Pawpulation Forecast

Final Presentation
August 6, 2024

Our Team

Erin Smith

Project Manager

Rob Chimel

Data Lead

Andre Gigena

Front End Lead

Jessica Huber

Model Lead

Table of contents

01.

Problem

Problem

Animal shelters have limited resources[1] and many animals are in need of temporary shelter[2], making it critical to optimize care. Knowing the length-of-stay of each animal would help shelters make better decisions on resource allocation[3].

1 https://www.usatoday.com/story/news/nation/2024/02/20/more-dogs-euthanized-than-cats-2023/72633423007/
2 https://www.aspca.org/helping-people-pets/shelter-intake-and-surrender/pet-statistics
3 Subject Matter Expert Interview with Melanie Sobel Director of Denver Animal Protection

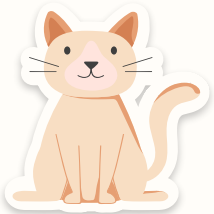
02.

Solution

Pawpulation Forecast is a length-of-stay prediction report generator for animal shelters

Shelters will be able to leverage these reports to make resource allocation decisions that will help them reduce the number of animals in their care at a given time

Floki
Puppy / Small
Stray
Not Neutered



Puja
Adult / Large
Owner Surrender
Treateable

Length of Stay Prediction:
0 - 3 Days

Pawpulation Forecast

Length of Stay Prediction:
30 - 100 Days

Prioritize Neutering Treatment

Prioritize Foster Home Placement

Our approach includes an automated data flow with XGBoost model training and self-serve report generation and export

Shelter Data System API

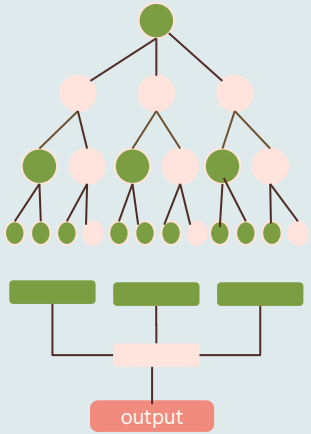
Raw Data

Name
Kennel Number
Type
Color
Sex
Intake_Type
Intake_Juris
Intake_Date
Outcome_Date

Feature Enrichment

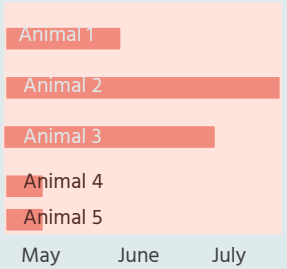
Is_Fixed
Has_Name
Is_Mixed_Breed
Is_MultiColor
Days_in_shelter
Bin_Label

XGBoost Model



Report

"Generate prediction for animals that arrived in May"



CSV Export

03.

Data

Data Source & Engineering

Animal Shelter Income and Outcome Dataset - *Sonoma Department of Health Services, Denver Animal Protection & Austin Animal Center*

- ~230K records & counting
- API for Sonoma and Austin (MVP uses Sonoma API)
- Train on Sonoma, test on Denver and Austin

Feature Engineering

- Handling all datasets, a challenge that paid off
- Added 14 new features by scraping keywords, embedding notes, summarizing dates

Data Pipeline Features

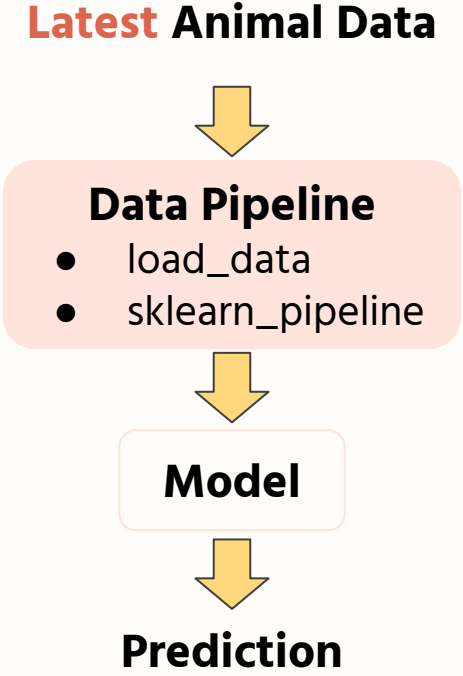
Simple

- *load_data* cleans user inputs and handles all feature engineering
- *sklearn_pipeline* prepares clean data for inference

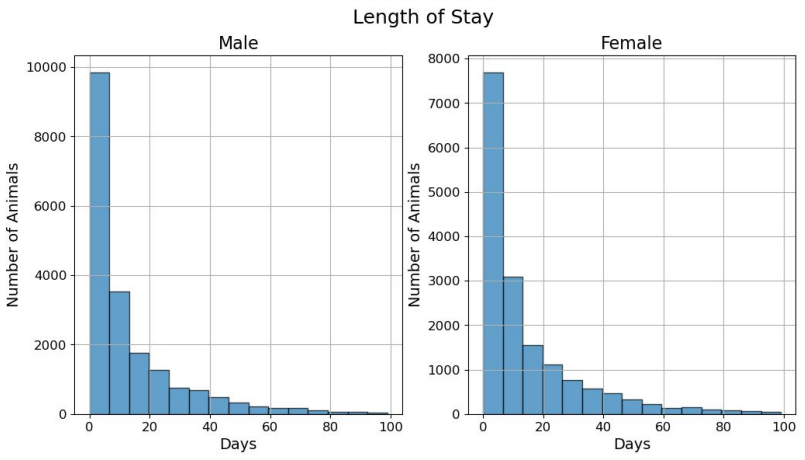
Functional

- Accepts multiple inputs, from API calls, bulk uploads and user inputs
- Lightning fast report generation* and intake results

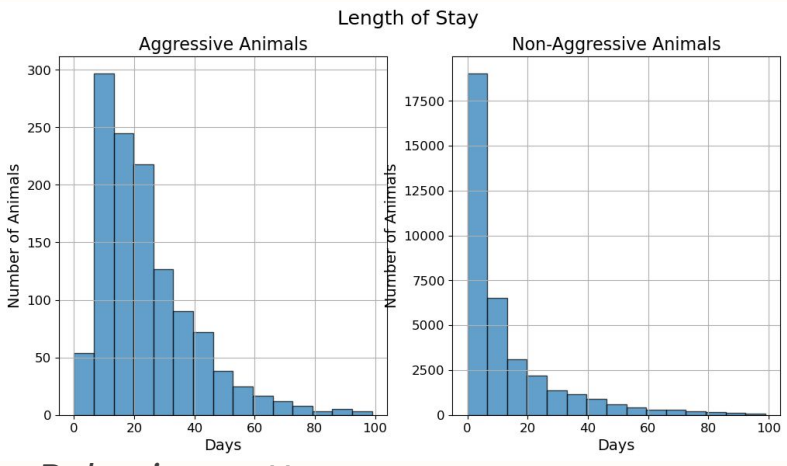
**1 month reports are processed in under a second in MVP*



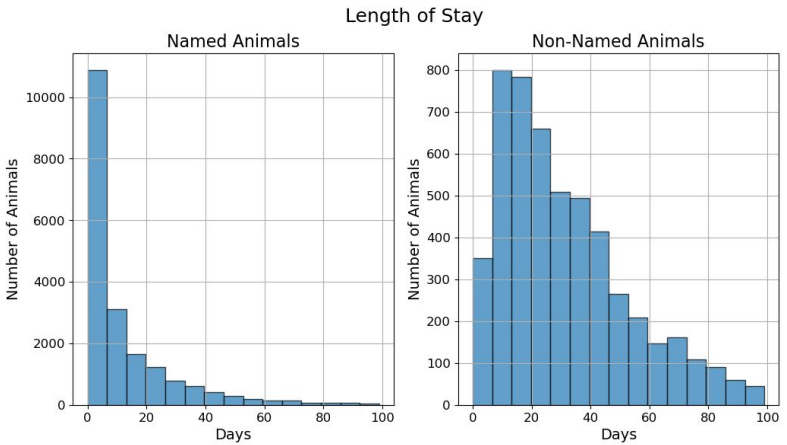
EDA Learnings



Gender is not a good predictor



Behavior matters



Named animals leave faster

04.

Model

Our approach builds on related work[5] which sought to predict animal exit outcome

Predicting the Outcome of Dogs at the Austin Animal Center

Presented by:

Giulia Bronzi, Hannah Kim, Willa Sun, Nikki Tong (Wellesley College)

Similarities

- Both applications intend to help shelters make better decisions upon intake
- Animal shelter intake and outcome data includes similar animal information
- Both model outputs are classification predictions

Differences

- Model objectives: research predicts exit outcome vs our model predicts days at shelter
- Animal shelter data: research used Austin data vs we used Sonoma data
- Model approaches: research used regression tree vs we selected XGBoost

5 https://www.causeweb.org/usproc/eusrc/2020/program/8

Our chosen model had the highest overall performance compared to alternative options

Explored baseline models with bucket labeling

- Logistic Regression: 0.45 accuracy, 0.37 f1-score
- Random Forest: 0.61 accuracy, 0.57 f1-score
- Gradient Boosted: 0.57 accuracy, 0.52 f1-score
- **XGBoost: 0.61 accuracy, 0.54 f1-score**

Label	Day Range	# Rows
0	[0, 1]	7170
1	(1, 5]	3502
2	(5, 13]	4474
3	(13, 30]	5214
4	(30, 298]	4875

We selected XGBoost as the model to continue with…

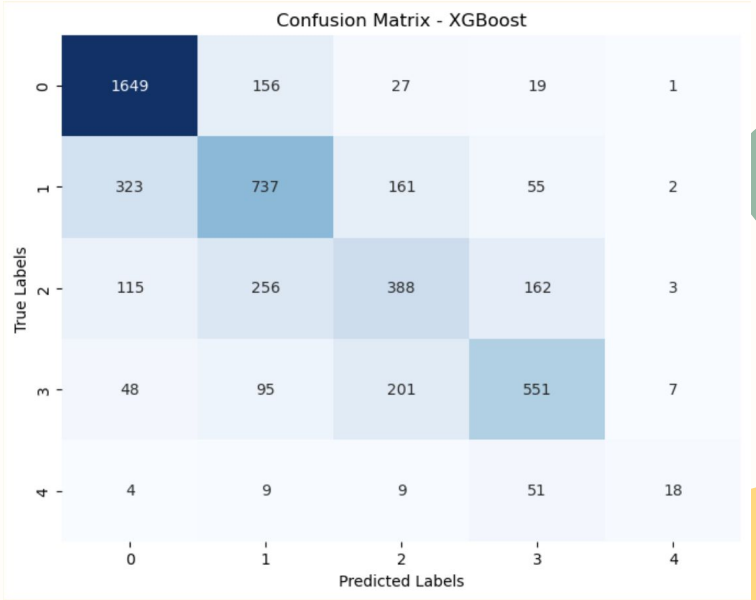
Bin adjustments helped increase the models usefulness for customers and slightly improved performance

Baseline XGBoost: 0.61 accuracy, 0.54 f1-score

Grid Search XGBoost: 0.66 accuracy, 0.56 f1-score

- Params:
  - colsample_bytree: 0.9
  - learning_rate: 0.1
  - max_depth: 7
  - n_estimators: 300
  - subsample: 0.8

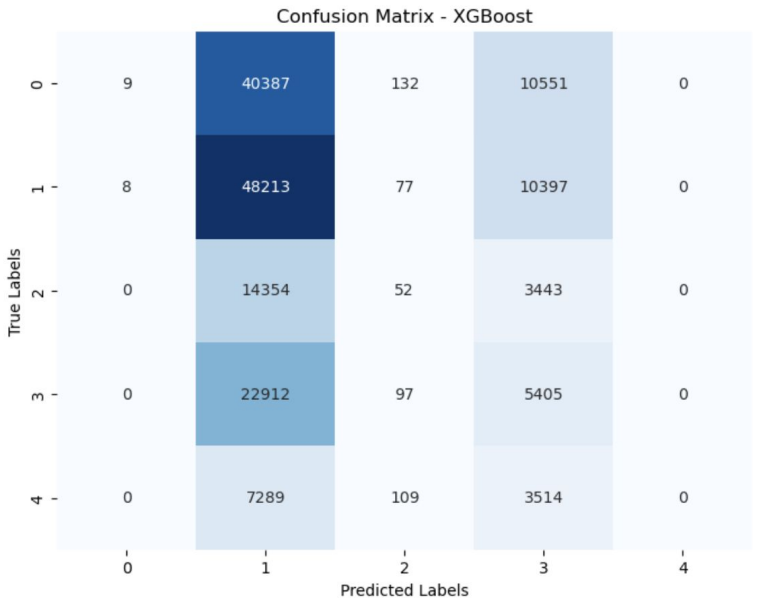
Label	Day Range	# Rows
0	[0, 3]	9362
1	(3, 14]	6327
2	(14, 30]	4671
3	(30, 100]	4398
4	(100, 298]	477



However, generalizability across shelters is low

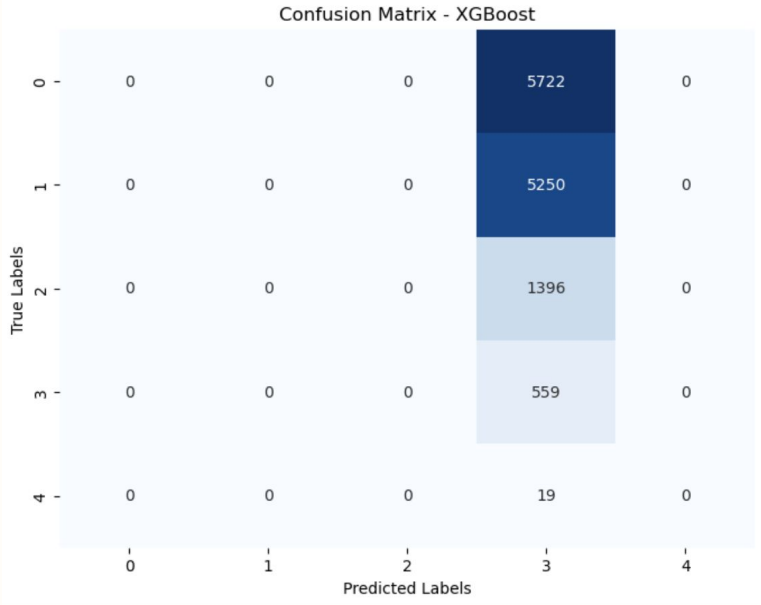
Austin Dataset

- XGBoost: 0.32 accuracy, 0.14 f1-score



Denver Dataset

- XGBoost: 0.04 accuracy, 0.02 f1-score



Low generalizability across shelters is due to data set inconsistencies and feature importance differences

Index	Feature Name (SONOMA)
f0	Type
f1	Breed
f2	Color
f3	Sex
f4	Size
f5	Kennel_Number
f6	Intake_Type
f7	Intake_Subtype
f8	Intake_Condition
f9	Intake_Jurisdiction
f10	Multipe_Visit_Count
f11	Age_inDays_at_Income
f12	Age_Group
f13	Is_Agressive
f14	Has_Name
f15	Is_Fixed
f16	Is_Mixed_Breed
f17	Is_Multicolor
f18	Color_Embedding_Cluster
f19	Breed_Embedding_Cluster
f20	Intake_Year
f21	Intake_Month
f22	Intake_Day
f23	Birth_Year
f24	Birth_Month
f25	Birth_Day

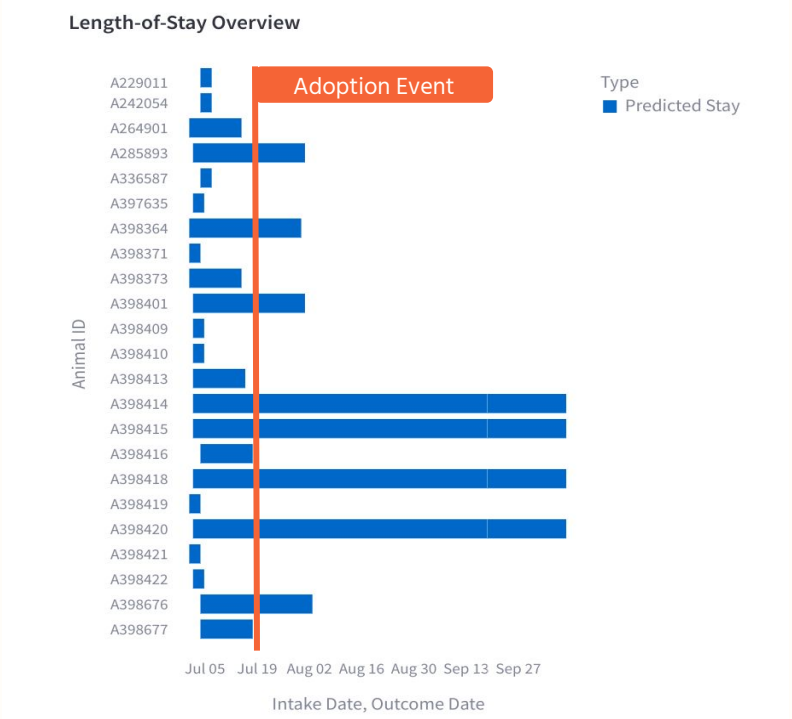
Index	Feature Name (AUSTIN)
f0	Type
f1	Breed
f2	Color
f3	Sex
f4	Size
f5	Kennel_Number
f6	Intake_Type
f7	Intake_Subtype
f8	Intake_Condition
f9	Intake_Jurisdiction
f10	Multipe_Visit_Count
f11	Age_inDays_at_Income
f12	Age_Group
f13	Is_Agressive
f14	Has_Name
f15	Is_Fixed
f16	Is_Mixed_Breed
f17	Is_Multicolor
f18	Color_Embedding_Cluster
f19	Breed_Embedding_Cluster
f20	Intake_Year
f21	Intake_Month
f22	Intake_Day
f23	Birth_Year
f24	Birth_Month
f25	Birth_Day

Index	Feature Name (DENVER)
f0	Type
f1	Breed
f2	Color
f3	Sex
f4	Size
f5	Kennel_Number
f6	Intake_Type
f7	Intake_Subtype
f8	Intake_Condition
f9	Intake_Jurisdiction
f10	Multipe_Visit_Count
f11	Age_inDays_at_Income
f12	Age_Group
f13	Is_Agressive
f14	Has_Name
f15	Is_Fixed
f16	Is_Mixed_Breed
f17	Is_Multicolor
f18	Color_Embedding_Cluster
f19	Breed_Embedding_Cluster
f20	Intake_Year
f21	Intake_Month
f22	Intake_Day
f23	Birth_Year
f24	Birth_Month
f25	Birth_Day

05.

Reports

Reports were designed for users who need to make resource allocation decisions for one or many animals



Hypothetical Example

As Sonoma Animal Shelter prepares for their Adoption Event on July 19, they may choose to dedicate more resources towards advertising the 4 animals who are predicted to remain at the shelter for another 2+ months

If we had more time, we would focus on making our solution more customizable to each shelter

- **Model training and bin adjustments** for each shelter
    - Denver: +.57 accuracy and +.37 f1-score once trained on the Denver data set
    - Austin: +.19 accuracy and +.24 f1-score once trained on the Austin data set.
- **Strategy recommendations** in generated reports based on shelter best practices
- **Intake and outcome data management system** options for shelters who don't have standardized data collection practices

Based on interviews, our reports show promise of being helpful to animal shelters

“Yes, this would help tremendously. We could potentially try to transfer pets to a foster based rescue that would be more suitable for their needs. Or even come up with different programs to keep the pet in their original home, like our training courses”

- *Humane Society of Memphis animal shelter*

06.

Wrap Up

Pawpulation Forecast aims to help animal shelters better allocate limited animal care resources by providing predictions for how long animals will remain in their care.

Acknowledgements

The following subject matter experts contributed heavily during the research and/or prototype testing phases of our project:

- Monica Dangler, Director of Pima Animal Care Center (PACC)
- Kaitlyn Pappas, PACC Employee
- Katie Hutchinson, PACC Animal Placement Manager
- Melanie Sobel, Director of Denver Animal Protection
- Sarah Siskin, Adoption Manager, Humane Society of Memphis
- Joscelyne Thompson, Intake Manager, Humane Society of Memphis

Thanks!