Animal Shelter Outcome Prediction Report

Jason Lee
jl78928

October 6, 2025

1 Data Preparation

The initial dataset, comprising 131,165 records and 12 columns, underwent several transformations to prepare it for machine learning. All 12 columns were initially read in as string data types, which involved extensive conversion. Cleaning began by eliminating 17 duplicate rows. Mode imputation was chosen to handle missing categorical data, including 40 nulls in Outcome Type and a massive 65,346 in Outcome Subtype. Temporal attributes were generated by extracting year, month, and day from the DateTime column. For consistency, all age entries were normalized to a single unit (days); this process converted many text-based units (e.g., '2 weeks', '1 year') to a standard numerical representation (e.g., 14, 365). Age outliers were also examined and fixed to reasonable biological limits for dogs and cats to prevent skewing the model with unrealistic values. The Name column was translated into a binary indicator (1 for named, 0 for unnamed), as preliminary analysis suggested that named animals might have increased adoption rates because of perceived personalization. Breed was reduced by consolidating 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 increased the dataset to 618 features. This significant increase in dimensionality was largely due to the high number of unique categories in the Color feature. Finally, the target variable, Outcome Type, was binarized (1 for Adoption, 0 for Transfer), and columns irrelevant or transformed—including Animal ID, Name, Breed, and the original date columns—were dropped. It was necessary to remove those fields in order to ensure the model was generalizable, as unique identifiers would otherwise lead to severe overfitting. For further data quality enhancement, a correlation analysis was performed after encoding to identify and remove highly collinear features so that multicollinearity would not artificially inflate variance in model predictions. The final cleaned data were saved in a compressed format for quick loading in subsequent modeling steps with less memory usage during training.

2 Data Insights

Exploratory analysis revealed several important patterns that informed the modeling strategy. One important pattern was class imbalance in outcomes, with more transfers compared to adoptions overall. A striking pattern was that spayed or neutered animals were adopted at far greater rates, while intact animals were nearly all transferred, highlighting the critical role of sterilization for adoptability. The age distribution was highly biased towards younger animals; the boxplot figure showed the median age was well under three years, with a compact interquartile range showing that shelter outcomes are dominated by this age group. Furthermore, outcome dynamics were different by species; dogs were adopted almost twice as often as they were transferred, while cats had more balanced outcomes. These initial results showed that attributes connected to age, sterilization status, and animal type would be effective predictors. Seasonal trends were also uncovered, with spring and summer months having peak adoptions, possibly linked to higher public engagement during warmer months. Color analysis indicated that certain patterns, like tabby for cats or black for dogs, were linked to longer shelter stay, which impacted transfer potential. Sex

and outcome cross-tabulations indicated refinement differences, with female animals being more adoptable, possibly a reflection of adopter preference.

3 Model Training Procedure

A formal modeling procedure was used for prediction. The data were split into a 70% training set and 30% test set, using stratification to preserve the original balance of adoptions to transfers in both sets to prevent biased evaluation. Three algorithms were tried: an SGDClassifier (chosen for its efficiency), a basic K-Nearest Neighbors (KNN) with k=5, and an optimized KNN model where GridSearchCV showed the ideal number of neighbors through 3-fold cross-validation. Given that the shelter's goal is to accurately classify successful adoptions, recall was used as the primary metric for evaluating the models. This was undertaken because it measures the model's ability to identify all existing adoptions, cutting down on the number of undetected adoptions by the model (false negatives). All models were trained solely on the training set and later tested on the com- whole unseen test data to provide an unbiased assessment of its true performance in the real world. Hyperparameter optimization for the SGDClassifier comprised learning rate and regularization parameter tuning to achieve maximum convergence rate. In KNN models, Euclidean and Manhattan distance metrics were tried out during the optimization phase to find the best fit in the high-dimensional space. Early stopping criteria were imposed in all the models to prevent overfitting, monitoring validation loss across cross-validation folds

4 Model Performance and Confidence

The models all had good performance. The highest test recall was seen in SGD-Classifier at 96.6%; however, its low precision (71.5%) indicates high false positives. The KNN models gave a more realistic balance, and the highest precision (82.2%) with a good recall of 88.5% was seen in optimized KNN. The F1-score, which gives a harmonic balance between precision and recall, also was best for optimized KNN (85.2%), which further reinforces its standing as the most balanced performer. There is strong confidence in the models, and deployment is suggested for the optimized KNN as it provides an improved trade-off between the metrics. Obstacles are the high-dimensional data (618 features), which incurs greater computational expense, and the class imbalance, which may bias a model. Future developments should be done through feature reduction techniques (like PCA) to improve efficiency and methods like class weighting to provide a balanced training environment. In addition, exploration of tree-based ensemble models like Random Forest could give further improvement in performance and give good feature importance. Model interpretability was established by techniques like permutation importance for the KNN, with age and sterilization being the most important feature contributors. Maybe in reality, incorporating the model into an intake system at a shelter could give real-time output, with periodic retraining with new data to maintain accuracy over time.