as age or breed.

# Execution and Management

## Project Plan

The project plan including intermittent peer reviews is included below. This plan is fluid and will be reviewed weekly and adjusted as needed to include more detailed steps.

| # | Milestone | Task | subTask | Start Date | Due Date | Week |
| --- | --- | --- | --- | --- | --- | --- |
| **1** | **Rough project Plan** |  |  | 6/1/2020 | 6/7/2020 | **1** |
| **2** | **Data Selection & Project Proposal** |  |  | 6/8/2020 | 6/14/2020 | **2** |
| **2** |  | Chose Data Source |  | 6/8/2020 | 6/9/2020 | **2** |
| **2** |  | Proposal | Introduction | 6/9/2020 | 6/9/2020 | **2** |
| **2** |  |  | Preliminary Requirements | 6/10/2020 | 6/12/2020 | **2** |
| **2** |  |  | Expected Results | 6/12/2020 | 6/14/2020 | **2** |
| **2** |  |  | Execution & Management of Project | 6/12/2020 | 6/14/2020 | **2** |
|  | **Week 2 Peer Review** |  |  | 6/12/2020 | 6/14/2020 | **2** |
| **3** | **Preliminary Analysis** |  |  | 6/15/2020 | 6/28/2020 | **3** |
| **3** |  | Review Week 2 Peer Reviews | Update project plan as necessary | 6/15/2020 | 6/15/2020 | **3** |
| **3** |  | Data Understanding |  | 6/15/2020 | 6/21/2020 | **3** |
| **3** |  |  | Combine datasets | 6/15/2020 | 6/15/2020 | **3** |
| **3** |  |  | Summary Statistics for Single Variables | 6/15/2020 | 6/16/2020 | **3** |
| **3** |  |  | Single Variable Visualizations | 6/16/2020 | 6/17/2020 | **3** |
| **3** |  |  | Multivariate Summary Statistics | 6/17/2020 | 6/18/2020 | **3** |
| **3** |  |  | Multivariate Visualizations | 6/18/2020 | 6/19/2020 | **3** |
| **3** |  |  | Complete Data Audit Documentation | 6/19/2020 | 6/20/2020 | **3** |
|  | **Week 3 Peer Review** |  |  | 6/19/2020 | 6/21/2020 | **3** |
| **3** |  | Review Week 3 Peer Reviews | Update project plan as necessary | 6/22/2020 | 6/22/2020 | **4** |
| **3** |  | Project Paper |  | 6/22/2020 | 6/28/2020 | **4** |
| **3** |  |  | Intro/Background of the Problem | 6/22/2020 | 6/22/2020 | **4** |
| **3** |  |  | Methods | 6/22/2020 | 6/23/2020 | **4** |
| **3** |  |  | Preliminary Results | 6/23/2020 | 6/28/2020 | **4** |
| **3** |  |  | Discussion | 6/24/2020 | 6/28/2020 | **4** |
| **4** | **Project Presentation and Status** |  |  | 6/28/2020 | 7/19/2020 | **5** |
| **4** |  |  | Update project plan as necessary | 6/28/2020 | 6/28/2020 | **5** |
| **4** |  | Data Preparation | Handle Missing Values, Outliers and Perform Transformations | 6/28/2020 | 6/30/2020 | **5** |
| **4** |  | Modeling |  |  |  | **5** |
| **4** |  |  | Feature Selection | 6/30/2020 | 7/5/2020 | **5** |
| **4** |  |  | Model Selection | 6/30/2020 | 7/5/2020 | **5** |
| **4** |  | Evaluation | Evaluation | 6/28/2020 | 7/12/2020 | **5** |
|  | **Week 5 Peer Review** |  |  | 7/2/2020 | 7/5/2020 | **5** |
| **4** |  |  | Review Week 5 Peer Reviews | 7/6/2020 | 7/6/2020 | **6** |
| **4** |  |  | Update project plan as necessary | 7/6/2020 | 7/6/2020 | **6** |
| **4** |  |  | Update applicable sections of paper | 7/5/2020 | 7/12/2020 | **6** |
|  | **Week 6 Peer Review** |  |  | 7/9/2020 | 7/12/2020 | **6** |
| **4** |  |  | Review Week 6 Peer Reviews | 7/13/2020 | 7/13/2020 | **7** |
| **4** |  |  | Update project plan as necessary | 7/13/2020 | 7/13/2020 | **7** |
| **4** |  |  | Create presentation slides | 7/13/2020 | 7/19/2020 | **7** |
| **5** | **Final project paper and presentation** |  |  | 7/20/2020 | 8/8/2020 | **8** |
| **5** |  |  | Update project plan as necessary | 7/20/2020 | 7/19/2020 | **8** |
| **5** |  |  | Complete paper | 7/20/2020 | 7/26/2020 | **8** |
|  | **Week 8 Peer Review** |  |  | 7/23/2020 | 7/26/2020 | **8** |
| **5** |  |  | Review Week 8 Peer Reviews | 7/27/2020 | 7/27/2020 | **9** |
| **5** |  |  | Update project plan as necessary | 7/27/2020 | 7/27/2020 | **9** |
| **5** |  |  | Complete presentation slides | 7/27/2020 | 8/2/2020 | **9** |
| **5** |  |  | Record Presentation | 8/2/2020 | 8/8/2020 | **10** |
|  | **Week 10 Peer Review** |  |  | 8/5/2020 | 8/8/2020 | **10** |

## Project Risks

If the model techniques employed do not identify any features that impact the length of stay for an animal that are within the control of the organization to adjust, this could render this project irrelevant. For example, the only factors that impact length of stay are age, weight, and breed. In this event, the focus will need to shift to identifying the appropriate adoption channels for animals with those characteristics.

# References

[1] Pet Statistics | Shelter Intake and Surrender | ASPCA. (n.d.). Retrieved June 13, 2020, from   
https://www.aspca.org/animal-homelessness/shelter-intake-and-surrender/pet-statistics