COE 379L – Project 1 Report

Keshav Bhargava

Data Preparation

The dataset used in this project was about sheltered animals in Austin. The dataset ranged in categories from species, breed, and age to adoption outcome. The goal of the project was to predict the outcome type, which in this case, whether the animal was transferred to another organization or adopted by someone.

I began looking at the dataset, checking the shape, size, and any missing values. The dataset contained over 130,000 entries with 12 different variables. Here are the techniques I did to clean up the data:

1. Duplicate Rows:
   a. Dropped 17 duplicate rows that were found
2. Missing Values:
   a. Converted all NaN values in ‘Name’ to “Unknown”
   b. Filled the Outcome Type and Outcome Subtype using their mode
3. Conversion of Types:
   a. For ‘Age Upon Outcome’ I converted it into a consistent numeric value represented in days which ensured that the model could interpret age as a continuous variable.
   b. Converted DateTime, Month Year, and DateofBirth to datetime variables
   c. Converted ‘Outcome Type', 'Outcome Subtype', 'Animal Type', 'Sex upon Outcome', 'Breed', 'Color’ to Categorical variables
4. Irrelevant Columns:
   a. Dropped Animal ID and name as there can be several different types of IDs and names. Typically this won't help you determine if the animal will be adopted or transferred. I also dropped the breed column later on.
5. Hot Encoding:
   a. Since we are using hot encoding for the next step for machine learning I determined what categorical variables are the most important. Outcome type Animal type, Sex upon outcome seemed to be the most important categorical features that impacted the result.

Insights for Data Preparation

The exploratory data analysis showcased several interesting features. Age seemed to have a strong relationship with the outcome of the animal.Typically younger animals were more likely to be adopted than transferred, while older animals tended to be transferred to other shelters. Animal type also played a significant role. Cats tended to be more likely to transfer while dogs had a higher adoption rate. Sex and reproductive systems had a moderate effect. Neutered or spayed animals were more likely to be adopted than intact ones. Adoptions were usually seasonal with most during the summer time. As for color, most animals were black or white.

Model Training Procedure

The target variable in this prediction model was the outcome type. First I used train_test_split to split with 80% train and 20% test data. I also stratified y to preserve class ratios. For the actual features that I included in the training data, I included all the hot encoded variables as well as age upon outcome. For the models I trained I did the following:

1. K-Nearest Neighbor Classifier (KNN) – I started with a baseline KNN model using k=5. The model was trained on the standardized feature set.
2. K-Nearest Neighbor with Grid Search CV – I performed hyperparameter tuning using GridSearchCV, exploring values of n_neighbors from 3 to 11 and weight options (uniform and distance).
3. Linear Classification – I implemented a Perceptron model to serve as a linear baseline. Perceptron trains efficiently on large datasets and provides interpretable coefficients.

Model Performance

Model	**Accuracy**	**Precision**	**Recall**	**F1**
KNN (n=5)	**0.8437**	**0.8432**	**0.8437**	**0.8404**
KNN (Grid Search)	**0.8504**	**0.8519**	**0.8504**	**0.8462**
Perceptron	**0.7877**	**0.7937**	**0.7877**	**0.7743**

Observations:

1. The baseline KNN model performed well with an accuracy of 84%.
2. The KNN grid search improved the performance with an accuracy of 85%, suggesting that neighborhood grid search does improve a model
3. The Perceptron model’s performance improved significantly after scaling and tuning, achieving an accuracy of 78.8%, which, while slightly lower than KNN, demonstrates solid linear classification capability on the dataset.

Model Confidence

I am moderately confident in the model’s prediction. The accuracy is solid for a real-world classification task. But the model is fairly simple and most likely doesn’t capture all the nuances that come with animal outcomes. F1 might be the most important metric, as it ensures that we minimize false negatives and positives, which could be critical depending on the context of the 'Outcome Type' predictions.