Solutions

2.1

- Using it for model choice: it becomes a second validation set, and its score is then optimistic ✓.

2.2

- Predicting every parcel on time: $158/200 = 0.79$ ✓.

2.3

- Scaling before splitting; imputing with test statistics; feature selection over the full data (and the label-encoding feature itself) ✓.

3.1

- Competitor: accuracy $(40 + 138)/200 = 0.89$ ✓.
- Precision $40/60 = 0.667$; recall $40/42 = 0.952$; F1 $= 2\times0.667\times0.952/(0.667+0.952) = 0.784$ ✓.
- Comparison: competitor wins recall (0.952 vs 0.857), loses precision (0.667 vs 0.783) and accuracy (0.89 vs 0.92) ✓.
- For catching lates (refunds depend on not missing them): the competitor – recall is the operative metric ✓.
- Demand: 5-fold CV with per-fold counts, and the same folds for both models ✓.

3.2

- Wrong order: fit scaler, imputer, and feature selector on ALL data before splitting – the test set's mean/variance, its missing-value distribution, and its label-feature correlations all leak into training ✓.
- Correct per-fold: split first; fit scaler, imputer, selector on the training fold only; transform the validation fold with those fitted objects ✓.
- The status code: a feature whose provenance is post-outcome – the data card's provenance column catches it before any split ✓.
- Unequal folds: fold accuracies weight small folds equally – pool TP/FP/FN/TN across folds and compute metrics on the pooled counts (or weight by fold size) ✓.