COE 379L Project 01 Report — Austin Animal Center Outcomes

Author: Manas Pathak
Course: COE 379L — Software Design for Responsible Intelligent Systems
Date: Fall 2025

The Austin Animal Center dataset contains more than 131,000 records of animals and their outcomes, making it one of the largest publicly available shelter datasets. For this project, I focused on predicting whether an animal's outcome would be Adoption or Transfer. Because the raw data came directly from operational systems, it required careful cleaning and transformation before modeling could be effective.

To prepare the data, I first converted the columns containing dates (`Date of Birth`, `DateTime`, and `MonthYear`) into proper `datetime` objects. This made it possible to extract calendar features that could capture seasonal trends in outcomes. I then standardized the `Age upon Outcome` column, which originally stored ages as strings like "2 years" or "3 weeks." By tokenizing the strings, mapping units to days, and converting them into a numeric field called AgeDays, I created a consistent measure of animal age. Missing or ambiguous values were filled with the median to preserve comparability while reducing the impact of outliers.

Duplicate detection revealed 17 redundant records, which were removed to avoid double-counting outcomes. I also handled missing values: outcome-related fields such as `Outcome Subtype` were imputed with the mode, and animals with no name were labeled "Unknown." To create meaningful predictors, I engineered new features including `OutcomeYear`, `OutcomeMonth`, `BirthYear`, and `BirthMonth`. These allowed the model to account for potential seasonal adoption patterns. Finally, I dropped columns that were either redundant or high-cardinal, such as `Breed`, `Name`, `Date of Birth`, and `DateTime`. Categorical fields like `Animal Type` and `Sex upon Outcome` were cast to the `category` type and encoded with one-hot encoding. This process left me with a clean, structured dataset suitable for machine learning.

The preparation also provided useful insights into the shelter population. Adoptions were more common than Transfers, making up about 64% of outcomes, so I ensured stratification during the train–test split to preserve class proportions. Dogs made up the majority of animals, followed by cats, while other species were rare. Most animals were spayed or neutered, while intact animals formed a smaller group. Age was strongly right-skewed, with most animals under two years but some much older, justifying the use of median imputation and suggesting that age

effects would not be strictly linear. I also found that outcomes peaked at certain times, notably mid-2016 and again in 2024–25, which may reflect adoption drives or seasonal shifts in shelter operations. These findings confirmed that age, sterilization status, species, and seasonality were likely to be influential predictors.

For modeling, I split the dataset into an 80/20 train–test split, stratified by the outcome label and fixed with `random_state=42` to guarantee reproducibility. Each model was built in a scikit-learn pipeline with a `StandardScaler(with_mean=False)` so that numeric inputs were standardized while sparse one-hot encodings were preserved. I began with a baseline K-Nearest Neighbors (KNN) classifier with k=5. To refine this, I applied a 5-fold `GridSearchCV` on a small stratified sample of 2,000 rows, testing k values of 3, 5, and 11, and selected the best setting by F1 score on the Adoption class. Finally, I trained a Logistic Regression model using the `lbfgs` solver and 200 iterations. While the solver gave a convergence warning, the model still performed well, and increasing iterations would stabilize it fully.

The models showed progressively stronger performance. The baseline KNN achieved roughly 86% accuracy and an F1 score of 0.90, performing especially well on recall but slightly weaker on precision. Tuning KNN improved balance, raising accuracy to about 87% and F1 into the 0.90–0.91 range. Logistic Regression achieved the best results, with 88% accuracy, an F1 of 0.91, and the strongest recall at 0.96. Because recall on Adoption was prioritized, Logistic Regression was the clear choice. In practical terms, this model is especially good at identifying animals that are likely to be adopted, which directly supports the shelter's mission.

In evaluating the models, recall was emphasized over precision. A false negative—predicting Transfer when the true outcome is Adoption—risks underestimating demand for adoption services and misallocating resources. While precision still matters, recall aligns more directly with shelter priorities. For this reason, I focused on the F1 score as the key evaluation metric, giving weight to models that achieved strong recall without sacrificing overall balance. Logistic Regression provided the best trade-off, combining interpretability with excellent recall.

I am confident in the final model, though with some caveats. Strengths include the size of the dataset, the systematic preprocessing, and the use of stratified sampling. Logistic Regression is both interpretable and robust, making it a strong baseline for operational use. Limitations remain: one-hot encoding expands dimensionality, dropping `Breed` removed a potentially informative feature, and logistic regression assumes linear decision boundaries. In addition, model evaluation was based on a random split rather than a time-based split, which may not fully capture temporal drift. Future work could address these issues by exploring non-linear models like gradient boosting, tuning probability thresholds to balance adoption metrics, and retraining regularly as data evolves. Despite these considerations, the model's high recall (0.96) and strong F1 score (0.91) provide confidence that it can effectively predict Adoption versus Transfer outcomes and serve as a practical tool for supporting shelter operations.