Data-driven Outcomes for Shelter Animals

Chloe, YeeCheng, Olivia

Problem & Data sets

Problem: 7.6 million animals enter US shelters annually.

Our solution

- Realise adoption trends (EDA)
- Predict shelter animal outcomes (ML)
- Recommend popular pet names (demonstration)

Conclusions

- Allocate care efficiently
- Tailor adoption strategies to communities

Data sets

- Austin
- Indiana
- California

Data Science Tools

RStudio	• Data visualization
Python	• Data cleaning • Machine Learning • Name Recommendation System

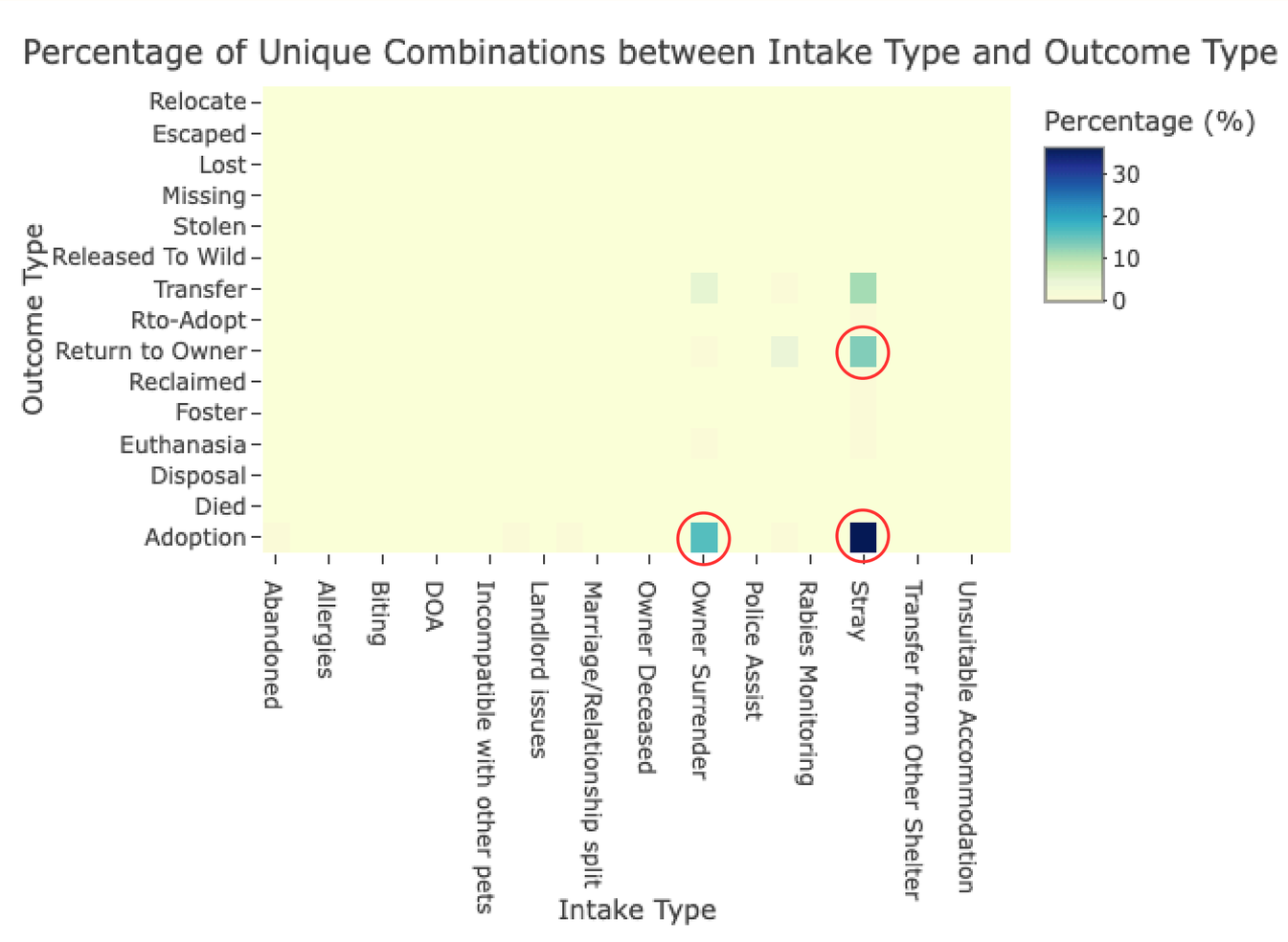
Data Cleaning

- **Feature selection**: Importance score
- **Merging**: Income and outcome tables (Austin), matching columns with other data sets
- **Discard**: California dataset

Exploratory Data Analysis

Packages: tidyverse, dplyr, tidyr, plotly (interactiveness), ggplot2.

Heatmap:



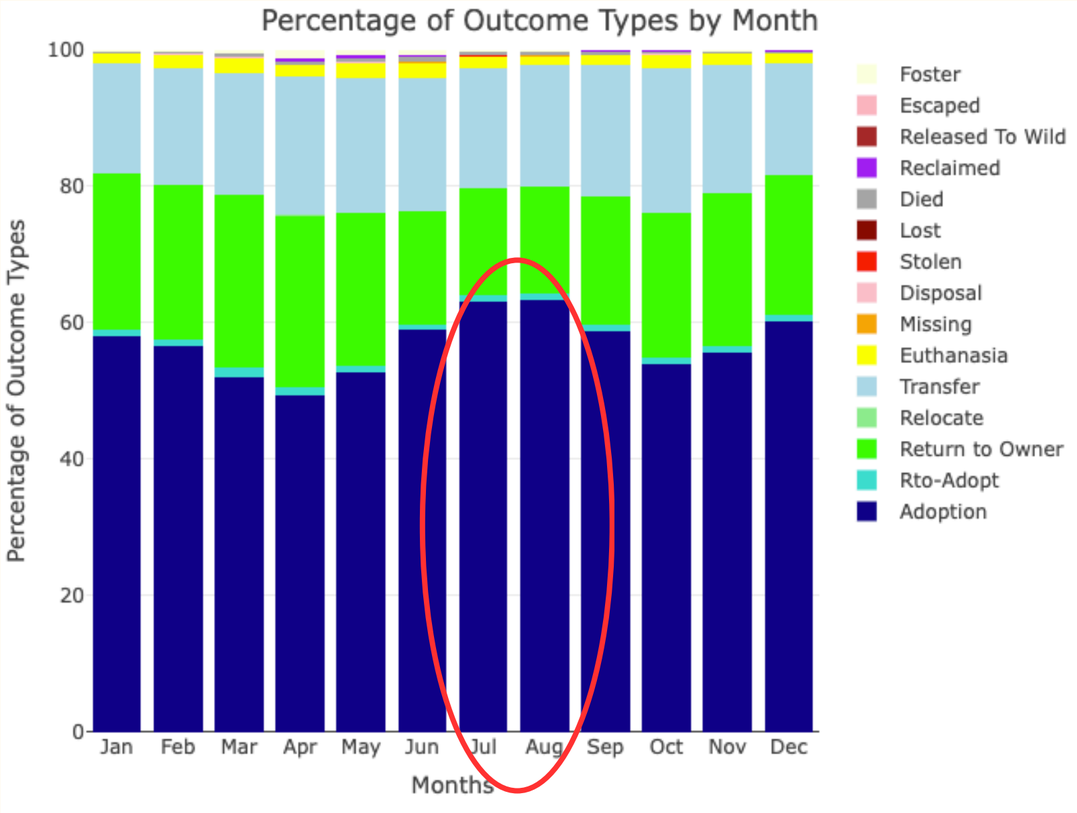
Visually Top 3 Unique Combinations

- **Stray & Adoption**:
  Least costly
  Without health and trauma issues

- **Owner surrender & Adoption**:
  Ready for adoption
  Established satisfactory conditions

- **Stray & Return to owner**:
  Lost pets categorized as stray
  Owners hopeful towards finding them

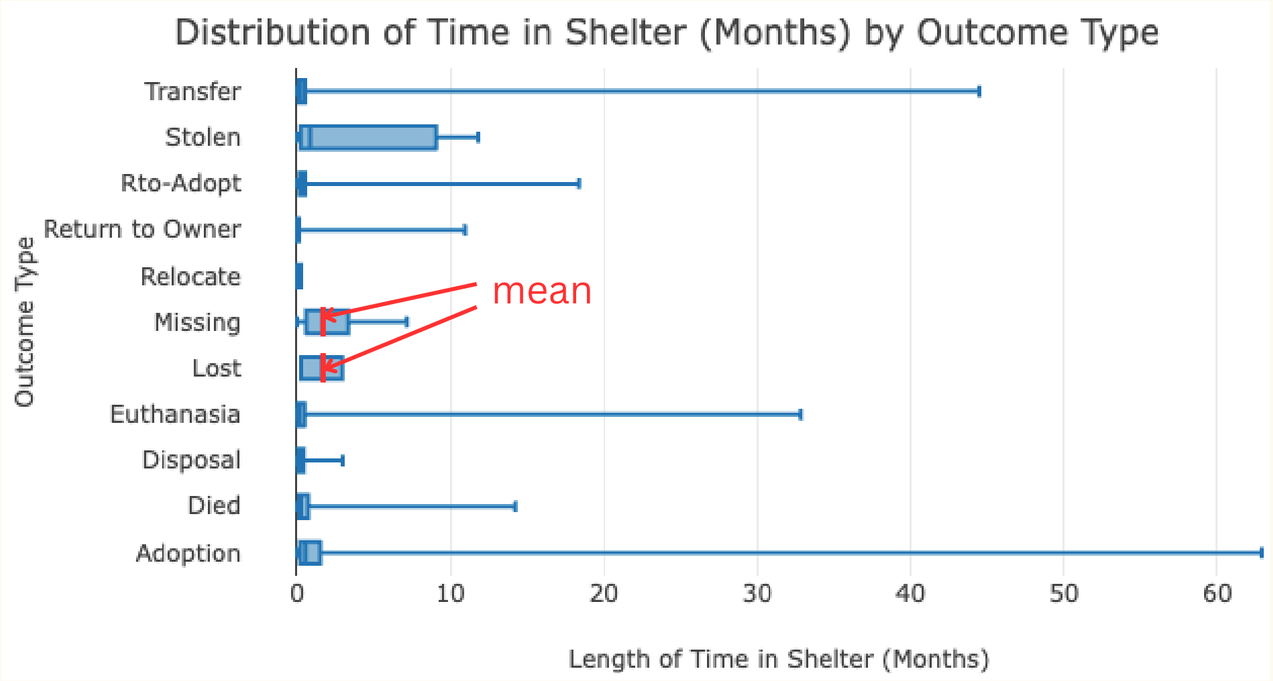
Time-series stacked bar graph:



- **Minimal (5~10%) fluctuations** over months

- **Higher adoption rates: July and August**
  - School holidays/summer vacation (US)
  - “Kitten/Puppy season”: Warmer weathers for mating & high influx of youngsters

Comparative box plot:



- **Shorter duration**:
  Transfer, RTO, relocate, euthanasia, disposal, died
  - Animals easily sorted and left shelter

- **Longer duration**:
  Stolen, missing, lost
  - Takes time to identify pets and their owners

  Adoption
  - Adoption depends on people’s varied preferences

Pre-processing - Feature Engineering

Technique	Feature
Label Encoding	Name
1-Indexed Numerical Mapping	Animal Type, Outcome, Sex, Intake and Outcome Type
One hot Encoding	Breed, colour
Numerical	Age
Pandas DateFrame, Radian Conversion	Date

Animal_ID	Colour
A1229291	Black/White
A7868294	Yellow Black
A8343929	White

Animal_ID	Colour_Black	Colour_White	Colour_Yellow
A1229291	1	1	0
A7868294	1	0	1
A8343929	0	1	0

Shelter Outcome Predictions via ML

1. Principle Component Analysis (PCA)

- 124,914 rows & 9 features - reducing dimensionality, retain 95% variance
- PCA applied to every ML model, compared with non-PCA
- Avoid overfitting on training data

2. Repeated 5-fold Cross Validation (CV)

- Evaluate model's ability to generalise to unseen data
- Splits training data into 5 parts, CV repeated 10 times
- Bar plots comparing 5 evaluative metrics across models

3. Testing & Evaluating

ML models:

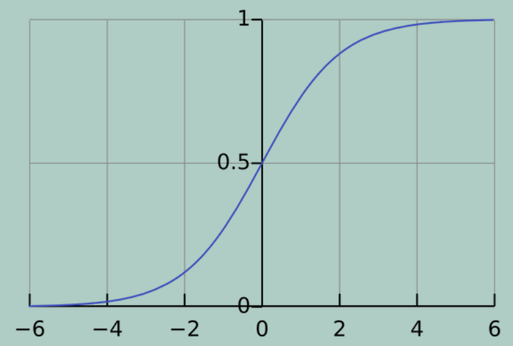
- Logistics Regression
- k-Nearest Neighbour
  Random Forrest
- Support Vector Machine
  XGBoost
- Neural Network

Evaluative metrics:

- Accuracy, precision, recall, F1 score, log loss

ML Models

1. **Logistics Regression**:
   - Sigmoid function: Probability of prediction class [0,1]



2. **kNN**:
   - Predicting based on “nearest” data points
   - Euclidean distance

$$|X - Y| = \sqrt{\sum_{i=1}^{i=n} (x_i - y_i)^2}$$

3. **Random Forest**:
   - Ensemble method
   - Multiple decision trees (randomly selected data samples and features)
   - Trees vote on the final prediction

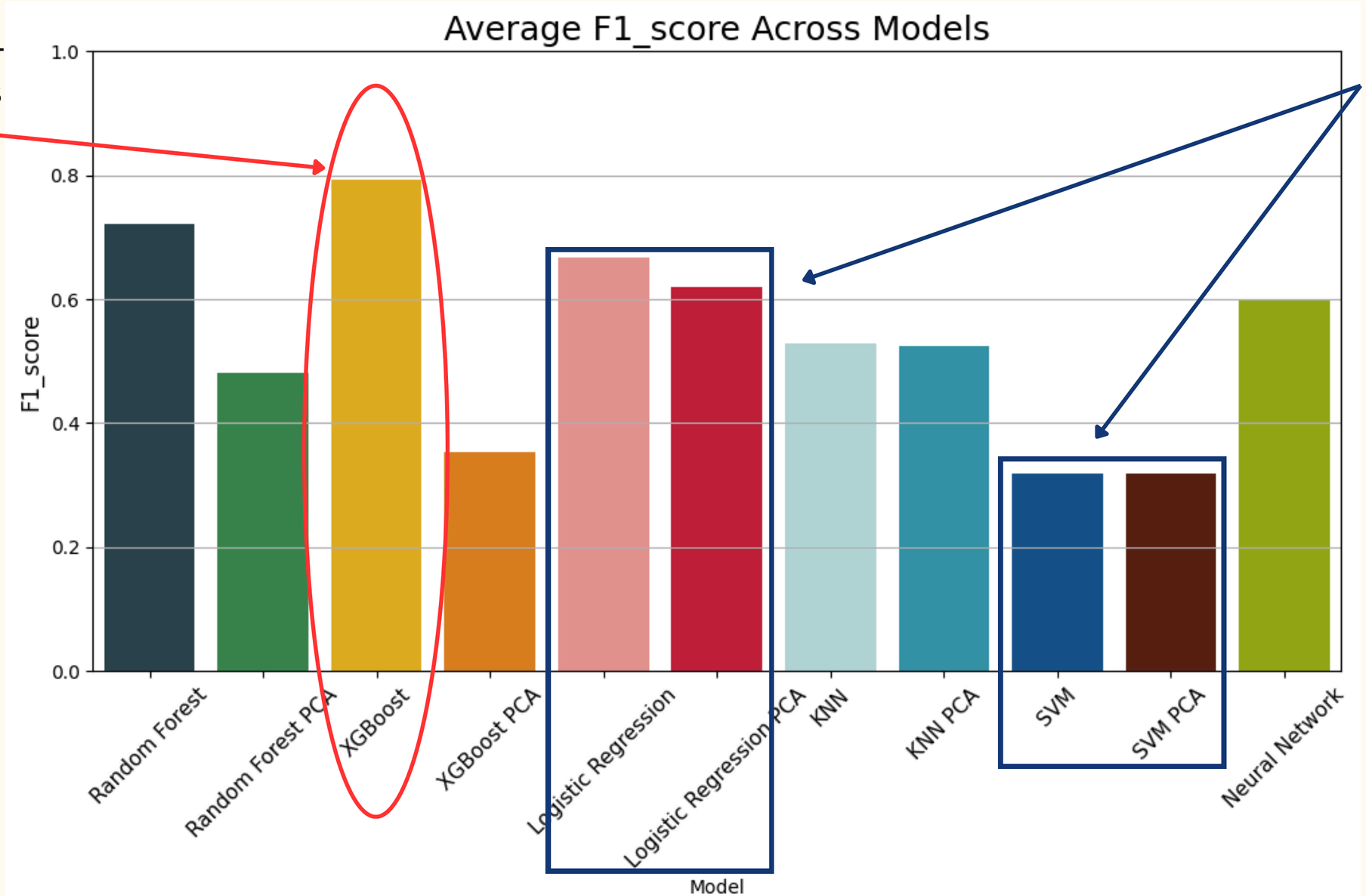
4. **XGBoost**:
   - Sequentially builds models
   - Each model corrects previous errors

5. **SVM**:
   - Supervised learning algorithm
   - Finds optimal hyperplane separating different classes in feature space
   - Maximize margin between the classes, ensuring best separation

6. **Neural Network**:
   - Interconnected layers of nodes (neurons)
   - Each node processes input data, passes data to the next layer

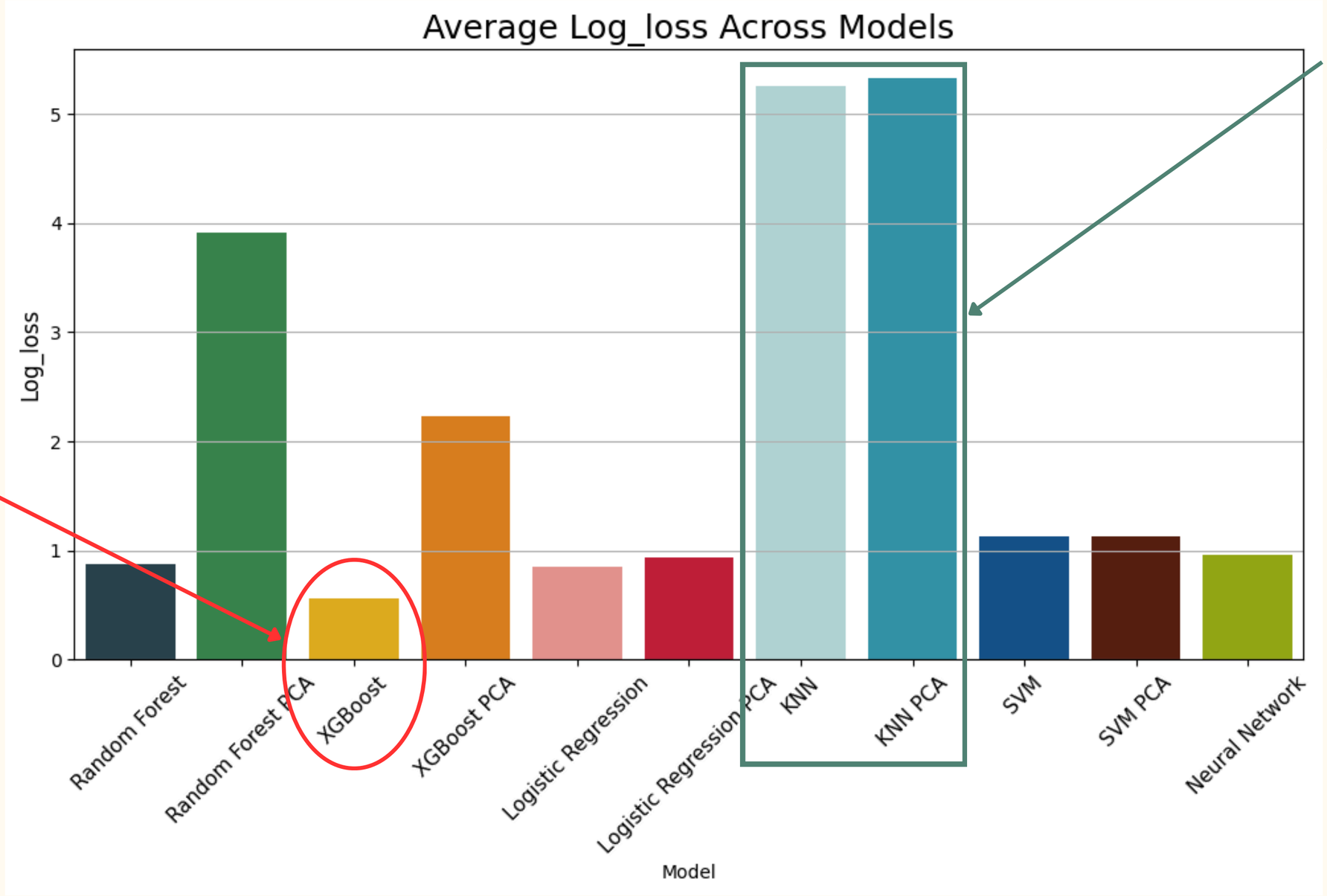
Repeated 5-fold CV Bar Plots - F1 Score

Highest F1 Score - XGBoost performs the best on positive and negative classes.



PCA only reduces running time for linear models, but minimal improvement to accuracies.

Repeated 5-fold CV Bar Plots - Log Loss



Lowest Log Loss - least penalisation in XGBoost indicates least amount of incorrect predictions.

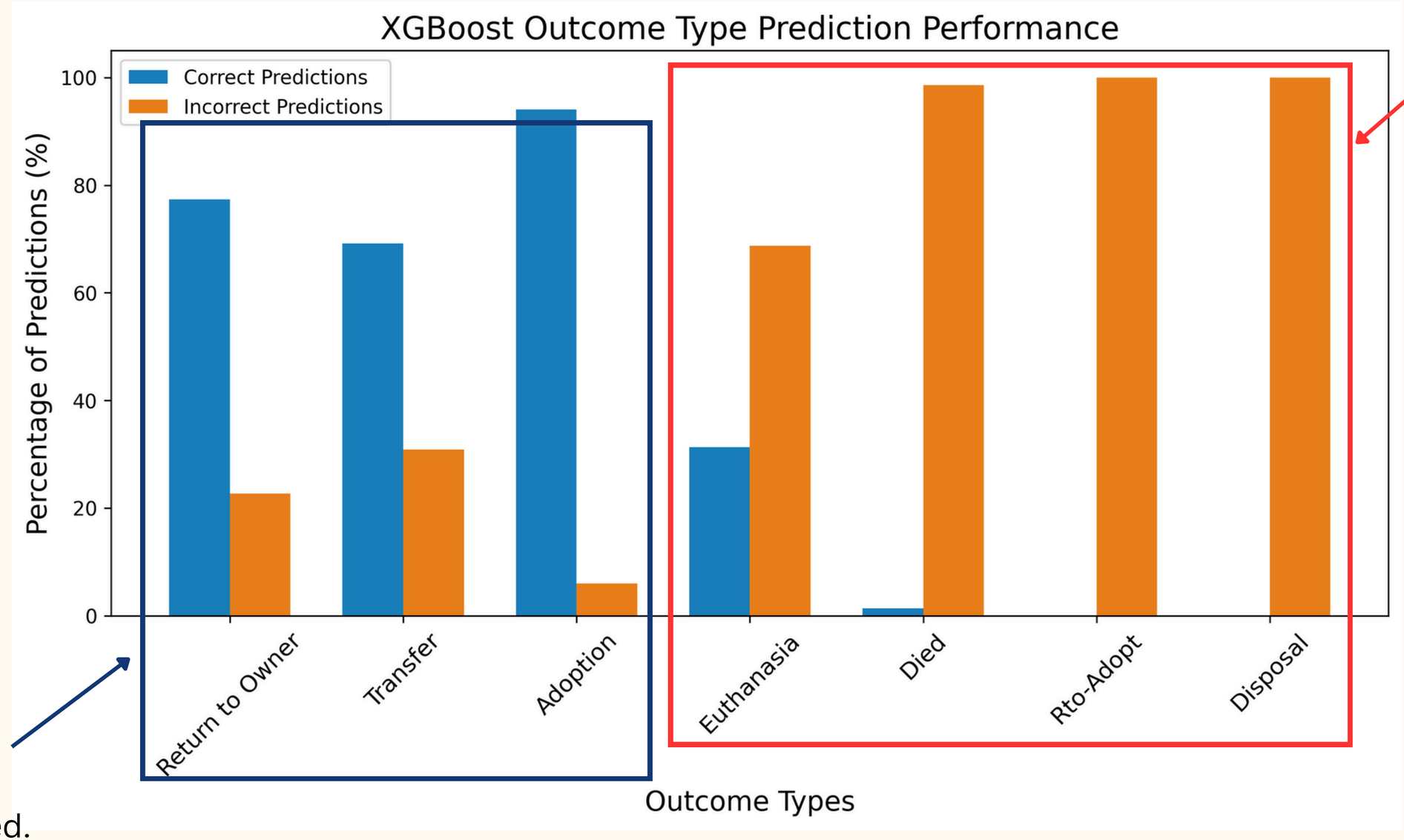
Highest Log Loss - In high-dimensional spaces, similar distances blur important features.

PCA distorts distances.

Evaluating ML performance on test data

Model	Scores				
	Accuracy	Precision	Recall	F1_Score	Log_loss
RF	0.8145	0.8057	0.8327	0.8143	0.8700
RF PCA	0.4892	0.5449	0.4903	0.5114	3.896
LR	0.7890	0.7532	0.7952	0.7601	0.6920
LR PCA	0.7258	0.6318	0.7560	0.6636	0.8708
XGB	0.8446	0.8273	0.8467	0.8352	0.5406
XGB PCA	0.2500	0.2614	0.2680	0.2138	4.950
KNN	0.7342	0.7198	0.7400	0.7261	3.927
KNN PCA	0.7342	0.7198	0.7400	0.7261	3.927
SVM	0.2209	0.4667	0.2296	0.2493	1.491
SVM PCA	0.6519	0.6385	0.7976	0.7091	1.078
NN	0.7714	0.7869	0.7886	0.7757	0.7509

A Look into XGBoost



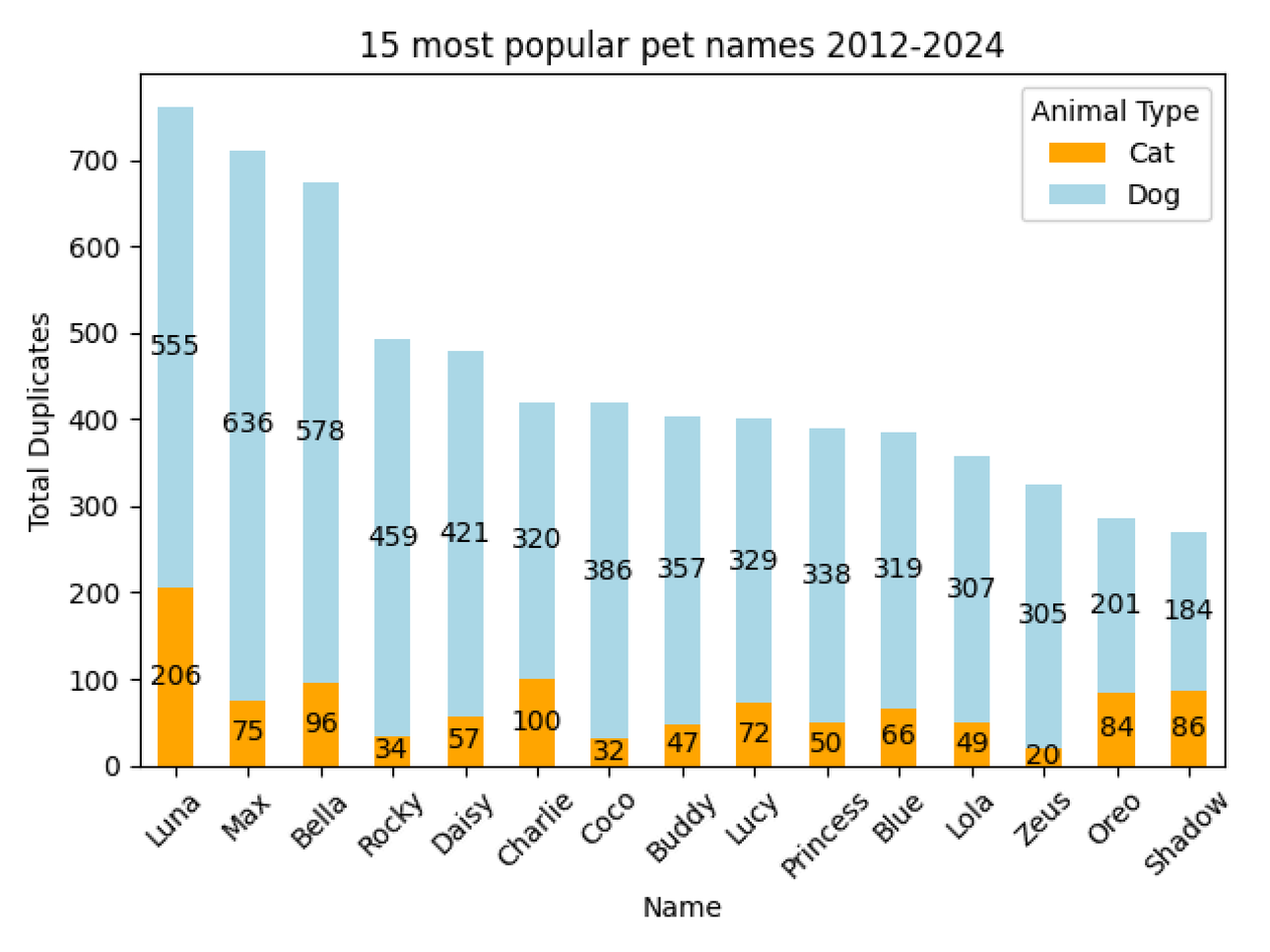
Many incorrect predictions for pets with health issues.

Many correct predictions for pets with owners/adopted.

Conclusions

	Insight	Solution	Effect
1. Allocate care more efficiently	Low adoption rate, poor ML predictions for pets with health issues	• Boost health to satisfactory level for adoption • Euthanisation	• Increase chances for adoption • Reduce shelter time • Make space for more pet intake
2. Tailor adoption strategies to communities	Highest adoption rates in July and August	• Advertise adoption in other months • Increase pet intake during July and August	• Adopters have diverse choice
	Longer shelter time for missing and lost pets	• Conduct surveys for community's pets	• Easier location of missing and lost pets • Reduce shelter time

Recommendation System



Motivation

- Pet names duplication
- Recommends uniquely tailored pet names

Recommendation System

Implementation Methods

kNN Distance Types

- Euclidean
- Manhattan
- Chebyshev
- Hamming
- Canberra
- Braycurtis

TF-IDF Encoding

$$w_{x,y} = tf_{x,y} \times \log\left(\frac{N}{df_x}\right)$$

TF-IDF

Term x within document y

$tf_{x,y}$ = frequency of x in y

df_x = number of documents containing x

N = total number of documents

TF-IDF balances word frequency in a document with how rare the word is across the dataset to represent its importance.

Recommendation System

Demonstration

Features:

- Animal Type: Dog
- Breed: Labrador Retriever/Pit Bull
- Sex: Spayed Female
- Colour: Black/White
- Age: 3

Give her a name!

Now let's use OptiPaw's recommendation system...

Image retrieved from
https://www.reddit.com/r/labrador/comments/1akghug/lifespan_of_lab_mixes/#lightbox

Thank you!