

Project 01 Model Report

Predicting Animal Outcomes

1. Project Overview

The primary objective of this project was to analyze a dataset about sheltered animals in Austin. The dataset had information on features such as date of birth, age, breed, color, and animal type for animals at a shelter. These features were used to develop and evaluate different classification models to assess their ability to predict the Outcome Type of animals at a shelter, categorized as either Transfer or Adoption.

Data Preparation

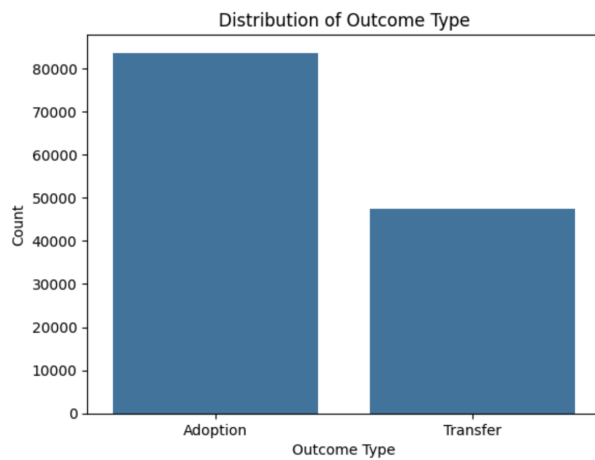
The data preparation phase began with an examination of the initial composition. The dataset had an initial shape of (131165, 12). Many columns contained missing or NaN values. The Name column had 37507 nulls, Outcome Type had 40, and Outcome Subtype had 65355. I replaced the null values with appropriate information and then identified and removed all duplicate rows.

After data cleaning, I completed univariate analysis to visualize the data columns. This analysis showed a majority of adoptions in the data instead of transfers. There were also a majority of dogs in the Animal Type category, and very few birds. The Sex upon Outcome distribution showed a large majority of neutered males or spayed females.

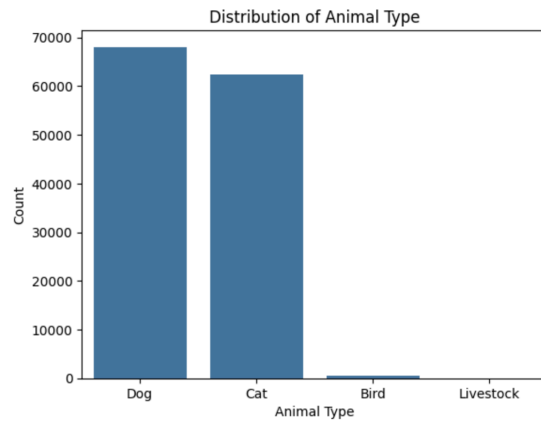
Next, feature transformation was performed on the Age upon Outcome column. This feature was originally a categorical variable with a range of possibilities from "2 years" or "2 weeks", which was then converted into a single numerical feature. The Age in Days column transformed all of these descriptions to a single number which represented the age in days.

Another important consideration was deciding what columns would need to be part of the final set. Columns with irrelevant information or high cardinality, meaning that they had too many unique categories, were removed. The columns dropped were: Animal ID, DateTime, Date of Birth, MonthYear, Name, Breed, and Color.

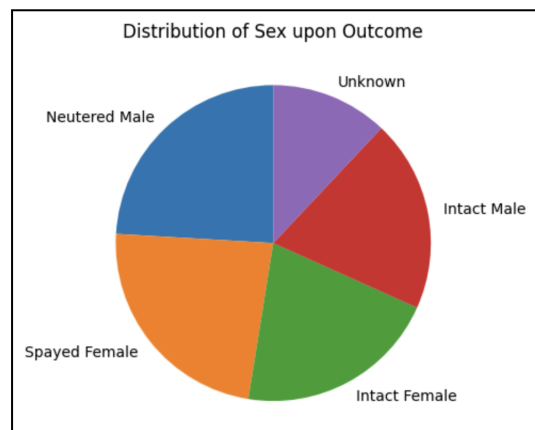
That left us with Age in Days, Outcome Type, Outcome Subtype, Animal Type, and Sex Upon Outcome. These were One-Hot Encoded to transfer all categorical variables into boolean variables.



Distribution of animals that were adopted vs transfers.



Distribution of animal type. There was a majority of dogs.



*Distribution of Sex upon Outcome.
Large majority of neutered male or spayed females.*

Insights from Data Preparation

A key insight from my initial data preparation is the importance of initial checks. For example, failing to drop duplicate rows caused errors and led to an unrealistic accuracy across all three classification models during my first runs. This demonstrated the necessity of carefully selecting which columns to drop, as redundant or high-cardinality features (like 'Breed') can negatively affect model performance, though dropping too many features does not allow for comprehensive inferences.

2. Model Training and Evaluation

Training Procedure

Splitting the Data

The dataset was split into 70% training data and 30% testing data. This was achieved using `train_test_split` from `scikit-learn` library. Reproducibility was ensured using a `random_state=42`, and the proportions of the classes selected were maintained by defining `stratify=y`, which reduced bias in the evaluation.

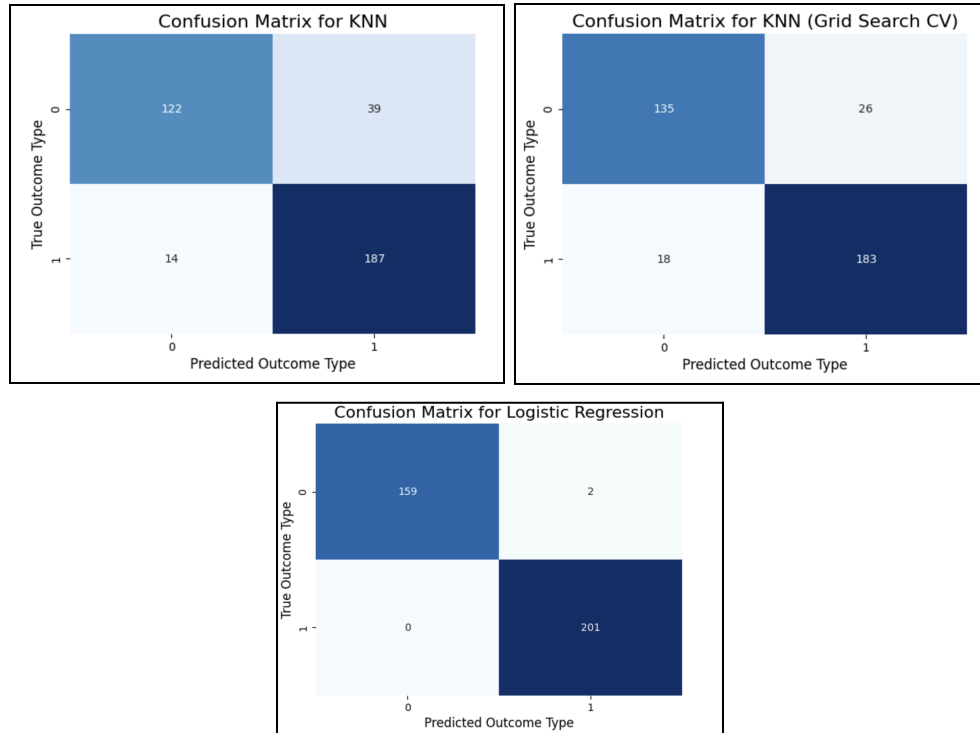
Model Selection and Training

I trained three classification models using the cleaned dataset.

- K-Nearest Neighbor (Basic):** I set `k` to the default value of 5.
- K-Nearest Neighbor (Grid Search CV):** Used grid search to find the best value of `k`, and then trained a kNN model with that value. The model concluded that the best value of `k` was 1.
- Logistic Regression (Scikit-learn):** I used logistic regression as the linear model.

2.2. Model Performance (for True Predictions)

Model	Accuracy	Precision	Recall	F1-Score
KNN (Basic)	0.85	0.83	0.93	0.88
KNN (GridSearch)	0.88	0.88	0.91	0.89
Logistic Regression	0.99	0.99	1.00	1.00



2.3. Class Prediction

How does the model perform to predict the class?

The models performed well in predicting the class. Particularly, the Logistic Regression model achieved a very high accuracy score of 0.99. GridSearch successfully tuned the K-Nearest Neighbor model, increasing its accuracy from the initial 0.85 with the Basic KNN to an optimized 0.88. The models successfully utilized the cleaned and encoded features (Outcome Subtype, Animal Type, and Sex upon Outcome) to predict the outcome.

Model Confidence: I am confident in the model because the accuracy scores are high. Since we cleaned the data, the model was able to extract what features could be used to predict the outcomes using the features Subtype, Animal Type, and Sex upon Outcome. We are able to see that logistic regression had the highest accuracy score.

Which metric is most important for this problem?

The most important metric for this problem is accuracy. The dataset is only moderately imbalanced, so we do not need to consider the F1-score so heavily. Overall, accuracy is the best measure for us to clearly understand the predictive performance of the model. This directly reflects how well the model classifies Adoption and Transfer outcomes. Accuracy is more relevant than Precision and Recall because it could bias the model from a balanced performance. Accuracy allows us to see the fairest assessment of the classifier's performance across the whole dataset.