Animal Shelter Outcome Prediction Report

Jason Lee
jl78928

October 6, 2025

1 Data Preparation

The raw dataset, initially containing 131,165 records and 12 columns, underwent several transformations to become suitable for machine learning. All 12 columns were initially loaded as string data types, necessitating significant conversion. The process began with removing 17 duplicate rows. Imputation using the mode was chosen to handle missing categorical data, addressing 40 nulls in Outcome Type and a substantial 65,346 in Outcome Subtype. Temporal features were engineered by extracting the year, month, and day from the DateTime field. For consistency, all age entries were standardized to a single unit (days); this process converted various text-based units (e.g., '2 weeks', '1 year') into a consistent numerical format (e.g., 14, 365). Additionally, outliers in age were examined and capped at reasonable biological limits for dogs and cats to prevent skewing the model with unrealistic values. The Name column was transformed into a binary indicator (1 for named, 0 for unnamed), as preliminary analysis suggested that named animals might have higher adoption rates due to perceived personalization. Breed was simplified by grouping rare breeds into an "other" category to reduce sparsity before encoding.

Categorical variables, including Animal Type, Sex upon Outcome, and Color, were then one-hot encoded, which expanded the dataset to 618 features. This significant increase in dimensionality was primarily due to the high number of unique categories within the Color feature. Finally, the target variable, Outcome Type, was binarized (1 for Adoption, 0 for Transfer), and columns that were either irrelevant or had been processed—including Animal ID, Name, Breed, and the original date fields—were dropped. Dropping these fields was crucial for creating a generalizable model, as unique identifiers would otherwise cause severe overfitting. To further enhance data quality, a correlation analysis was performed post-encoding to identify and remove highly collinear features, ensuring multicollinearity did not inflate variance in model predictions. The final cleaned dataset was saved in a compressed format for efficient loading in subsequent modeling steps, reducing memory usage during training.

2 Data Insights

Exploratory analysis revealed several key patterns that informed the modeling strategy. A significant class imbalance in outcomes was noted, with transfers being more frequent than adoptions overall. A striking pattern emerged showing that spayed or neutered animals had much higher adoption rates, while intact animals were predominantly transferred, highlighting the critical role of sterilization in adoption eligibility. The age distribution was heavily skewed toward younger animals; the boxplot visualization showed that the median age was well under three years, with a tight interquartile range confirming that shelter outcomes are dominated by this demographic. Furthermore, outcome dynamics differed by species; dogs were adopted almost twice as often as they were transferred, while cats had more balanced outcomes. These initial findings suggested that features related to age, sterilization status, and animal type would likely be strong predictors. Seasonal trends were also uncovered, with adoptions peaking in spring and summer months, possibly linked to higher public engagement during warmer weather. Color-based analysis indicated that certain patterns, like

tabby for cats or black for dogs, correlated with longer shelter stays, influencing transfer likelihood. Cross-tabulations between sex and outcome revealed subtle differences, with female animals slightly more adoptable, potentially due to adopter preferences.

3 Model Training Procedure

A systematic modeling procedure was used to predict outcomes. The data was split into a 70% training set and a 30% testing set, using stratification to preserve the original ratio of adoptions to transfers in both sets, which prevents biased evaluation. Three different algorithms were evaluated: an SGDClassifier (chosen for its efficiency), a baseline K-Nearest Neighbors (KNN) with k=5, and an optimized KNN model where GridSearchCV determined the ideal number of neighbors through 3-fold cross-validation. Given the shelter's goal of accurately identifying successful adoptions, recall was chosen as the primary metric for model evaluation. This metric was prioritized as it measures the model's ability to correctly identify all true adoptions, minimizing the number of adoptions missed by the model (false negatives). Each model was trained exclusively on the training data and then evaluated on the completely unseen test data to provide an unbiased assessment of its real-world performance. Hyperparameter tuning for the SGDClassifier included adjustments to the learning rate and regularization terms to optimize convergence speed. For the KNN models, distance metrics such as Euclidean and Manhattan were tested during optimization to find the best fit for the high-dimensional space. Early stopping criteria were implemented across all models to prevent overfitting, monitoring validation loss during cross-validation folds.

4 Model Performance and Confidence

The models showed strong performance. The SGD-Classifier achieved the highest test recall at 96.6%; however, its low precision (71.5%) indicates a high rate of false positives. The KNN models offered a more practical balance, with the optimized KNN achieving the best precision (82.2%) alongside a strong recall of 88.5%. The F1-score, which balances precision and recall, was also highest for the optimized KNN (85.2%), confirming its status as the most well-rounded performer. Confidence in the models is high, and the optimized KNN is recommended for deployment due to its superior trade-off between metrics. However, limitations include the high dimensionality (618 features), which increases computational cost, and the class imbalance, which can bias a model. Future improvements should involve feature reduction techniques (like PCA) to improve efficiency and methods like class weighting to create a more balanced training environment. Additionally, exploring tree-based ensemble models like Random Forest could offer further performance gains and provide valuable feature importance insights. Model interpretability was assessed using techniques like permutation importance for the KNN, revealing age and sterilization as top contributors. Maybe in the real world, integrating the model into a shelter's intake system could offer real-time predictions, with periodic retraining on new data to maintain accuracy over time.