Ayushi Kate
COE379L
October 8 2025

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

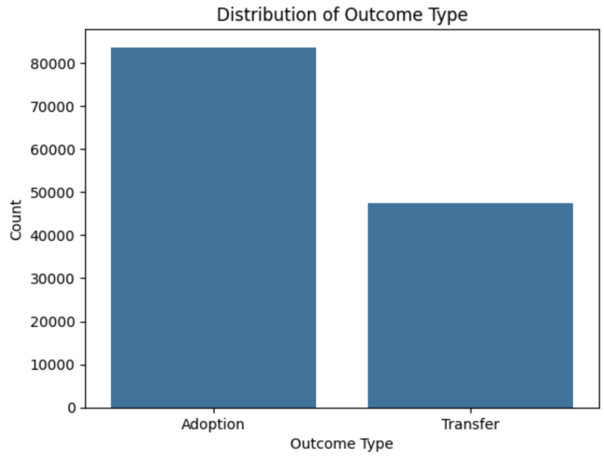
The data preparation phase began with an examination of the initial composition. The dataset had an initial shape of (131165, 12). Many columns contained missing or NaN values. The Name column had 37507 nulls, Outcome Type had 40, and Outcome Subtype had 65355. I replaced the null values with appropriate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

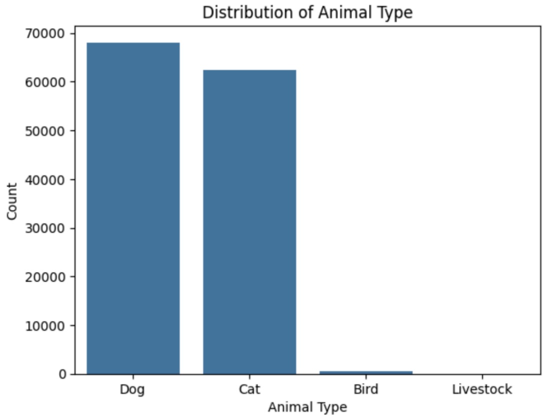
Next, feature transformation was performed on the Age upon Outcome column. This feature was originally a categorical variable with a range of possibilities from "2 years" or "2 weeks", which was then converted into a single numerical feature. The Age in Days column was the transformation of all the descriptions 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 values, were removed. The columns dropped were: Animal ID, DateTime, Date of Birth, Outcome Subtype, Name, Breed, and Color.

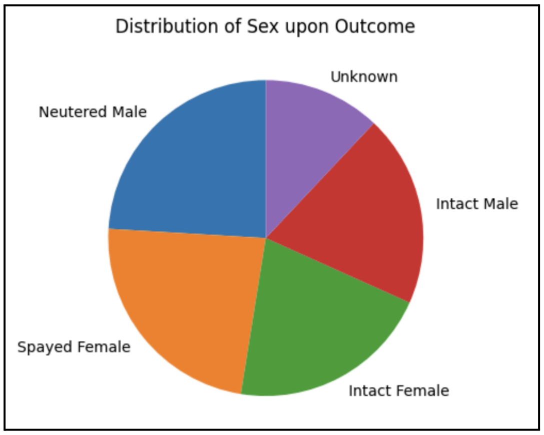
That left us with Age in Days, MonthYear, Outcome Type, Animal Type, and Sex Upon Outcome. These were One-Hot Encoded to transfer all categorical variables into boolean variables.



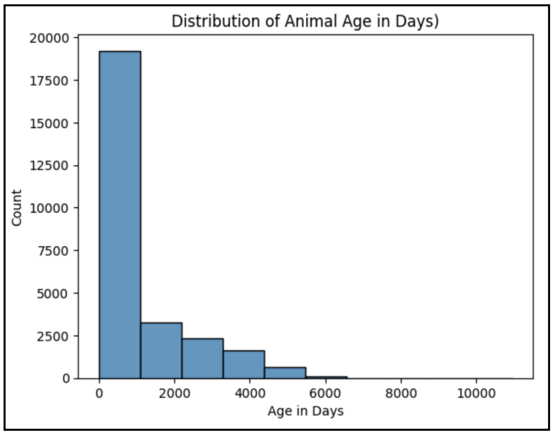
Distribution of animals that were adopted vs transfers.



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



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



Most of the animals were between 0 to 1000 days old.

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

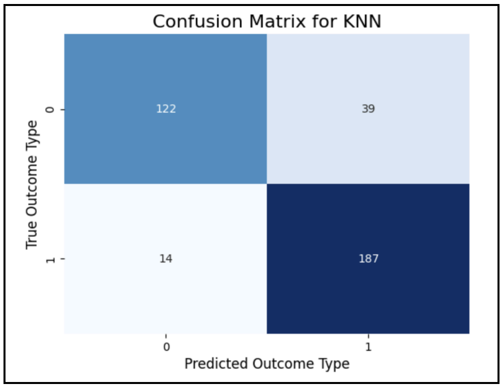
Splitting the Data
The dataset was split into 70% training data and 30% testing data. This was achieved using train_test_split from scikit-learn library. Reproducibility was ensured using a random_state=42, and the proportions of the classes selected were maintained by defining stratify=y, which reduced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

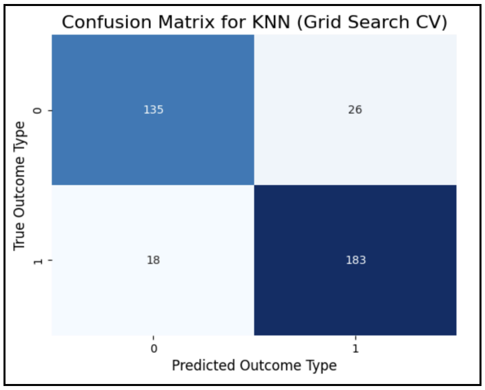
Model Selection and Training
I trained three classification models using the cleaned dataset.

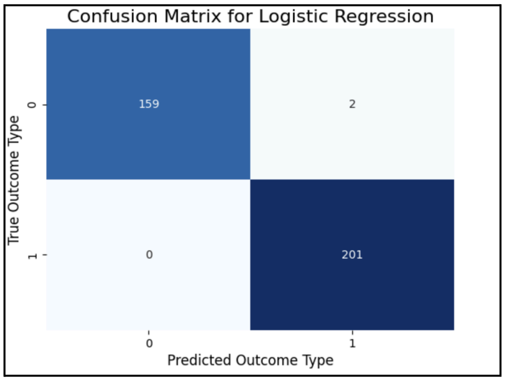
1. **K-Nearest Neighbor (Basic):** I set k to the default value of 5.
2. **K-Nearest Neighbor (Grid Search CV):** Used grid search to find the best value of k, and then trained a kNN model with that value. The model concluded that the best value of k was 1.
3. **Logistic Regression (Scikit-learn):**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

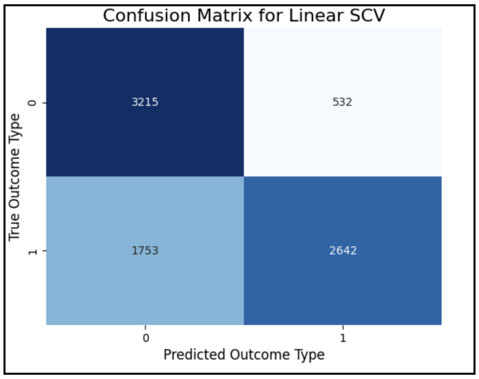
2.2. Model Performance (for True Predictions)

Model	Accuracy	Precision	Recall	F1-Score
KNN (Basic)	0.60	0.62	0.64	0.63
KNN (GridSearch)	0.69	0.73	0.68	0.70
Logistic Regression	0.72	0.81	0.62	0.70
Linear SVC	0.72	0.83	0.60	0.70









2.3. Class Prediction

How does the model perform to predict the class?

The models performed well in predicting the class. They achieved an accuracy range of 0.60 to 0.72. Particularly, the Logistic Regression model and Linear SVC achieved the highest accuracy score of 0.72. GridSearch successfully tuned the K-Nearest Neighbor model, increasing its accuracy from the initial 0.60 with the Basic KNN to an optimized 0.69. The models successfully utilized the cleaned and encoded features (Age Upon Outcome, Animal Type, MonthYear, and Sex upon Outcome) to predict the outcome. The models had overall high numbers in precision but lower scores for recall.

Model Confidence: I am moderately confident in the model. The data cleaning, removing leakage, and balance in the data selected ensures that the statistics are fairly reliable. However, the mode's predictive power is limited because there are so few categories to consider. I removed the 'Breed' column to simplify analysis, but it is possible that it will limit the model development. It is also pretty important to note that the dataset only has information about Austin shelter animals, so it cannot be generalizable. Yet, the accuracy scores were pretty high overall, so I have moderate confidence in the model.

Which metric is most important for this problem?

The most important metric for this problem is accuracy. The dataset is only moderately imbalanced, meaning the classes (Adoption vs. Transfer) occurs substantially. Thus, we do not need to consider the F1-score so heavily. Our primary goal for this problem is to maximize the overall correctness, and accuracy is the best measure to understand the aggregate predictive performance of the model.

There is no need to prioritize precision or recall in this scenario, and over-optimizing could actually significantly bias the model from a balanced performance. Therefore, accuracy is the fairest assessment of the classifier's performance across the whole dataset.