CMSC 35300 Final Project

Andrew Dunn
University of Chicago
Chicago, IL 60637
andrewdunn@uchicago.edu

Betty Fang
University of Chicago
Chicago, IL 60637
bfang@uchicago.edu

Abstract

Using 8 years of data from the City of Austin's Animal Center, we predict the outcomes of cats held at the Center. A multinomial logistic regression gradient descent model predicts the 6 possible outcomes with about 50 % accuracy. The model can be improved further with more feature engineering and refined oversampling.

1 Problem

We sought to predict the outcomes of cats (n=67,436) in the City of Austin's Animal Center from October 2015 to November 2023 using information collected about each cat at the time of intake. We used a multinomial logistic regression to develop a multi-label classification model to predict 6 outcomes: Adoption, Foster, Return to Owner, Transfer, Euthanasia/Died, and Non-Outcome.

1.1 Context

A core function of animal shelters is the intake of lost, abandoned, or surrendered domesticated animals such as dogs and cats. Animal shelters promote animal welfare by reuniting lost pets with owners, adopting out animals to interested owners, and placing animals with foster families. The capacity of shelters, however, is constrained by resources such as kennel space, staff, and funding. In addition to positive shelter outcomes (reunification, adoption, fostering), negative outcomes include euthanasia, which can be a consequence of behavior, medical needs and costs, or lack of space.

While U.S. animal shelters strive to minimize negative outcomes, intakes and euthanasia rates have spiked significantly post-COVID-19. This trend has been documented across the U.S. in cities such as Chicago, Los Angeles, Austin, and New York. Economic downturns, behavioral issues due to under-socialization during COVID-19, low animal sterilization rates, and non-licensed breeding are all factors that can contribute to over-crowding and, by extension, higher euthanasia rates. Although alleviating over-crowding requires upstream policy interventions, optimizing the existing resources at shelters can help reduce negative outcomes. With predictive modeling, shelters can better allocate resources towards factors associated with positive animal shelter outcomes and in turn increase these positive outcomes.

1.2 Literature Review

A review of research in this area suggests that predicting animal shelter outcomes is well-suited for multi-label classification problems in machine learning and a topic with contemporary applications in animal welfare and public policy.

Predictive modeling methods that have been applied to animal shelter data include logistic regression, neural networks, random forest algorithms (Bradley & Rajendran, 2023), structural equation modeling (Kilgour & Flockhart, 2022), decision trees, and gradient boosting (Zadeh et al., 2022). Researchers have used public data containing intake information about animals and online adoption profiles to build models that predict shelter outcomes. Because this topic can be approached as a binary classification problem or a multi-label classification problem, different researchers have sought to optimize variations of the same problem.

Bradley & Rajendran (2023) used categorical features (species, breed, color, gender, age, location, and outcome type) to predict length of stay (low, medium, high, very high) for dogs and cats. Kilgour & Flockhart (2022) predicted cat outcomes in a Washington, D.C. shelter using characteristics (coat, color, age), location, intake date and type, and length of stay. Zadeh et al. (2022) featurized information from online pet adoption profiles (fee, amount of photos and videos on the profile, sentiment score, name, fur length, and word count) to predict length of stay.

Mitrović et al. (2019) compared 5 algorithms (K-Nearest Neighbors (KNN), Naive Bayes, C4.5, Random Forest, Support Vector Machines (SVM)), and concluded that KNN and C4.5 performed the best on the Austin Animal Center data. They also grouped outcomes into only 3 labels: Good (Returned to Owner, Adopted, Returned To Owner-Adopted), Neutral (Relocated, Transferred), and Bad (Died, Disposed, Missing, and Euthanized).

2 Methodology

2.1 Data

The two datasets used can be found on the City of Austin's Open Data Portal: Austin's Animal Center Intakes and Austin's Animal Center Outcomes. We focused exclusively on the subset observations for cats (n=67,436).

The label classes were fairly imbalanced, with "Adoption" and "Transfer" outcomes occurring much more frequently than "Euthanasia", "Foster", "Return", or "No Outcome."

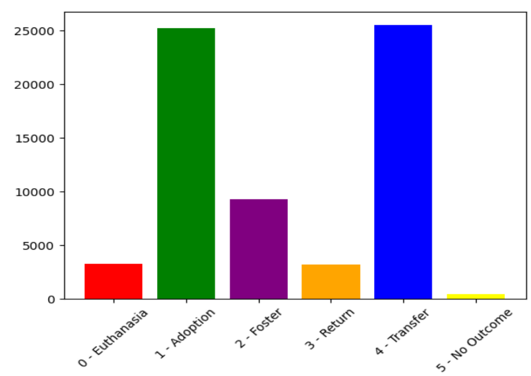


Figure 1: Distribution of Label Classes

2.2 Data Processing

The Austin Animal Center's Intake and Outcomes datasets are updated in real-time. We downloaded all available data from October 19, 2015 though November 25, 2023, and joined the two datasets on 'Animal ID'. For our labels, we grouped Outcome Type and Outcome Subtype combinations into 6 primary outcomes: Euthanized (Euthanasia, Died), Adopted, Fostered, Returned to Owner, Transferred, and Non-Outcome (animals without an outcome and/or still available for adoption).

For our features, we one-hot encoded 4 fields: Intake Type (Stray, Owner Surrender, Public Assist, Abandoned, Euthanasia Requested), Intake Condition (Normal, Aged, Behavior, Neonatal, Nursing, Pregnant, Med, Other), Sex upon Intake, and Breed.[1] Out of the 108 unique fields under Breed, we one-hot encoded the top 15, which accounted for 99% of observed cat breeds. Lastly, we standardized Age into months to create a continuous variable. In total, we had 33 features (see Table 1).

We used an 80/20 train/test split and stratified across the label classes. Given a sizable imbalance across the labels, we tried oversampling in the training data to improve the ability of the model to

[1]Although Color has been found to be highly predictive of animal shelter outcomes (Brown & Morgan, 2014; Carini et al., 2020), we excluded this field due to cleaning/processing complexity and time constraints.

predict the less common classes. However, this resulted in a worse performing model. Therefore, we present results without oversampling.

2.3 Multinomial Logistic Regression

Multinomial regression is a natural fit for a mutually exclusive multi-class prediction problem. The model is efficient, and the resulting coefficients are interpretable, unlike other multi-class machine learning techniques.

Given feature matrix X of shape (N, P), which consists of N observations and P features, each data point is represented by X_i, where $i = 1, 2...N$. Given Y label vector of shape $(N, 1)$, each observation's label is represented by Y_i. We create a model to predict the probability that an observation Xi belongs to class k, refining that model over a number of epochs. The multinomial regression formula can be represented as follows:

$z = f(k, i) = \beta_{0,k} + \beta_{1,k}x_{1,i} + \beta_{2,k}x_{2,i} + ...\beta_{P,k}x_{P,i}$, where K is the total number of classes such that $k = 1, 2, ...K$ and $\beta_{p,k}$ is the regression coefficient/weight associated with the p-th feature and k-th outcome. In our data, K = 6 and P = 33.

The softmax function is used to compute the probabilities: $\sigma(z) = P(Y = k|X) = \frac{e^z_k}{\sum_{j=1}^{K} e^z_j}$

Finally, the weight gradient used to update the weights in each epoch is: $\frac{1}{n}X^T(\sigma(z) - y)$

We repeat this process for a given number of epochs or until the weights converge, then select the class with the highest probability as the predicted label for each observation.

Pseudo code in Python:

```
X = feature matrix
y = labels

# randomly initialize weights (w) and bias (b)
# set t to learning rate, n to len(X), n_epochs to number of epochs, and epsilon as a small value close to 0
# set y to one-hot-encoded array of y

def softmax(z):
    exp_mat = np.exp(z - np.max(z)) / np.sum(
            np.exp(z - np.max(z)), axis=1)
        return exp_mat

for i in n_epochs:
    z = X.T @ weights + b
    w_gradient = 1/n * X.T @ (softmax(z) - y)
    b_gradient = 1/n * sum((softmax(z) - y)
    w_new = w - t * w_gradient
    b = b - t * b_gradient

    if np.abs(w- w_new) < epsilon:
        break
    else:
        w = w_new
```

Listing 1: Training the Model

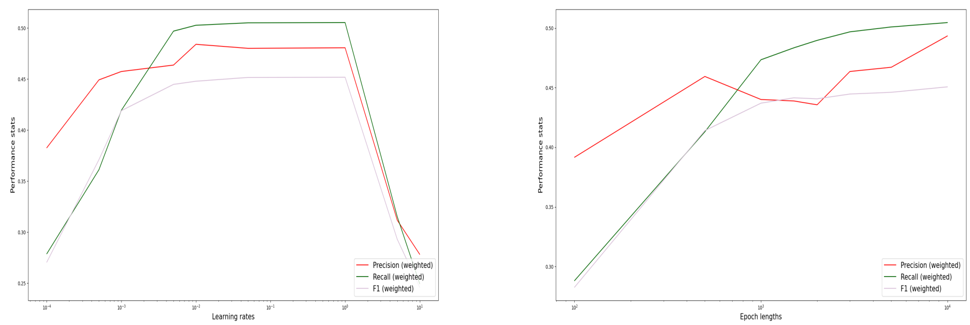
After training the model, we assessed the predictions using the training and test data. In both instances, we multiplied the output weights and bias from training by the X matrices to obtain raw y-hat values. The largest predicted class label determined the final predicted label, and we compared these predicted labels with the true labels to assess model performance.

```
# use previously learned weights (w) and bias (b) to predict outcomes using test data
z = X @ w + b
y_hat = softmax(z)
prediction = np.argmax(y_hat, axis=1)
```

Listing 2: Deploying the Model

3 Results

We tested models with different values for learning rates and the number of epochs on the training data. Model performance generally peaked with learning rates between .01 and 10, and performance showed signs of plateauing at 10,000 epochs.



(a) Different Learning Rates

(b) Different Numbers of Epochs

Figure 2: Model Performance with Different Hyper-parameters

Using the hyper-parameter values from the best performing models, we ran this model on the test data. The model's loss stabilized after about 1,000 epochs. This model had precision of 48.1%, recall of 50.5%, and F1 score of 45.2%. While these figures are lower than ideal, they demonstrate far higher predictive power than random guessing. We could likely increase model performance with more feature engineering (e.g adding Color), by reducing the number of potential predicted outcomes from 6 and grouping into fewer labels, testing alternative oversampling techniques, and by testing other machine learning methodologies.

The confusion matrix compares predicted labels against true labels. The figure highlights the label imbalance of the data, as the model primarily predicted labels 1 and 4, reflecting those classes' disproportionate frequency in the data.

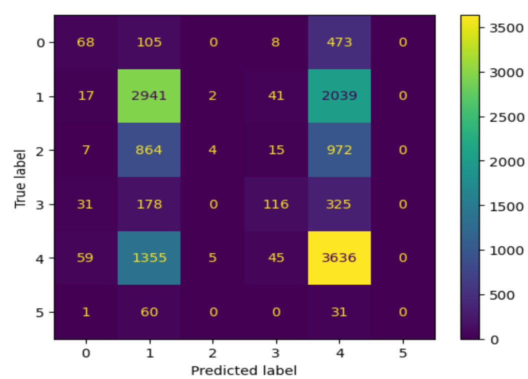


Figure 3: Confusion Matrix of True Labels vs. Predicted Labels

4 Conclusion

The multinomial logistic regression model improves prediction accuracy, even after minimal data processing and feature engineering. We show that such a model, built only in Python numpy, yields prediction accuracy of about 50% in a real-life dataset. Further development of this approach should focus on processing the data, refining an oversampling technique, and creating predictive features.

References

[1] Bradley, J. B., & Rajendran, S. (2021). Increasing adoption rates at animal shelters: a two-phase approach to predict length of stay and optimal shelter allocation. *BMC Veterinary Research*, **17**(1). https://doi.org/10.1186/s12917-020-02728-2.

[2] Brown, W. P., & Morgan, K. T. (2014). Age, breed designation, coat color, and coat pattern influenced the length of stay of cats at a No-Kill shelter. Journal of Applied Animal Welfare Science, 18(2), 169–180. https://doi.org/10.1080/10888705.2014.971156.

[3] Carini, R. M., Sinski, J., & Weber, J. D. (2020). Coat color and cat outcomes in an urban U.S. shelter. Animals, 10(10), 1720. https://doi.org/10.3390/ani10101720.

[4] Kilgour, R. J., & Flockhart, D. T. T. (2022). Direct and indirect factors influencing cat outcomes at an animal shelter. *Frontiers in Veterinary Science*, 9. https://doi.org/10.3389/fvets.2022.766312.

[5] Mitrović, K., Milošević, D., & Greconici, M. (2019, May 29-31). *Comparison of Machine Learning Algorithms for Shelter Animal Classification*. [Paper presentation]. IEEE 2019: 13th International Symposium on Applied Computational Intelligence and Informatics (SACI), Timisoara, Romania. https://doi.org/10.1109/saci46893.2019.9111575

[6] Zadeh, A. H., Comb, K., Burkey, B., Dop, J., Duffy, K., & Nosoudi, N. (2022). Pet analytics: Predicting adoption speed of pets from their online profiles. *Expert Systems With Applications*, *204*, 117596. https://doi.org/10.1016/j.eswa.2022.117596.

Table 1: Features

Intake Type
Stray Owner Surrender Public Assist Abandoned Euthanasia Requested
Sex upon Intake
Female Intact
Age upon Intake
Age in Months
Intake Condition
Cond Normal Cond Aged Cond Med Cond Behavior Cond Other Cond Feral Cond Neonatal Cond Nursing Cond Pregnant
Breed (Top 15)
American Shorthair Mix Domestic Longhair Domestic Longhair Mix Domestic Medium Hair Domestic Medium Hair Mix Domestic Shorthair Domestic Shorthair Mix Himalayan Mix Maine Coon Mix Manx Mix Other Ragdoll Mix Russian Blue Mix Siamese Siamese Mix Snowshoe Mix