Data splitting

Group-aware cross-validation splitting

Purpose: Group-aware cross-validation is designed to evaluate model performance in datasets where observations are not independent, because multiple observations belong to the same economic entity (e.g. firms observed over several years). Its goal is to prevent information leakage and to obtain stable and realistic out-of-sample performance estimates.

While a single group-aware split produces one train–test partition. Group-aware cross-validation repeats this process multiple times, holding out different sets of firms, and averages performance across these splits.

1. Group by ID. Each firm is treated as a single group, identified by id. All observations associated with the same firm (multiple years of balance-sheet data) are considered inseparable for validation purposes.

2. Construct a Firm-Level Representation

The panel dataset is collapsed to the firm level to determine how firms are assigned to folds.

- The default indicator is aggregated using an ever-default rule:
- A firm is labeled as defaulting if it defaults in any observation.

Time-invariant firm characteristics (sector, size, group membership) are retained.

3. Define Stratification Variables

To preserve structural heterogeneity across folds, firm-level stratification variables may be defined (e.g. Sector $\times$ Default).

These variables are used to guide fold assignment.

Stratification is applied at the firm level, not at the observation level.

4. Define Assign Firms to Folds

- Firms are randomly partitioned into K mutually exclusive folds. Each firm is assigned to exactly one fold.
- All observations belonging to that firm inherit the same fold assignment.
- When stratification is used, the algorithm attempts to preserve the distribution of stratification variables across folds.

5. Perform cross validation

For each fold $k = 1 \ldots K$

- Training set: all observations belonging to firms not in fold k
- Validation set: all observations belonging to firms in fold k

Clarificatiokn

Should the data on the distribution of firms across categories (and the defaults within the categories) be explicitly used in the data-splitting procedure?

Such information is primarily used as a diagnostic check rather than as a strict splitting constraint. After applying group-aware splitting and multivariate stratification on key variables (e.g. sector and default status), category distributions and default rates can be compared across samples to verify that no major distortions occur. Explicitly forcing category-specific default rates is generally unnecessary and may bias the evaluation sample.