Predicting Risk in Shelter Animals

Course Project – Data Mining 2025

Kfir Diamond & Daniel Tugendhaft
Lecturer: Louisville Metro KY Analysis Team

Section 1:
The Problem & Its Significance

The Challenge of Animal Welfare

Animal shelters across the U.S. face constant overpopulation and limited resources. Decisions at the moment of intake are critical.

The Core Problem: Identifying animals at high risk of euthanasia, death, or disposal immediately upon arrival.

Unlike inventory problems, the cost of error here is *irreversible and ethical*. Missing a high-risk animal means missing a chance for life-saving intervention.

Original Project Objectives

System Architecture

Creating an automated process to translate noisy administrative logs into structured ML-ready data.

Handling Imbalance

Implementing techniques like cost-sensitive learning to prevent bias toward the majority "Safe" class.

Risk Profiling

Using Unsupervised Learning to discover latent patterns and characteristics common to negative outcomes.

Dataset and Core Features

Feature Category	Attributes Included	Count / Context
Static Profile	Breed, Color, Sex, Animal Type	64,000+ Records
Situational	Intake Type, Condition, Intake Date	Louisville Metro Data
Operational	Kennel ID, Jurisdiction	Proxy for health status
Target (Binary)	is_at_risk	Positive: Euthanasia/Death

Engineering Meaningful Signals

- **Temporal Seasonality:** Extracted Month and Season to capture trends like "Kitten Season" (high summer intakes).
- **Breed Complexity:** Created a binary *mixed_breed* feature by parsing string data for keywords like "Mix" or "/".
- **Recidivism (first_time):** Identified returning animals, often linked to behavioral issues or chronic health problems.
- **Age Standardization:** Normalized units (days/years) into a continuous *Age_Years* and binned into life stages.

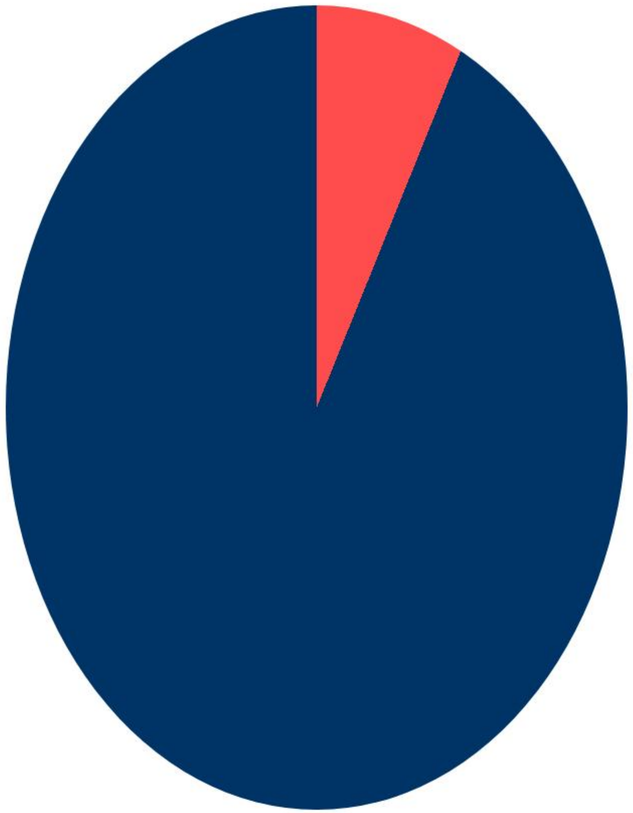
Methodology: LightGBM

Primary Model Choice

We pivoted from SVM/MLP to **LightGBM** for its leaf-wise tree growth strategy and native categorical handling.

Rationale: Handles missing values natively and avoids the "curse of dimensionality" inherent in One-Hot Encoding high-cardinality features like Breed.

Addressing the Class Imbalance



At Risk (6.2%): Primary focus.

Not At Risk (93.8%): Majority class.

Strategy: Cost-Sensitive Learning using inverse frequency weights to penalize misclassified risk cases.

Unsupervised: K-Modes Clustering

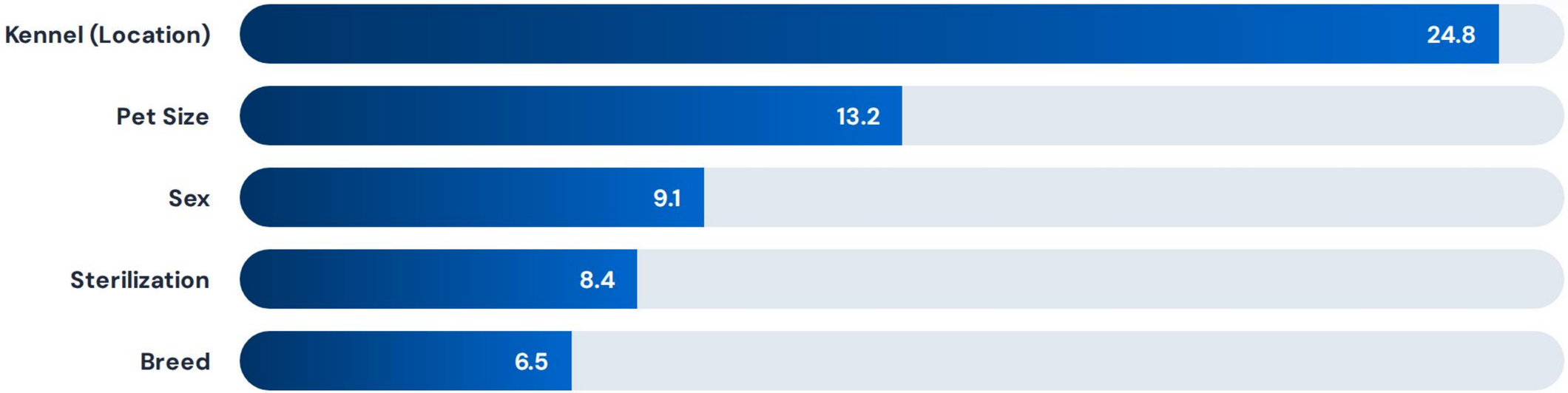
Why K-Modes?

Standard K-Means relies on Euclidean distance, which fails for categorical data like *Color* or *Breed*. K-Modes uses matching dissimilarity and cluster **Modes**.

Objective

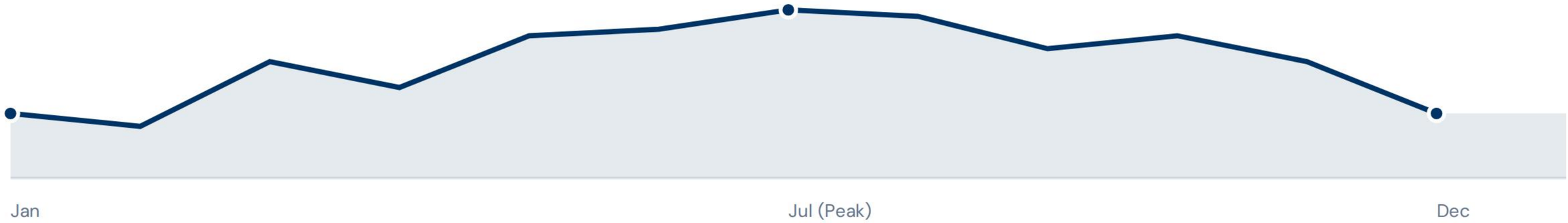
Identify "Risk Archetypes"—combinations of situational factors (e.g., Intake Type + Season) that correlate with high mortality without labeled data bias.

Key Predictors of Risk



Kennel Location is a powerful proxy for initial medical triage status.

Seasonal Risk Trends (Mortality %)



Cluster Profile (Risk Group 3): Stray Black Cats in Summer exhibit a 16.6% mortality rate, validating "Kitten Season" strain.

Conclusions & Insights

Success Metric: Achieved ~75% Recall. We successfully flag 3 out of 4 truly at-risk animals.

Future Work: Applying NLP to free-text medical notes and deploying the model as a triage API for shelter staff.

Thank You! Questions?

Kfir Diamond & Daniel Tugendhaft | Data Mining 2025

Image Sources

https://media.istockphoto.com/id/531610325/photo/homeless-dog-behind-bars.jpg?s=612x612&w=0&k=20&c=Z041oPng2ag3md5bfOdZbO2138K724sRNz9EutkmEqo=

Source: www.istockphoto.com

https://img.freepik.com/premium-vector/abstract-futuristic-technology-network-background-with-glowing-blue-lines-nodes-dark-background-connecting-geometric-shapes-data-points-innovation-concepts_1222051-791.jpg

Source: www.freepik.com

https://preview.redd.it/i-failed-a-stray-cat-that-i-loved-and-i-think-about-him-all-v0-e856w4c4cxie1.jpeg?width=640&crop=smart&auto=webp&s=8cbefa067a4794269b756fd3ef26b0f15b573a59

Source: www.reddit.com