Addressing Animal Shelter Adoption with Supervised Methods

Jeffrey Clancy, Gretchen Forbush, Quinn Johnson, Lexi Lin
August 13, 2024

Prepared for the Executive Board of PAWS Chicago

Roadmap

1. Background, Objective, Plan
2. Data Description & Data Cleaning
3. Modeling Methodology & Results
4. Insights & Recommendations

Executive Summary

Animal shelters across the United States, like PAWS Chicago, are struggling with the side effects of overpopulation. After a pandemic adoption frenzy, many owners are returning to their pre-lockdown lifestyles and finding that they are ill-equipped to care for their animals, forcing them to return them to shelters. In order to support PAWS's No Kill Mission, we have produced two models to guide policy to increase adoption rates at the shelter. These models collectively identify the importance of animal age and time in shelter in determining their likelihood of adoption. Using these insights, we recommend that PAWS invest in marketing & advertising and continue to study the characteristics in their population to identify actions that can be taken to improve every animal's chance of finding their forever home.

Background, Objective, Plan

Pets Are Worth Saving: Advancing PAWS No Kill Mission

Background: Approximately **6.3 million** companion animals enter adoption shelters each year; sadly, nearly **1 million** of these animals are euthanized. PAWS Chicago's current No Kill Mission is focused on controlling overpopulation.

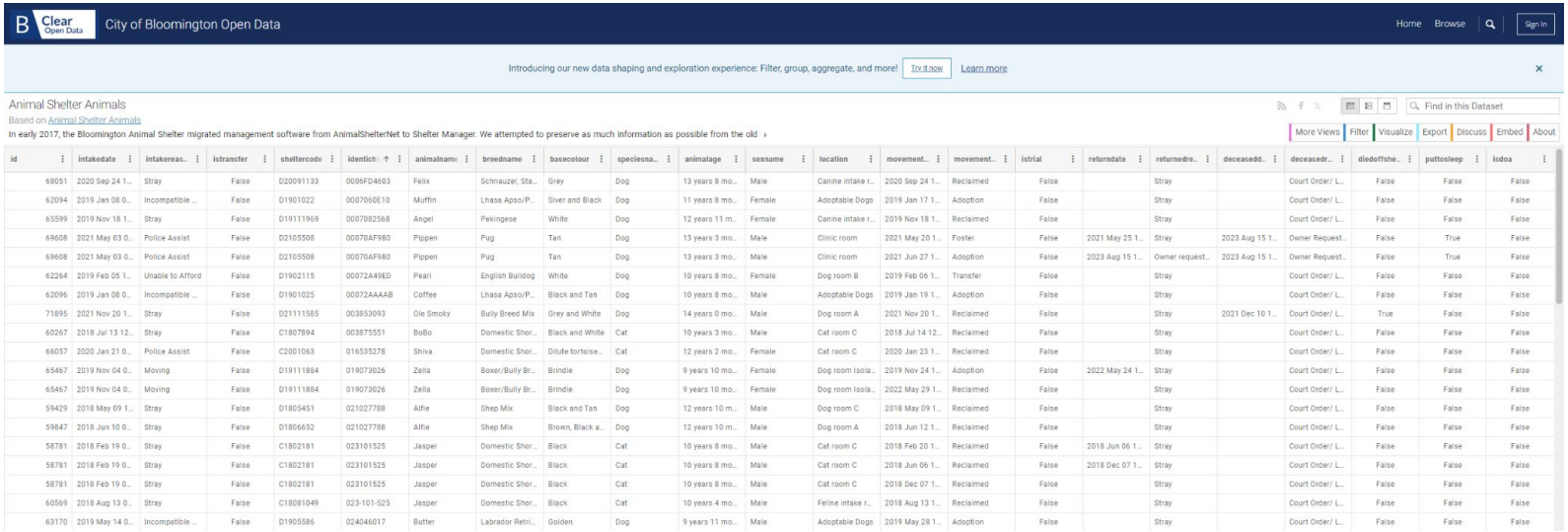
Objective: Apply Data Science to help advance PAWS No Kill Mission.

- Identify characteristics that lead to successful adoptions to forever homes.
- Provide insights for raising adoption rates through animal treatment and promotion.

Using Public Data to Build Our Analysis

Organization: Animal Care & Control | City of Bloomington, Indiana

Data Set: Animal Shelter Animals (2018 - 2022) from the City of Bloomington Open Data Portal



Applying Data Science to Advance PAWS No Kill Mission by Increasing Adoption Rates

Clean & Aggregate the Data

- Shelter database is designed for tracking intake & outtake movements, not for analysis
  - Response variable does not exist in the data and must be defined & created
  - Categorical variables need to be aggregated to prevent size imbalances, causing some data loss during model optimization

Run the Models - Logistic Regression and Regression Tree

- Logistic Regression
  - Provides a clear interpretation of the impact of each variable on the response
  - Highlights only the relevant variables following stepwise optimization
- Regression Tree
  - Provides insight into levels of continuous variables and different factors that make the largest impact
  - Allows for easy visualization of how most important factors work together in the model

Uncover Insights & Provide Recommendations

- Use insights from models to provide recommendations for next steps to PAWS

Data Description & Data Cleaning

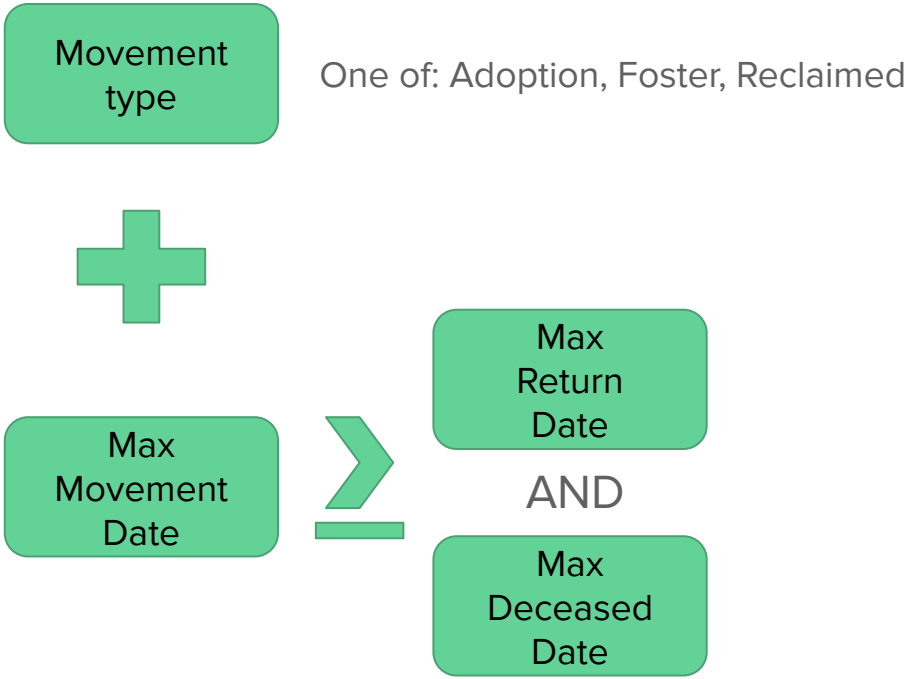
Our Data Requires Us to Define and Create Our Response Variable and to Perform Feature Engineering to Reduce Dimensionality

Movement type	Movement date	Animal age	Intake date	Intake reason	Gender
Return date	Deceased date	Breed name	Base colour	Species name	
location	Is transfer	Shelter Code	Is trial	ID	Returned reason
Identichip Number	Deceased Reason	Diedoff Shelter	Put to Sleep	Is Doa	Animal name

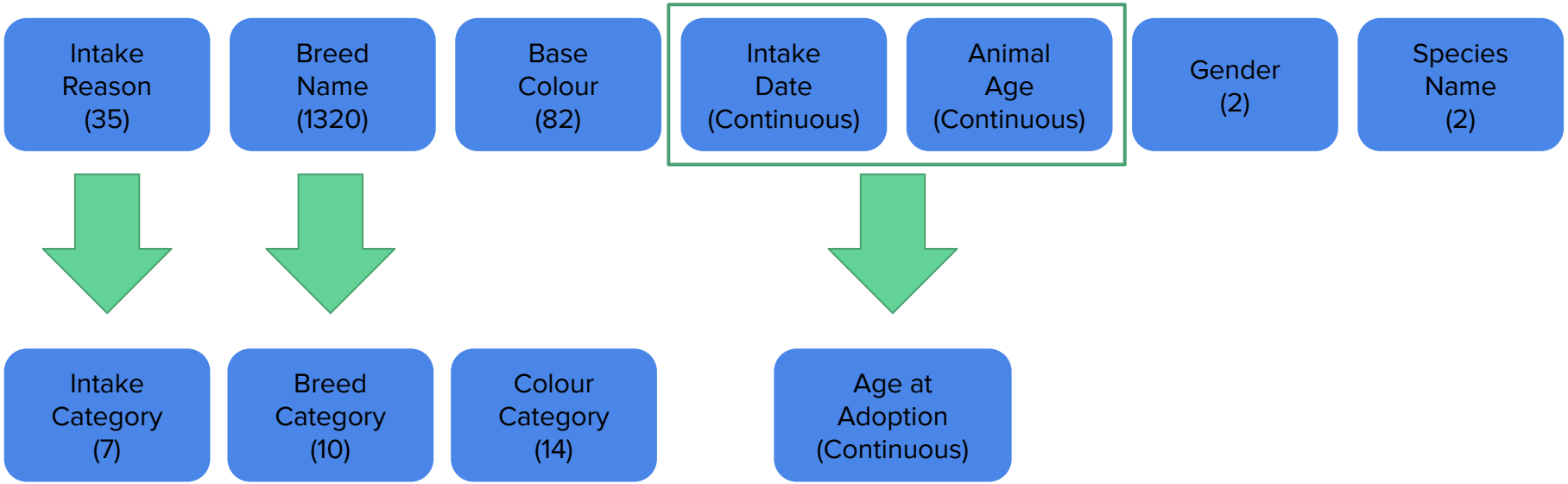
Legend: Used to Create Response Variables | Used to Create Predictor Variables | Not Used

Defining Our Response Variable: Is the Animal Adopted

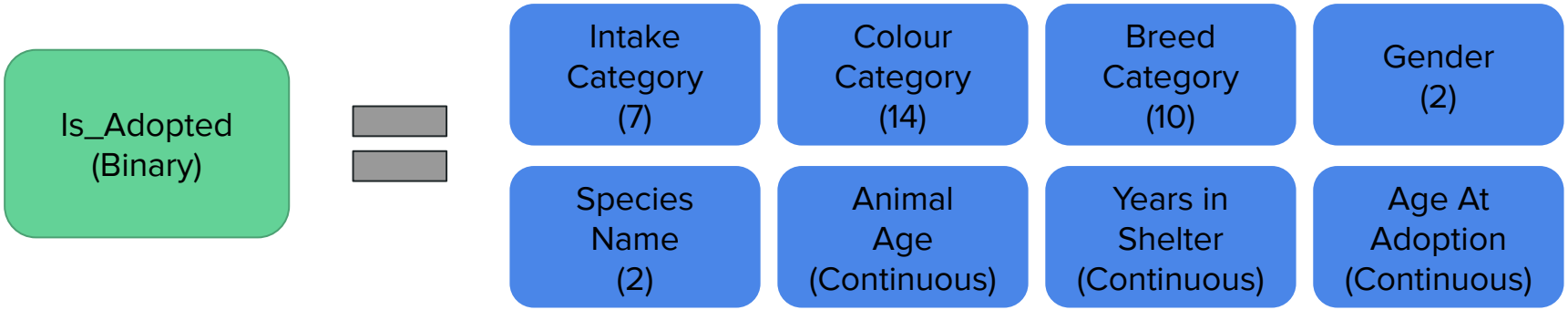
- Challenges:
  - Animals could be coded for movement type “adoption” even if they were subsequently returned
  - Columns for animals passing away were coded for every animal, even if the animal had no deceased date
  - Dataset contained many redundant rows for the same animal, requiring the data to be aggregated at the animal level to be analyzed
- Solution
  - Use the maximum date of both positive (adoption, foster, reclaimed) and negative (returned, deceased) to effectively categorize positive vs negative outcomes



Feature Engineering: Reducing Categorical Variables and Creating New Predictor Variables



Final Data Set Contains a Binary Response Variable and Both Categorical and Continuous Numeric Predictor Variables



Legend: Response Variables | Predictor (Continuous or Number of Categories)

Modeling Methodology & Results

Logistic Regression Significantly Outperforms *A Priori* Expectations

Regression Output

- Following stepwise model selection, our model retains 21 of the 33 variables, of which 19 are statistically significant at the 5% level
- It is notable that neither of the species variables, cats or dogs, was found to be statistically significant, highlighting that the two species have similar rates of adoption, all else being equal

Confusion Matrices

- Model shows high predictive accuracy , much higher than the No Information Rate

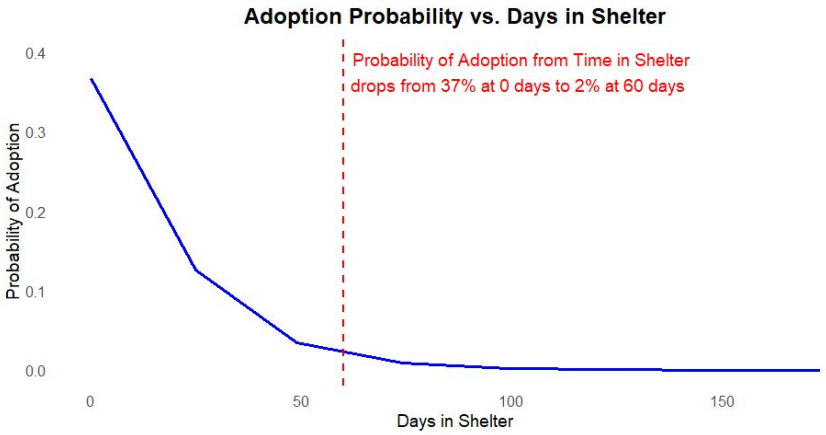
Final Model Summary

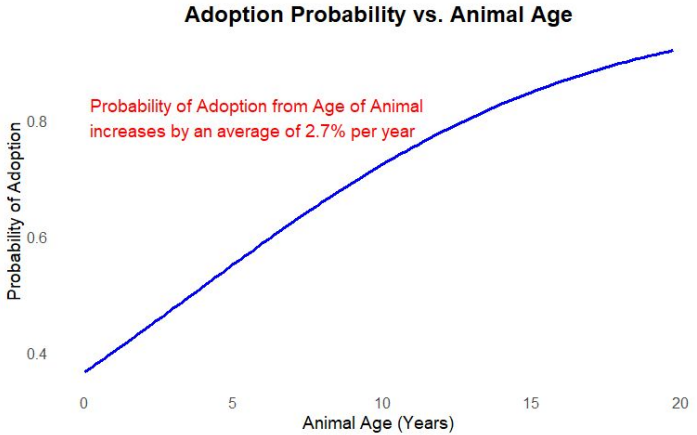
```
Call:
glm(formula = is_adopted ~ animalage_numeric + years_in_shelter +
    Abandonment_Stray + Behavioral_Issues + Health_Issues + Housing_Financial_Issues +
    Incompatibility_Issues + Other_Intake_Reason + Black_Variants +
    Mixed_Variants + Other_Colour + Seal_Variants + Domestic_Long_Hair +
    Domestic_Medium_Hair + Domestic_Short_Hair + Herding_Dogs +
    Mixed_Breeds + Other_Breed + Retrievers + Terriers + Female,
    family = binomial, data = data.train)

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)              -0.54513    0.23125  -2.357 0.018410 *
animalage_numeric         0.15080    0.00819  18.412  < 2e-16 ***
years_in_shelter        -20.59788    0.44956 -45.818  < 2e-16 ***
Abandonment_Stray         1.35791    0.16328   8.317  < 2e-16 ***
Behavioral_Issues         1.44104    0.22208   6.489 8.65e-11 ***
Health_Issues             1.65443    0.21101   7.840 4.49e-15 ***
Housing_Financial_Issues  1.15100    0.16686   6.898 5.28e-12 ***
Incompatibility_Issues    1.74037    0.17885   9.731  < 2e-16 ***
Other_Intake_Reason       1.02867    0.43449   2.368 0.017908 *
Black_Variants            0.11914    0.05232   2.277 0.022782 *
Mixed_Variants            0.97188    0.34742   2.797 0.005151 **
Other_Colour              0.19930    0.08533   2.336 0.019509 *
Seal_Variants             0.92514    0.46684   1.982 0.047512 *
Domestic_Long_Hair        0.34357    0.20058   1.713 0.086740 .
Domestic_Medium_Hair      0.66643    0.19194   3.472 0.000517 ***
Domestic_Short_Hair       0.67131    0.16215   4.140 3.47e-05 ***
Herding_Dogs              0.36873    0.19587   1.883 0.059761 .
Mixed_Breeds              0.72937    0.17051   4.278 1.89e-05 ***
Other_Breed               0.49734    0.17834   2.789 0.005291 **
Retrievers                0.52315    0.18608   2.811 0.004931 **
Terriers                  0.60944    0.20383   2.990 0.002790 **
Female                   -0.14858    0.04985  -2.980 0.002880 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 16471  on 12001  degrees of freedom
Residual deviance: 10457  on 11980  degrees of freedom
AIC: 10501

Number of Fisher Scoring iterations: 8
```

Confusion Matrices

```
[1] "Confusion Matrix for Training Data:"
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 3796  708
         1 1497 6001

               Accuracy : 0.8163
                 95% CI : (0.8092, 0.8232)
    No Information Rate : 0.559
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.6214

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.8945
            Specificity : 0.7172
         Pos Pred Value : 0.8003
         Neg Pred Value : 0.8428
             Prevalence : 0.5590
         Detection Rate : 0.5000
   Detection Prevalence : 0.6247
      Balanced Accuracy : 0.8058

       'Positive' Class : 1

[1] "Confusion Matrix for Test Data:"
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1609  303
         1  661 2571

               Accuracy : 0.8126
                 95% CI : (0.8017, 0.8232)
    No Information Rate : 0.5587
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.6136

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.8946
            Specificity : 0.7088
         Pos Pred Value : 0.7955
         Neg Pred Value : 0.8415
             Prevalence : 0.5587
         Detection Rate : 0.4998
   Detection Prevalence : 0.6283
      Balanced Accuracy : 0.8017

       'Positive' Class : 1
```

Time is of the Essence and Age is More Than Just a Number





The model highlights how dramatically the likelihood of adoption drops for animals that are in the shelter for even a relatively short period of time, with the probability contribution from time in shelter dropping close to 0 once an animal has been in the shelter for 2 to 3 months.

The model indicates that older animals are more likely to find permanent homes through adoption, as the likelihood that an animal is adopted steadily rises with the age of the animal.

Animal Characteristics Have Limited Impact on Adoptability, With A Few Notable Exceptions

Intake

- All intake categories significantly increase adoption probability compared to Shelter Transfer.

Color

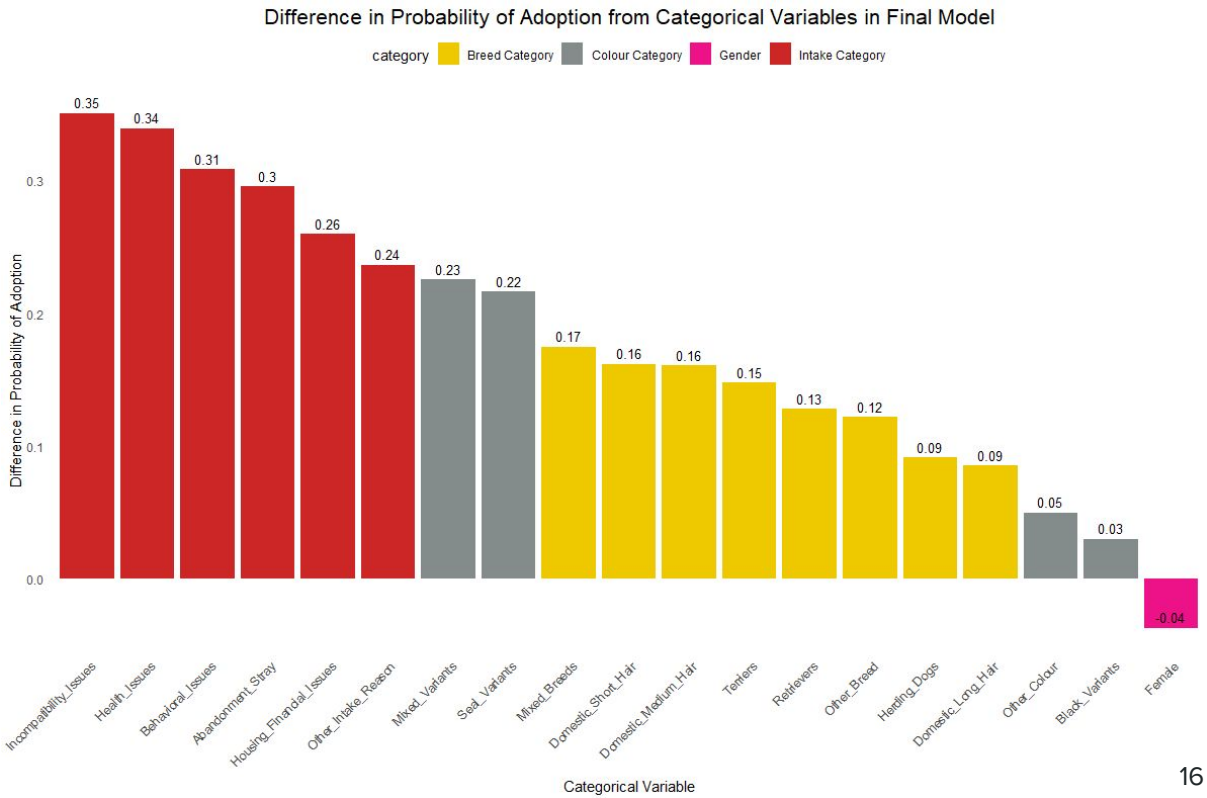
- Only 4 of 14 color categories significantly impact adoption probability, suggesting color generally isn't a major factor, with the exception of Mixed and Seal variants.

Breed

- Toy Breeds and Bulldogs were not identified as significant by the model, indicating a notable gap between these and other breeds.

Gender

- Female animals have a slightly, but significantly, lower adoption rate than male animals.



Pruned Tree Model Produces Even Stronger Predictions

- Confusion matrices for the pruned tree show stronger balanced accuracy across both the train and test data sets than the logistic regression model

- In particular, the pruned tree has a much stronger specificity, indicating that it is a stronger predictor of animals that are not able to find permanent adoption

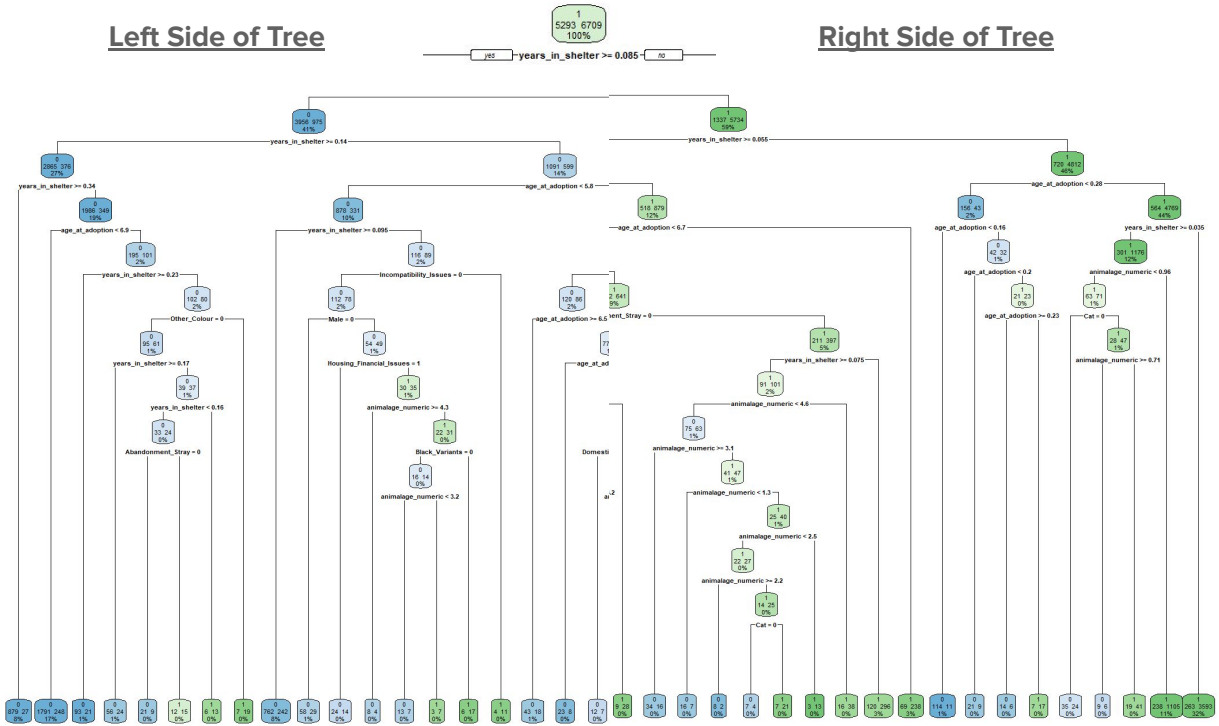
```
[1] "Confusion Matrix for Training Data:"
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 4370  880
         1  923 5829

               Accuracy : 0.8498
                 95% CI : (0.8433, 0.8561)
    No Information Rate : 0.559
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.695

 Mcnemar's Test P-Value : 0.3226

            Sensitivity : 0.8688
            Specificity : 0.8256
         Pos Pred Value : 0.8633
         Neg Pred Value : 0.8324
             Prevalence : 0.5590
         Detection Rate : 0.4857
   Detection Prevalence : 0.5626
      Balanced Accuracy : 0.8472

       'Positive' Class : 1
```

```
[1] "Confusion Matrix for Test Data:"
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1818  431
         1  452 2443

               Accuracy : 0.8283
                 95% CI : (0.8178, 0.8386)
    No Information Rate : 0.5587
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.6515

 Mcnemar's Test P-Value : 0.5009

            Sensitivity : 0.8500
            Specificity : 0.8009
         Pos Pred Value : 0.8439
         Neg Pred Value : 0.8084
             Prevalence : 0.5587
         Detection Rate : 0.4749
   Detection Prevalence : 0.5628
      Balanced Accuracy : 0.8255

       'Positive' Class : 1
```

Pruned Tree Model Highlights the Importance of Time in Shelter and Animal Age

- Both sides of the tree model highlight the importance of time in the shelter on the probability of adoption, with age of adoption being the second most-important variable

- While some categorical variables are present, they are not nearly as important as the two continuous variables mentioned above, as the majority of animals can be put into branches without evaluating the categorical variables



Insights & Recommendations

Data Science Can Provide Useful Improvements to PAWS No Kill Mission

Time is of the Essence

- Both models show that it is imperative to move animals out of the shelter and into a home quickly

Age is More Than Just a Number

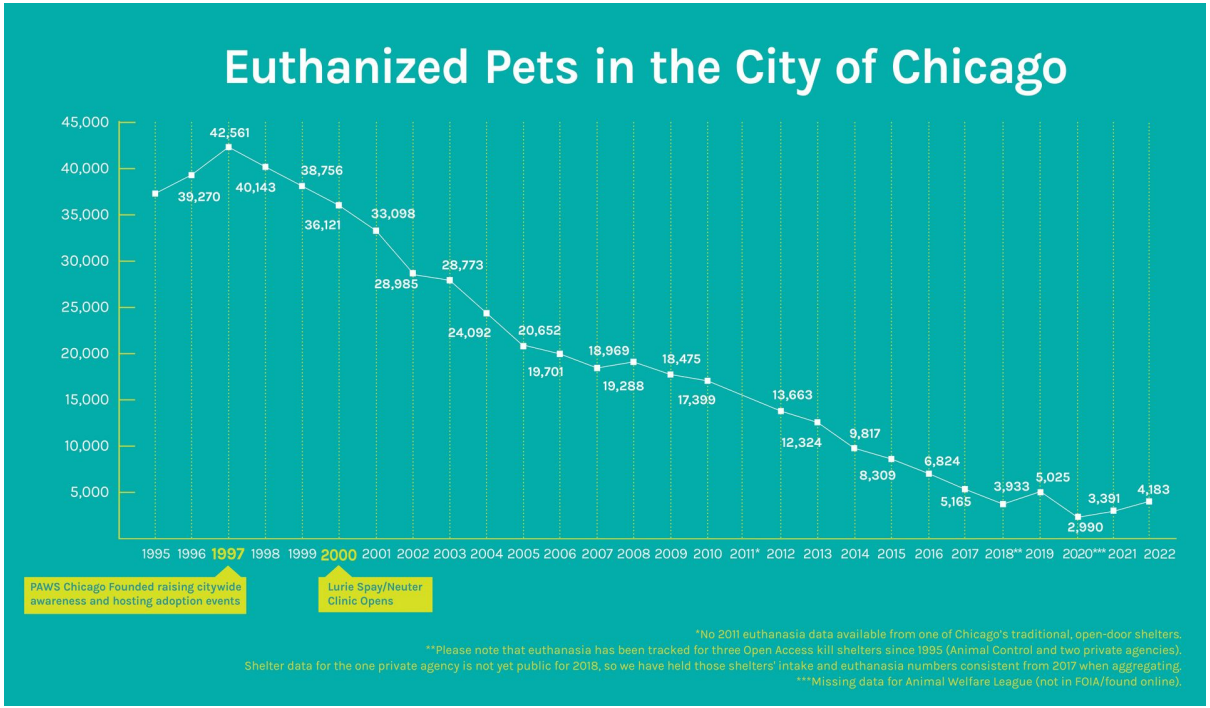
- Both models show that older animals are more likely to be adopted to their forever home

Animal Breed, Color, and Demeanor Are Not Key Determinants for Adoption

- Both models indicate that an animal characteristics do not affect probability of adoption with some minor exceptions

PAWS Chicago's No Kill Mission Has Made Great Strides

- Collaborate to build a comprehensive database allowing improvements to the No Kill Mission
  - PAWS Chicago's No Kill Mission, geared mostly towards Spay/Neuter, has been effective in reducing the number of euthanized animals



The Future of the No Kill Mission Is Even Brighter

- Raise money for marketing and advertising to increase opportunity for animals to be adopted in the first 60 days in the shelter
  - The results of both models highlight the return on investment that donors would realize on their contribution to PAWS if the speed at which animals are adopted can be increased through awareness
- Create a Database that allows time series analysis to evaluate adoptability
  - While the model allowed us to understand the positive impact of quickly moving animals out of the shelter, it did not provide the means to understand **why** animals had such low probability of adoption after 60 days in the shelter
- Study psychological effects of time in shelter
  - Monitor compatibility/adoptability traits while in shelter
  - Analyze whether an animal's disposition changes over time while in shelter