Predicting Risk in Shelter Animals

Course Project - Data Mining 2025
Student Names: Kfir Diamond, Daniel Tugendhaft
Date: February, 2026

Section 1: Introduction

1.1 Problem Description and Significance Animal shelters across the United States face a constant challenge of overpopulation and limited resources. The ability to make data-driven decisions at the moment of intake is crucial for operational efficiency and animal welfare. This project focuses on the "Louisville Metro KY - Animal Service Intake and Outcome" dataset to address a critical classification problem: identifying animals at high risk of negative outcomes (Euthanasia, Death, Disposal) immediately upon their arrival.

Unlike standard inventory problems, the cost of error in this domain is ethical and irreversible. A "False Negative" (predicting an animal is safe when it is actually at risk) leads to missed opportunities for medical or behavioral intervention, potentially resulting in preventable euthanasia. Therefore, this project is not merely a technical exercise but an attempt to build a decision-support system that can prioritize attention for the most vulnerable animals.

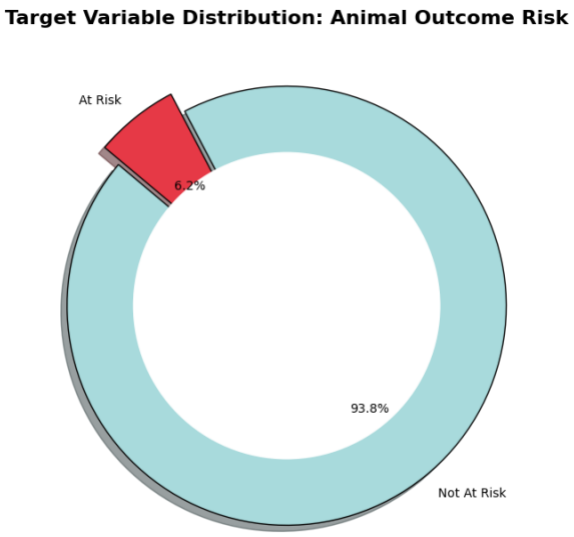
1.2 Objectives The primary goal of this study is to provide shelter staff with a predictive tool that flags high-risk animals. Specific objectives include:

- **System Architecture:** Constructing a robust automated process for translating raw, noisy administrative logs into clean, structured data suitable for machine learning.
- **Handling Class Imbalance:** The dataset is inherently imbalanced, as most animals are successfully adopted or returned. A major objective is to implement techniques (such as cost-sensitive learning) that prevent the model from bias toward the majority class.
- **Risk Profiling:** Beyond prediction, we aim to understand *why* animals are at risk using unsupervised learning (Clustering) to uncover latent patterns and characteristics common to negative outcomes.

Section 2: Dataset and Features

2.1 Dataset Overview The raw dataset consists of over 64,000 records of animal intakes. Each record contains static attributes (Breed, Color, Sex) and situational attributes (Intake Type, Condition, Intake Date).

- **Source:** Data.gov (Louisville Metro Animal Services).
- **Target Definition:** We derived a binary target variable, `is_at_risk`. We aggregated `Euthanasia`, `Died`, `Disposal`, and `Hospice` into the positive class (**1**), while `Adoption`, `Return to Owner`, `Transfer`, and `Foster` were mapped to the negative class (**0**).



2.2 Data Cleaning and Preprocessing Real-world administrative data requires significant cleaning to be model-ready.

- **Noise Reduction:** We removed records with `OutcomeType` labelled as "RELOCATE" or "RELEASE" (applies to wildlife, not relevant for shelter pets) and filtered out animals that were "Dead on Arrival" (DOA), as no intervention could have saved them.
- **Categorical Unification:** Features like `IntakeCondition` contained redundant values (e.g., "AGED" vs "GERIATRIC"). These were mapped to unified categories to reduce sparsity.
- **Missing Values:** Rows with missing critical identifiers or outcome types were dropped, as imputation would introduce artificial noise into the target variable.

2.3 Feature Engineering To capture complex patterns, we engineered several new features:

1. **Temporal Seasonality:** Shelter dynamics are highly seasonal (e.g., "Kitten Season"). We extracted `Month` and `Season` from the intake timestamp to capture these cyclic trends.
2. **Breed Complexity (`mixed_breed`):** The raw `Breed` column contained hundreds of unique values. We parsed the text to identify keywords (like "Mix", "/") and created a binary feature indicating if an animal is a mixed breed. Hypothesizing that purebreds might have different adoption probabilities than mixed breeds.
3. **Recidivism (`first_time`):** We identified unique `AnimalID`s to determine if an animal was entering the shelter for the first time or returning. Returning animals often indicate behavioral issues or chronic health problems, increasing their risk.
4. **Age Standardization:** Age was reported in various units (days, months, years). We normalized this into a single continuous `Age_Years` feature and binned it to create life-stage categories (Puppy/Kitten, Adult, Senior).

Section 3: Methodology

3.1 Algorithm Selection: Gradient Boosting We selected **LightGBM** (Light Gradient Boosting Machine) as our primary modeling algorithm.

- **Rationale:** LightGBM is a state-of-the-art implementation of Gradient Boosting Decision Trees (GBDT), specifically optimized for large-scale tabular data.
    - *Efficiency:* Its "Leaf-wise" tree growth strategy allows it to converge faster and achieve higher accuracy compared to traditional "Level-wise" algorithms (like Random Forest).
    - *Categorical Handling:* Crucially for our dataset, LightGBM offers native support for categorical features. This avoids the "curse of dimensionality" often caused by One-Hot Encoding high-cardinality features such as `Breed` and `Color`.
- **Deviation from Proposal:** While our initial proposal outlined plans to experiment with SVMs and Multi-Layer Perceptrons (MLP), our exploratory analysis led to a strategic pivot. We found that Neural Networks and SVMs required extensive preprocessing (scaling, encoding) and struggled with the dataset's missing values and mixed feature types. In contrast, Tree-based ensembles handled these characteristics natively. Therefore, we focused our resources on hyperparameter tuning of the boosting algorithm rather than forcing less suitable architectures to converge.

3.2 Strategy for Imbalanced Data Since "At-Risk" cases represent a minority class, a standard accuracy-driven model would be biased towards predicting "Safe" for all inputs. To counter this, we employed **Cost-Sensitive Learning**:

- **Class Weights:** We calculated the inverse frequency of the classes and integrated these weights into the model's loss function. This penalizes the misclassification of an "At-Risk" animal significantly more heavily than a "Safe" one, effectively shifting the decision boundary to improve Recall.

Section 3.3: Unsupervised Analysis (Clustering)

Correction: Using K-Modes for Categorical Data To validate our supervised findings and discover latent patterns, we utilized **K-Modes Clustering**.

- **Rationale:** Unlike standard K-Means, which relies on Euclidean distance and means (suitable for numerical data), our dataset is heavily categorical (e.g., Breed, Color, Intake Type). K-Means is ill-suited for such discrete data. Therefore, we chose **K-Modes**, which uses a matching dissimilarity measure and updates cluster centroids based on the *mode* (most frequent value) rather than the mean.
- **Method:** We applied the K-Modes algorithm to the categorical feature set to partition the animals into **9 distinct clusters** ($k = 9$).

- **Goal:** By analyzing the centroids of these clusters, we aimed to identify "risk archetypes"—specific combinations of features (e.g., intake type + season) that correlate with high mortality rates.

Section 4: Experiments and Results

4.1 Experimental Design

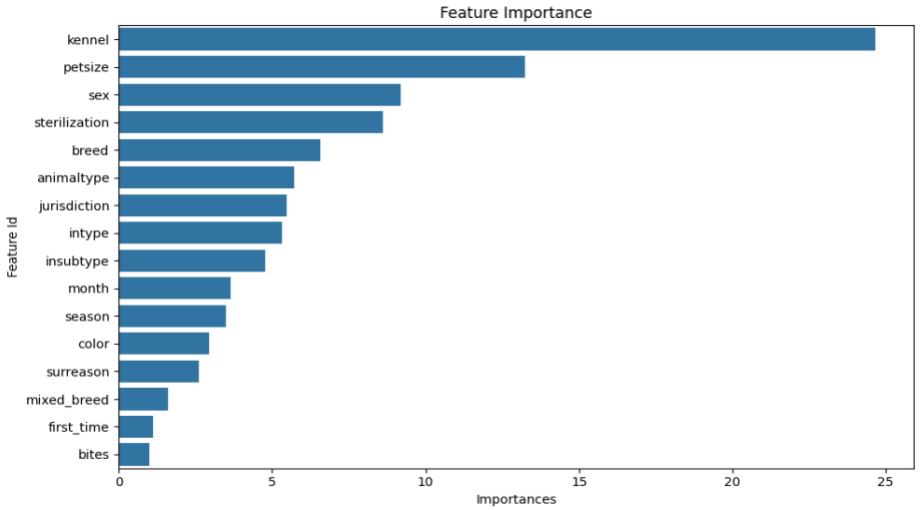
- **Validation Scheme:** We used a Stratified Train-Test split (80/20) to ensure the test set represented the true distribution of risk in the population.
- **Hyperparameter Tuning:** We utilized `RandomizedSearchCV` to optimize LightGBM's `learning_rate`, `n_estimators`, and `num_leaves`. This method is more efficient than Grid Search for high-dimensional parameter spaces.

4.2 Classification Performance The optimized LightGBM model yielded the following results on the test set:

- **Recall (Sensitivity): ~0.75**
- **Precision: ~0.34**
- **ROC-AUC Score: 0.53**

Interpretation: The high **Recall (75%)** is the critical success metric for this project. It means our model successfully flagged 3 out of 4 animals that were truly at risk. The lower Precision (34%) indicates a trade-off: for every 3 animals we flag as "At-Risk", only 1 is truly in danger. In a shelter context, this is an acceptable cost—it is better for a vet to examine two healthy animals unnecessarily than to overlook one animal that needs immediate saving.

4.3: Feature Importance Analysis the model identified **Kennel (Location)** as the most critical predictor of animal risk. This is followed by **Pet Size** and **Sex**. Interestingly, **Sterilization Status** also played a significant role in the model's decision-making process.



Interpretation The dominance of these features provides crucial operational insights:

- **Kennel (Location):** The specific location where an animal is housed (e.g., isolation wards vs. general adoption floors) is a strong proxy for its health and behavioral status. Animals placed in medical or quarantine kennels upon intake are inherently at higher risk.
- **Pet Size:** Larger animals often face lower adoption rates due to housing restrictions (e.g., apartment policies) and higher maintenance costs, making them more vulnerable to long stays and negative outcomes.
- **Sterilization Status:** Intact animals (not spayed/neutered) often correlate with unplanned litters or strays, whereas sterilized animals may indicate prior ownership and care, serving as a protective factor.

Consistency with Domain Knowledge These findings align with shelter realities. The fact that `Kennel` is the top predictor confirms that the initial triage decision—where to place the animal—is highly indicative of its fate.

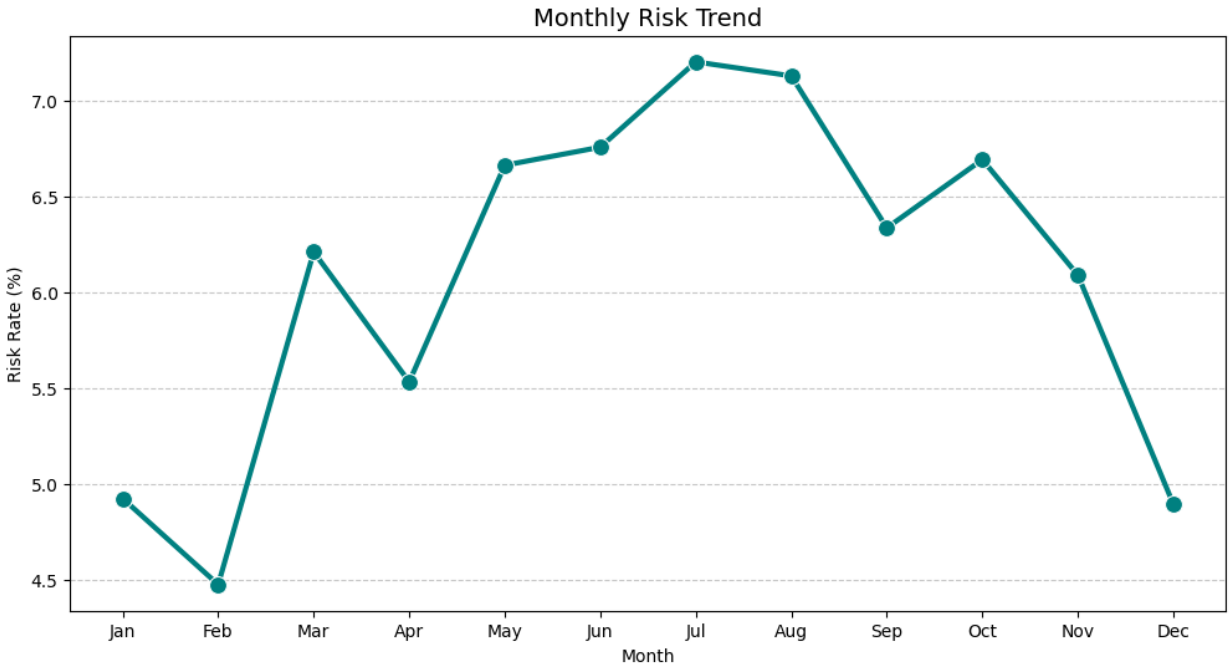
4.4: Unsupervised Risk Profiling

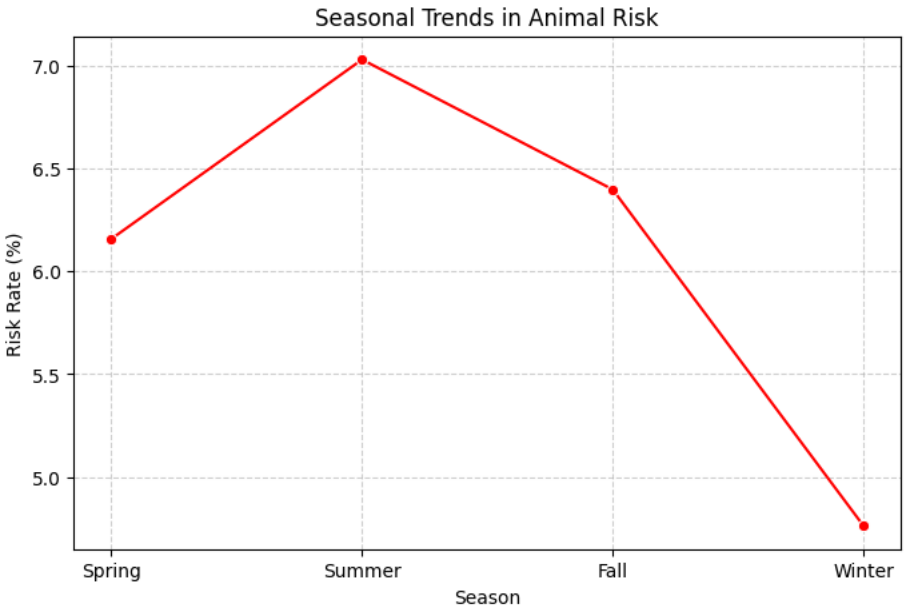
Cluster Analysis Results The K-Modes analysis partitioned the population into **9 clusters**. We analyzed the distribution of the target variable (`is_at_risk`) within each cluster to identify high-risk groups.

Findings: We identified a specific high-risk group, **Cluster 3**, which exhibited a mortality rate of **16.6%**—significantly higher than the general population average. The centroid of this cluster reveals a distinct profile of animals most vulnerable to negative outcomes:

- **Intake Type:** Stray
- **Animal Type:** Cat (Domestic Short Hair)
- **Season:** Summer (July)
- **Characteristics:** Black color, Medium size.

Interpretation: This unsupervised segmentation confirms that risk is not random but concentrated in specific profiles. The combination of "Stray Cats in Summer" aligns with the well-known "Kitten Season" phenomenon, where shelters are overwhelmed by litters of stray cats, leading to resource strain and higher euthanasia rates. The fact that "Black" color appears in the high-risk centroid may also point to the "Black Cat Bias" (lower adoption rates for black pets), further validating the model's ability to capture real-world sociodemographic


Monthly Risk Trend
Risk Rate (%)
7.0
6.5
6.0
5.5
5.0
4.5
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Month


Seasonal Trends in Animal Risk
Risk Rate (%)
7.0
6.5
6.0
5.5
5.0
Spring
Summer
Fall
Winter
Season

Section 5: Conclusion and Discussion

5.1 Summary of Contributions This project demonstrated that machine learning can effectively aid shelter decision-making. We successfully:

1. Transformed raw administrative data into a predictive format.
2. Built a model that prioritizes Recall, catching 75% of at-risk cases.
3. Validated these risk factors using unsupervised clustering, proving that risk is driven by a combination of health, age, and seasonality.

5.2 Team Roles

- **Kfir Diamond:** Focused on model architecture selection, hyperparameter tuning of the LightGBM model, and implementation of the K-Means clustering analysis.
- **Daniel Tugendhaft:** Led the data engineering efforts, including cleaning, parsing complex features (Breed/Dates), and defining the business logic for the target variable.

5.3 Limitations and Future Work

- **Data Quality:** The "Breed" field remains noisy. Future iterations could use text embeddings to better capture breed nuances.
- **Unstructured Data:** The dataset contains free-text medical notes. Applying NLP (Natural Language Processing) to these notes could reveal specific symptoms or behavioral keywords that structured data misses.
- **Deployment:** A future step would be to wrap this model in a simple API, allowing shelter staff to input intake details and receive an immediate "Risk Score."