# Dog Adoption Modeling

## Final Report

Daniel Rayl, Dieter Erben Vasconcelos, Joe Newhall & Lauren Weldon

ITAO-30220-01 Predictive Analytics

May 6, 2018

# Problem definition & pre-processing

## Introduction

The Austin Animal Center (AAC) runs the largest 'No Kill' municipal animal shelter in the United States, providing shelter to more than 18,000 animals each year and animal protection services to all of Austin and Travis County.

The AAC is an open-intake facility where lost and surrendered animals from all of Travis County in need of shelter are accepted regardless of age, health, species or breed. The mission of the AAC is to "place all adoptable animals in forever homes."


In fiscal year 2016 (FY16), the Austin Animal Center had a General Fund budget of $11.5 million which was allocated towards shelter services (57.7%), field services (13.4%), transfers (12%), support services (7.7%), prevention services (7.3 %) and other requirements (1.8%).[1] Each year in its proposed budget, the AAC outlines a number of unmet needs ranging from medical and behavioral programs to service and care workers. Like many shelters of all sizes around the country, the AAC faces unique challenges in resource allocation for the thousands of animals it serves as a non-profit organization.

While the majority of human resources are volunteer-based and non-technical, the center has strived to embrace data-informed decision-making by making its service data open source and available on Data.gov, which is managed and hosted by the U.S. General Services Administration's Technology Transformation Service. For this predictive analytics project, we chose to analyze the AAC's intake and outcome data from 2013 to present. We chose these datasets because we as a group love animals and wanted to explore opportunities for model implementation in non-profit organizations.

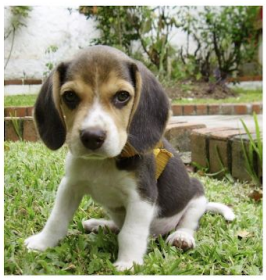
[1] Austin Animal Services FY 2016 Budget Overview

# Challenge

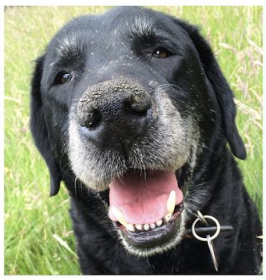
We framed our project around the following research question:

*Can we predict the length of time a dog will spend at Austin Animal Center between intake and adoption based on a historical dataset of intakes and outcomes?*

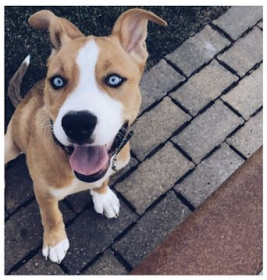
While the datasets included records for a variety of animal and outcome types, we chose to focus our research specifically around adoptable dogs. Our goal was to create a useful model that could be implemented to help shelters better serve dogs upon intake. For contextual framing, we created four case study dogs that represent the wide variety of dogs the AAC serves each year. We returned to these cases throughout our model building process for scoring and qualitative understanding:



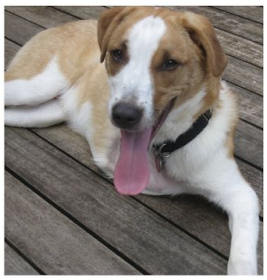
**Ellie**
3 months old
Beagle
Female, not spayed



**George**
14 years old
Labrador
Male, neutered
Hip problems



**Goose**
1.5 years old
Pitbull Terrier
Male, not neutered



**Marshmallow**
6 years old
Mixed breed
Female, spayed
Tri-pawed

# Framing the predictive analytics problem (DIDA)

**Data:** Historical intake & outcome data from Austin Animal Shelter from 2013-Present (filtered to only dogs that have had an outcome).

**Information:** Predicted length of time spent at AAC from intake to outcome.

**Decision:** IF predicted length of time is greater than [threshold] THEN Austin Animal Center should prioritize dog for increased resource allocation (marketing, medical/behavioral care).

**Advantage:** Better distribution of resources, decreased costs, better understanding of future capacity, informed planning, and helping dogs find forever homes.

# About the data

We chose to merge the two AAC datasets from Data.gov for our analysis. The first dataset includes animal intakes and the second includes outcomes, and both are from October 2013 to present. The original files are attached with this report.

We picked these datasets because we are interested in how predictive analytics can help non-profit organizations like animal shelters with better helping animals find homes. During our preliminary preprocessing we merged the two files on ID and, given the sheer amount of data in the two original files, filtered the data to only include dogs with records in both the intakes and outcomes datasets. We will go over each step of these preliminary modifications in our Excel and SAS EM pre-processing sections. The resulting merged and modified CSV file we used for our analysis (AAC_CLEANED_FINAL.csv) is attached with this report.

Our finalized dataset includes more than 8,000 records for dog adoption data. While most Animal_IDs are unique, there are a few exceptions (such as one dog showing up three times in the data). Each record represents a unique combination of dog, intake and outcome. The merged dataset has 19 variables, and 10 were kept after dropping the ones that were not relevant or violate *ex-ante* rules (would occur *after* the dog is adopted). The variable types include binary, interval, date/time, categorical and calculations based on the original variables.

# Data pre-processing in Excel

Before we could import our data into SAS EM, we needed to do some preliminary filtering, merging, and cleaning of the raw data. We did this in Excel.

| AAC Intakes Raw Data (81670 intake entries) | |
|---|---|
| Animal ID | String |
| Name | String |
| DateTime | Date |
| MonthYear | Date |
| Found Location | String |
| Intake Type | Categorical |
| Intake Condition | Categorical |
| Animal Type | Categorical |
| Sex upon Intake | Categorical |
| Age upon Intake | String |
| Breed | String |
| Color | String |

→

| AAC Outcomes Raw Data (81820 outcome entries) | |
|---|---|
| Animal ID | String |
| Name | String |
| DateTime | Date |
| MonthYear | Date |
| Date of Birth | String |
| Outcome Type | Categorical |
| Outcome Subtype | Categorical |
| Animal Type | Categorical |
| Sex upon Outcome | Categorical |
| Age upon Outcome | String |
| Breed | String |
| Color | String |

→

| AAC Merged Data (8305 entries after filter/clean) | |
|---|---|
| Animal_ID | String |
| Time_AAC_days | Integer |
| Animal_type | Categorical |
| Purebreed | Binary |
| Blacklisted_breed | Binary |
| Color | Categorical |
| Intake_date | Date |
| Intake_weekday | Categorical |
| Intake_month | Categorical |
| Intake_time_of_day | Binary |
| Found_in_Austin | Binary |
| Intake_type | Categorical |
| Intake_condition | Categorical |
| Sex | Binary |
| Reproductive Condition | Binary |
| Intake_age_years | Integer |
| Date_birth | Date |
| Outcome _type | Categorical |
| Adopted | Binary |

| **Legend** |
|---|
| Merge key |
| Calculated/revised field |
| Untouched field |

**Summary of Excel pre-processing:**

- First, we filtered both datasets to only include dogs, as the datasets were so large they would have required sampling and we wanted to focus our analysis.
- Next, we merged the two datasets with a left join based on the Animal IDs in the Intake dataset. This further reduced the size of the dataset, but also ensured that every record would have both intake and outcome information with minimal missing data.
- Next, we dropped variables that were unchanged from intake to outcome (such as redundant name, breed, birth date, and color variables).
- Next, we standardized several variables, including Intake_age and Outcome_age since the original values were strings in the form of weeks, months, or years. These variables now have numerical floating point values based on a common unit of years. We calculated these from the original Birthdate variable and its corresponding Intake_date and Outcome_date variables with Datedif().
- Next, we standardized and simplified several categorical variables. We kept the variable Color but mitigated inconsistencies of data entry by reducing the number of categories and merging some original ones (example: original colors included ‘grey’ and ‘gray’).
- We also used the variable Breed to create calculated binary variables Purebreed and Blacklisted_breed with ‘Yes’ and ‘No’ values. Purebreed filtered any use of a slash or variation on the word ‘mix’ to categorize as ‘Yes’ with all remaining values as ‘No.’
    - The variable Blacklisted_breed is based off of the most common guidelines from Homeowners Alliances (HOAs) and rental agreements for breeds not allowed in residencies. This is meant to be a proxy for (often incorrect) public perceptions of ‘dangerous’ or ‘non-residential’ dog breeds (example: Pitbulls).
- Next, we created calculated fields including a binary variable called Found_in_Austin to indicate whether a dog was brought in within the city or from outside.
- We next created two binary variables called Reproductive_Condition to indicate whether a dog is spayed/neutered and Sex to indicate the sex of the dog from the original field Sex upon Intake. The Sex upon Outcome field was not used because it would not be known at the time of intake.
- Finally, we created calculated fields out of the intake and outcome date variables to isolate the day of week (Monday-Sunday), month, and time of day (morning or afternoon) that could be potentially used for predictions. Since we are focusing on future dog intakes, we realized we cannot use specific past years/dates for variables on their own (example: if dogs were to have a better chance in 2014, it would not be of predictive use for a dog brought into the shelter today).

Below, we have included sample data from the first row of our merged dataset:

| Variable | Entry 1 |
|---|---|
| Animal_ID | A748291 |
| Time_AAC_days | 126 |
| Animal_type | Dog |
| Purebreed | No |
| Blacklisted_breed | Yes |
| Color | Black |
| Intake_date | 5/1/17 |
| Intake_weekday | Monday |
| Intake_month | May |
| Intake_time_of_day | Afternoon |
| Found_in_Austin | Yes |
| Intake_type | Stray |
| Intake_condition | Normal |
| Sex | Female |
| Reproductive Condition | Intact |
| Intake_age_years | 0.92 |
| Date_birth | 6/1/16 |
| Outcome _type | Transfer |
| Adopted | No |

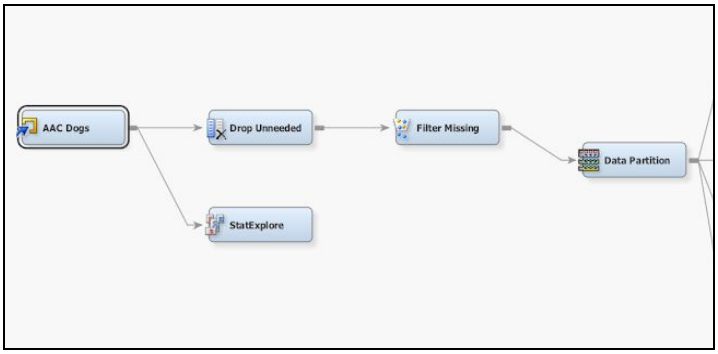
## Data pre-processing in SAS Enterprise Miner

In the data node we imported the cleaned dataset from Excel. Next, we dropped variables that were not relevant to our model such as “Animal_ID” and “Adopted” in the drop node. We used the filter node to remove observations that were missing the dependent variable, “Time_AAC_days.” Lastly, in order to compare the results of our models, we partitioned the data (60% training, 40% testing) to guarantee that each model would use the same training and validation set to assess performance.

AAC Merged Data

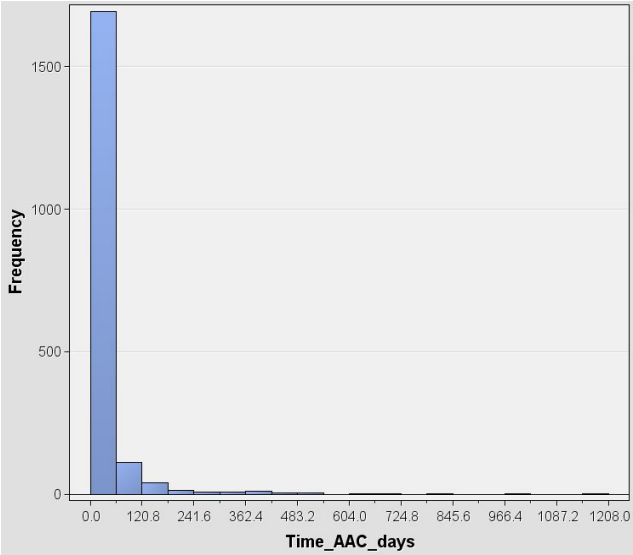
(8305 entries after filter/clean)

| Variable Name | Type | Role |
|---|---|---|
| Animal_ID | String | Dropped |
| Time_AAC_days | Integer | Target |
| Animal_type | Categorical | Dropped |
| Purebreed | Binary | Input |
| Blacklisted_breed | Binary | Input |
| Color | Categorical | Input |
| Intake_date | Date | Dropped |
| Intake_weekday | Categorical | Dropped |
| Intake_month | Categorical | Dropped |
| Intake_time_of_day | Binary | Dropped |
| Found_in_Austin | Binary | Input |
| Intake_type | Categorical | Input |
| Intake_condition | Categorical | Input |
| Sex | Binary | Input |
| Reproductive Condition | Binary | Input |
| Intake_age_years | Integer | Input |
| Date_birth | Date | Dropped |
| Outcome _type | Categorical | Dropped |
| Adopted | Binary | Dropped |

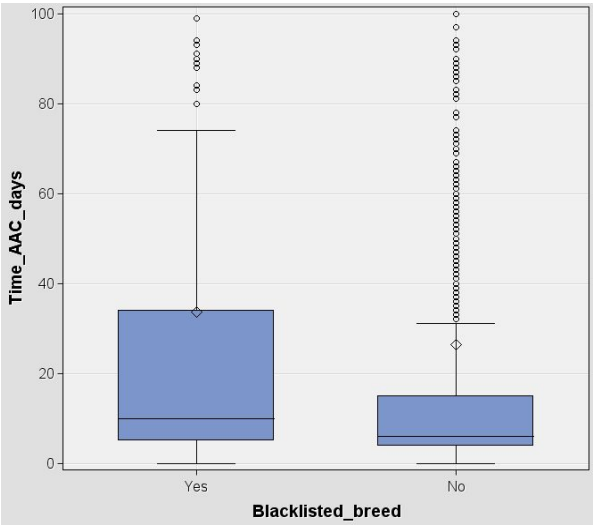


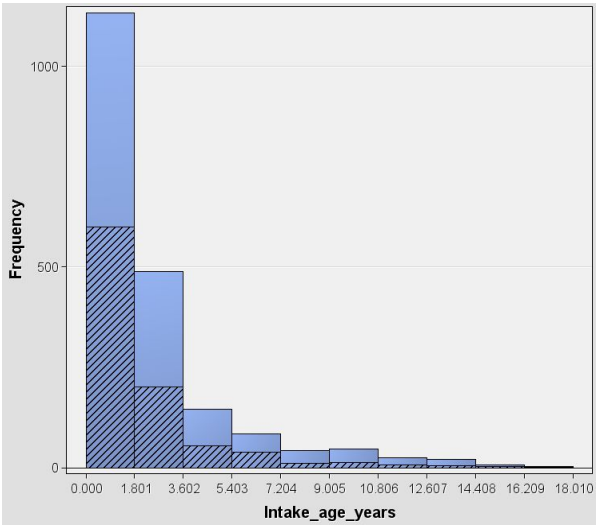
# Preliminary insights & exploration

This initial exploration gave us a sense of how we wanted to analyze the data. Our first major insight was that a significant portion of dogs stayed at the AAC for less than two months. We decided to further explore this by assessing our dependent variable, “Time_AAC_days,” against different benchmarks such as two weeks, one month, and three months.




In our initial exploration we wanted to get a sense for the distribution of the data and identify possible early indicators for our models. Two variables that showed potential for impact on the dependent variable were “Blacklisted_breed” and “Intake_age_years” because of their disparity in distributions. For the blacklisted breeds variable, there were nearly twice as many dogs that were blacklisted than were not. For the intake age variable, the distribution is skewed right with dogs less than 1.8 years old representing a large proportion of the data.

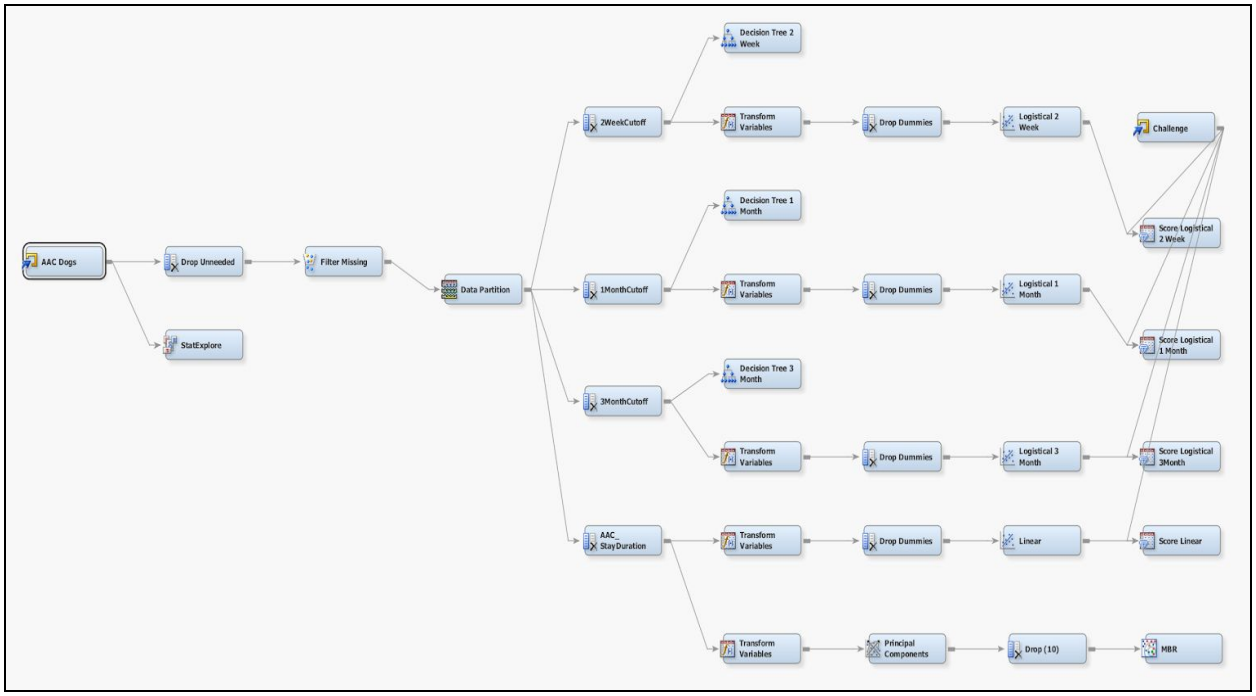


# Model implementation

## Overview of models tried

To address our research question, we used our historic intake/outcome data from the AAC to train a number of different models including Linear Regression, k-Nearest Neighbor, Logistic Regression (multiple cut-offs), and Classification Tree (multiple cut-offs). These models were all used to predict the expected length of stay for dogs at the shelter (either the number of days or if they would leave before a predetermined cut-off) and their performances were compared against each other.




## Linear regression model

**Predictor:** Estimated length of stay at shelter (number of days).

**Implementation:** We standardized values for numerical variables and converted categorical variables to a series of dummy variables and dropped redundant ones.

**Model:** With the training partition, we trained a linear regression model through backward elimination. The final model had 9 coefficients and an intercept. The most influential positive coefficients that increased the number of days stayed at the AAC were: Intake Condition - Sick (+0.1111 days), Intake Weekday - Monday (+0.0537 days), and Intake Age (+0.0374 days). The most influential coefficients that decreased the number of days were: Intake Type - Public Assist (-0.2004 days), Intake Month - June (-0.0907 days), and Found in Austin (-0.0611 days). The most counterintuitive coefficient was the reduction on the number of days stayed that being blacklisted had (-0.0373 days). We had thought being a blacklisted breed would extend the stay of a dog, but our model suggested otherwise.

**Case Study Outcome:** With our four case study dogs, we derived an expected stay of 35 days for puppy Ellie, 64 for grandpa George, 45 for young pitbull Goose, and 53 for tri-pawed Marshmallow.

**Performance:** RMSE = 0.97

## Nearest neighbor model

**Predictor:** Estimated length of stay at shelter (number of days).

**Implementation:** We performed principal component analysis to reduce the number of inputs. We tested different k-values, but the RMSE remained roughly the same for k's larger than eight, so in pursuit of a parsimonious model, k was set equal to eight.

**Model:** Using the training data partition, we created a Nearest Neighbor model with k = 8 nearest neighbors. This model outputs standardized values of the Length of Stay at the AAC that must be unstandardized to be meaningful.

**Case Study Outcome:** Using our four case study dogs, we derived an expected stay of 60 days for puppy Ellie, 20 days for grandpa George, 49 days for young pitbull Goose, and 18 days for tri-pawed Marshmallow. These examples are very different than those generated by the Linear model. This makes sense as each of our cases has one or two unusual characteristics (Ellie is a puppy, George is relatively old, Goose is a blacklisted breed, and Marshmallow has a serious health complication) that might make the cases meaningfully different than the local structure that surrounds them.

**Performance:** RMSE = 1.02

## Classification tree model

**Predictor:** Yes/No if a dog will leave the shelter before three cut-offs: 2 weeks, 1 month and 3 months.

**Implementation:** We calculated a binary field for each of the specified cut-offs in the historical data. The maximum number of depth levels varied depending on the cut-off. For the 2 weeks cut-off, the depth was three. For the one month cut-off, it was four. Lastly, for the three months cut-off, the maximum depth was six. The max depths were set at these levels to reduce the complexity of the trees produced without increasing the misclassification rates an undue amount.

**Models:** The English rules of the two early indicators of the 2 week cut-off model are: first, Intake age < 1.065 years & Intake Age >= 0.085 year and, second, Intake Age >= 1.065 years & Blacklisted Breed = No or Missing. The English Rules for the two early indicators of the 1 month cut-off model are: first, Intake Age < 1.065 years & Intake Age >= 0.095 years and, second, Intake Age >= 1.065 years & Blacklisted Breed = Yes. The English Rules for the two early indicators of the 3 month cut-off model are: first, Intake Age < 0.915 years and, second, Intake Age >= 0.915 years & Intake Month = NOT May, February, or October. From this we can tell that Intake Age is very important for preliminary groupings, with a significant cut-off point of 1.065 years old appearing as the first node in two of three models.

**Case Study Outcome:** Using our four case study dogs and the 2 week cut-off model, all of the dogs are expected to leave within two weeks with the following expected probabilities: 86.7% for Ellie, 75.5% for George, 58.9% for Goose, and 75.5% for Marshmallow. With the 1 month cut-off model, all of the dogs are again expected to leave within the cut-off time period, now set to a month, with the following expected probabilities: 87.8% for Ellie, 78.6% for George, 67.1% for Goose, and 78.6% for Marshmallow. Using the final model, with a 3 month cut-off, all of the dogs are again expected to leave within the cut-off with the following expected probabilities: 95.2% for Ellie, 92.2% for George, 92.2% for Goose, and 92.2% for Marshmallow. As the cut-off time increases, all of the probabilities also increase. Interestingly, Ellie, the puppy, has the highest chance of a short stay across all models while Goose, who is a Blacklisted Breed, has the lowest or is tied for the lowest chance at a short stay in every model. This finding seems to contradict the earlier finding that Blacklisted Breeds spend less time at the AAC.

**Performance:** For the 2-Week cut-off, the misclassification rate was 28%. This dropped to 19% for the 1 month cut-off and to 9% for the last cut-off of 3 months. These results are only slightly better than a Naive decision rule: 35% for the 2 week cut-off misclassification, 22% for the 1 month misclassification, and 9% for the 3 month misclassification.

## Logistic regression model

**Predictor:** Yes/No if a dog will leave the shelter before three cut-offs: 2 weeks, 1 month and 3 months.

**Implementation:** We calculated a binary field for each of the specified cut-offs in the historical data. Numerical values were standardized and categorical ones were coded into dummy variables, dropping one redundant variable.

**Models:** Three separate logistic regression models were created from the Training partition of the dataset using Backward Elimination. The 2 week model has 14 coefficients and the most impactful are: Intake Type - Public Assist (+1.73 days), Intake Type - Stray (+ 1.30 days), Intake Condition - Sick (-1.14), Intake Type - Owner Surrender (+1.13 days), and Intake Condition - Aged (-0.98). The 1 month model has 33 coefficients and the most impactful are: Intake Condition - Ages (-1.29), Intake Condition - Sick (-1.17 days), and Reproductive Condition - Intact (+0.81 days). The 3 month model has 20 coefficients with the most impactful being: Intake Condition - Sick (-0.55 days), Intake Type - Public Assists (+0.47 days), and Color - Blue (-0.38 days).

**Case Study Outcome:** using the 2 week logistic regression model, it was predicted that all of the dogs except for George will stay for less than 2 weeks at the AAC. The chances for each of the dogs to meet this cut-off is: 75.3% for Ellie, 45.2% for George, 54.9% for Goose, and 61.2% for Marshmallow. The 1 month model predicts that all of the dogs will stay less than a month at the AAC. The predicted chances for each dog to leave before a month is up is as follows: 82.8% for Ellie, 60.2% for George, 72.7% for Goose, and 74.3% for Marshmallow. The discrepancy between this and the first model may be attributed to the the predicted chances of George making the two cut-offs being very close to the probabilistic cut-off that is employed by the models and a small variance can result in a drastic

reclassification. With the 3 month model, all of the dogs are predicted to leave before 3 months are over with a high level of certainty: Ellie has a 93.2% chance, George has a 83.2%, Goose has a 88.8%, and Marshmallow has a 89.0% respectively to leave before the 3 month cut-off.

**Performance:** For the 2 week cut-off, the misclassification rate was 30%. This dropped to 22% for the 1 month cut-off and to 9% for the last cut-off of 3 months. These results are only slightly better than a Naive decision rule: 35% for the 2 week cut-off misclassification, 22% for the 1 month misclassification, and 9% for the 3 month misclassification.

## Model selection

To select a final model, we started by looking at our DIDA framework. Our main focus was to predict the length of an expected stay at the shelter when given a dog's information. While the classifier techniques helped us understand the likelihood of a dog staying at a shelter beyond a set period of time, the numeric techniques gave us a better understanding of the number of days we should expect a dog to be in the shelter. Our main dependent variable was a numerical value, so the numeric techniques fit better for our project.

| | Numeric Output? | Accuracy | Speed | Robustness | Presentation |
|---|---|---|---|---|---|
| Linear Regression | ✔ | RMSE = 0.97 | Quick | ✘ | Okay |
| k-Nearest Neighbors | ✔ | RMSE = 1.02 | Slow | ✔ | Bad |
| Classification Tree | ✘ | 2 week: 28%<br>1 month: 19%<br>3 month: 9% | Medium (Slow Update) | ✔ | Good |
| Logistic Regression | ✘ | 2 week: 30%<br>1 month: 22%<br>3 month: 9% | Quick | ✘ | Okay |

When comparing the different models, we first considered the performance of each model. To do this, we compared the error and misclassification rates of each model and found the linear regression had the highest accuracy. Next, we looked at the run-time speeds of the different models. Both regressions, linear and logistic, had the highest speeds. This was important for this project because the original dataset is relatively large and dynamic, so speed was an important factor to consider. For robustness, the classification tree was considered the best model due to its ability to handle non-monotonicity. Lastly, for presentation, the classification tree was superior to the rest of the models due to its visual diagram output. However, in comparison, the linear regression was still simple enough that explaining the inputs, contributions of individual effects (coefficients), and prediction values is not a major problem.

Taking these five comparison criteria into consideration, we decided to select the linear regression model. The linear regression provides granularity that allows the model user to see the contributions of the input to the total. It is also easy to compute and simpler than nearest neighbor and logistic regression, while also allowing some insight into the model through an interpretation of the effects of the coefficients.

## Case study testing

After selecting linear regression, we returned to the four case study dogs to test our model. Below is the data for our four dogs that was put into our models for testing:

| | **Ellie** | **George** | **Goose** | **Marshmallow** |
|---|---|---|---|---|
| Animal_ID | 1 | 2 | 3 | 4 |
| Animal_type | Dog | Dog | Dog | Dog |
| Purebreed | Yes | Yes | No | No |
| Blacklisted_breed | No | No | Yes | No |
| Color | Multi | Black | Multi | Multi |
| Intake_date | 4/17/2018 | 4/17/2018 | 4/17/2018 | 4/17/2018 |
| Intake_weekday | Tuesday | Tuesday | Tuesday | Tuesday |
| Intake_month | April | April | April | April |
| Intake_time_of_day | Morning | Morning | Morning | Morning |
| Found_in_Austin | Yes | Yes | Yes | Yes |
| Intake_type | Stray | Stray | Stray | Stray |
| Intake_condition | Normal | Injured | Normal | Injured |
| Sex | Female | Male | Male | Female |
| Reproductive Condition | Intact | Fixed | Intact | Fixed |
| Intake_age_years | 0.25 | 14 | 1.5 | 6 |
| Date_birth | 1/23/2018 | 4/23/2004 | 10/23/2017 | 4/23/2012 |

We saw in our selected Linear Regression model that our case study dogs showed prediction outcomes similar to those expected by industry expertise. Ellie, the young purebred Beagle, had the shortest expected stay, and we could predict that if she's spayed while in the shelter, her stay will be even shorter. George and Marshmallow, on the other hand, had significantly longer stays due to their more advanced age and health challenges. Most importantly, by creating testing data based on realistic dog profiles, we were able to better envision an implementation plan that would be easily adaptable to the AAC's current services.

**Ellie**
3 months old
Beagle
Female, not spayed

Predicted: 35 days

**George**
14 years old
Labrador
Male, neutered
Hip problems

Predicted: 64 days

**Goose**
1.5 years old
Pitbull Terrier
Male, not neutered

Predicted: 45 days

**Marshmallow**
6 years old
Mixed breed
Female, spayed
Tri-pawed

Predicted: 53 days

# Conclusion

We framed our project from the beginning around the following research question:

*Can we predict the length of time a dog will spend at Austin Animal Center between intake and adoption based on a historical dataset of intakes and outcomes?*

To address this question, we used historic intake and outcome data from the Austin Animal Center to train a number of different models including Linear Regression, k-Nearest Neighbor, Logistic Regression (multiple cut-offs), and Classification Tree (multiple cut-offs). These models were all used to predict a variation of the expected length of stay for dogs at the shelter (either the number of days or if they would leave before a predetermined cut-off) and their performances were compared against each other. We selected the Linear Regression model due to its ease of interpretation of effects, granularity, speed, and presentability. This would help the shelter to determined which dogs need to be prioritized and allocated more resources. This would allow the shelter to run more efficiently, better distribute resources, decrease costs, better understand future capacity, and ultimately help dogs find their forever homes.

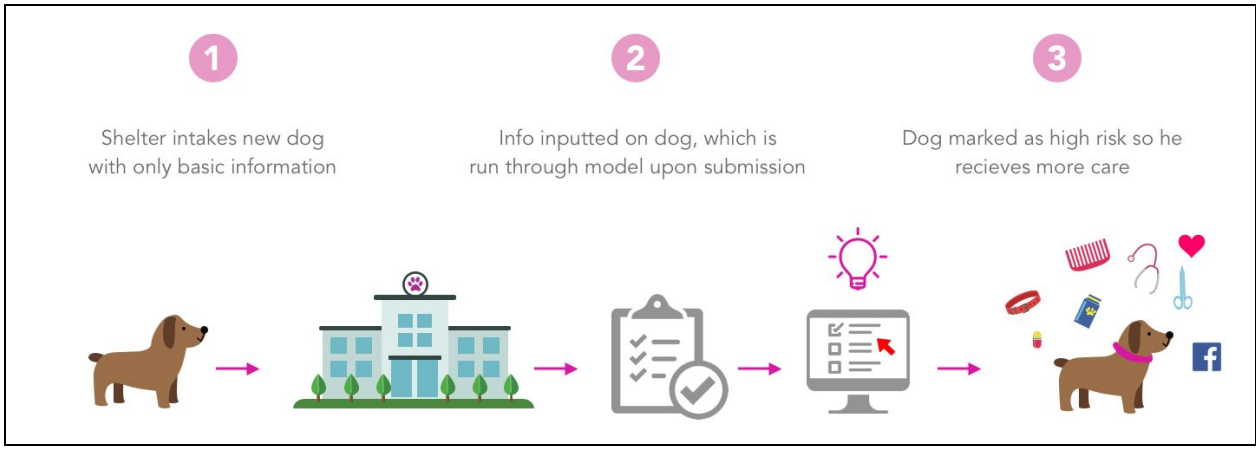
## Project challenges

The first challenge we encountered was the quality of the AAC data. In the world of non-profit organizations, it is common to see messy data. This includes human/input errors when entering the data, as well as lack of proper documentation regarding variables and a data dictionary. Besides, in order to get more insights from this data, we had to pre-process the data in Excel and SAS to create

calculated fields that helped us go deeper into the data and answer the questions we were interested in. Some variables, such as location or breed, were string inputs so we had to modify them in order to use them for analysis. An example of this is when we looked at the homeowner's association (HOA) list of "banned dogs" to determine if these dogs were less likely to be adopted or not. So to do this, we had to process the data to create a new binary variable that represents membership in this group. Lastly, we had to designate ex-ante fields. The goal of this model was to predict at the point where a dog is received in the shelter, so we had to remove the variables that are not present at that point in time. This allowed us to test "new data" that represents what future data will look like.

## Real-world implementation

There are unique challenges when implementing predictive models in real-world scenarios, especially those of non-profits. Animal shelters are largely volunteer-run, and those inputting, maintaining, and using the data varies considerably; the person who inputs the intake fields for dogs on one day may a new person the next day or at a different time. As such, we designed an implementation plan based on the current procedures that the Austin Animal Shelter uses. When a dog enters the shelter (stage 1), there is often minimal information about the dog available. Experienced shelter workers use their best intuition to inform the breed and age estimations of the dog, input this data into their current data management system (DMS), and then receive a length-of-stay prediction.




As the AAC has moved to open data and largely digital record management as part of Austin's digital initiatives, the organization would be ideal for model implementation. When the information is inputted to the DMS (step 2), our selected linear regression model could be used to predict, in real-time, what the dog's estimated stay at the AAC will be. The shelter can, in collaboration with the database managers, create thresholds for low-risk, medium-risk, and high-risk dogs, such that the predicted length of stay can be categorized for an insightful care plan. In this implementation, a high-risk dog would be flagged upon intake and given a color-coded collar or tag, and thus marked for all other shelter workers for additional marketing and care (step 3). Through this coding, the model would complement the AAC's existing process and be easily understood by all shelter workers.

## Closing thoughts

This project was an opportunity for us to apply predictive modeling to a real-world challenge faced by non-profits such as the Austin Animal Center. This type of model could be easily translated to other animal shelters around the country, and even encourage more precise data collection. The advantage of implementation, as previously envisioned, could include a reduction of costs, better allocation of resources, and higher precision planning based on capacity expectations. Most importantly for the AAC, the low-cost implementation of a linear regression predictive model could help them place more dogs in forever homes, creating a lasting impact in the Austin metro area.