COE 379L Project 01 Report -- Austin Animal Center Outcomes

1. Data Preparation Workflow

I began by loading the raw Austin Animal Center dataset (`project1.csv`) into pandas and auditing its s

Duplicate detection revealed several repeated entries (roughly 150 rows). I removed duplicates while p

`Age upon Outcome` required special handling because values were stored as strings like `"2 years"` (

After parsing the timestamp fields, I engineered calendar-based features (`OutcomeYear`, `OutcomeM

2. Exploratory Data Analysis Insights

- **Outcome balance.** Adoption and Transfer outcomes are both common. Transfers have a slight ed
- **Species mix.** Dogs dominate the records, followed by cats. Birds and other species appear rarely
- **Sterilization status.** Most outcomes involve spayed or neutered animals. Intact animals are a mino
- **Age distribution.** The `AgeDays` histogram is sharply right-skewed: the bulk of animals are young
- **Temporal trends.** Aggregating outcomes by month highlights peaks in mid-2016 and a resurgence

3. Modeling Procedure

- **Train/test split.** Using the engineered dataset, I separated features and target (`Outcome Type`) a
- **Pipelines.** Each model used a `Pipeline` with `StandardScaler(with_mean=False)` so the one-hot
- **Baseline KNN.** Fit a K-Nearest Neighbors classifier with `k=5` as a simple baseline.
- **KNN with tuning.** To satisfy the grid-search requirement without incurring excessive runtime, I dre
- **Linear model.** Trained a logistic regression classifier (`solver='lbfgs'`, `max_iter=500`) on the scal

4. Model Performance Summary

Model	Accuracy	Precision (Adoption)	Recall (Adoption)	F1 (Adoption)
KNN (k=5)	0.86	0.87	0.93	0.90
KNN (GridSearch)	0.87	0.88	0.91	0.89
Logistic Regression	0.88	0.87	0.96	0.91

Notes: Metrics are computed on the 26,230-record test set with Adoption treated as the positive clas

5. Priority Metric Justification

For an animal shelter, recall on the Adoption class is particularly important. A false negative (predicting

6. Confidence and Limitations

- **Strengths.** The dataset is large and stratified; multiple algorithms were evaluated; cross-validation
- **Limitations.** One-hot encoding creates a high-dimensional feature space, which can be sensitive t
- **Future work.** Incorporating more nuanced feature interactions, trying calibrated probability thresho

Given the strong F1 score from logistic regression and the alignment between its high recall (0.96) and