Vienna University of Economics and Business

OeNB ILAB

OeNB ILAB:
Data-Splitting

Author:
Tristan Leiter

Submission Date:
December 6, 2025

Chapter 1

Stratified Sampling

1.1 Issues with univariate stratified sampling

1.1.1 Heterogeneity in Sector Default Rates

- **Issue:** Default risk is not uniform across the economy; it is structurally different between industries. Univariate sampling ignores these differences, treating a "Retail" default the same as an "Energy" default during the split.
- **Literature Insight:** Corporate diversification strategies lead to "divergent profitability, capital allocation, and risk profiles across business units" (Wei, 2025). Consequently, financial institutions must prioritize "cross-sectoral risk monitoring" (Wei, 2025). Furthermore, empirical evidence shows that rating heterogeneity is driven by market-specific factors (such as transition vs. non-transition economies), implying that risk models must account for these sub-market characteristics to be accurate (Hornik et al., 2010).
- **Advantage of Multi-Variate:** It forces the Training set to learn the specific risk drivers of *every* sector, rather than allowing the model to be dominated by the majority sector.

1.1.2 Out-of-Sample (OOS) "Blind Spots"

- **Issue:** In datasets with extreme class imbalance (e.g., 1% defaults), univariate sampling can randomly assign *all* defaults from a small sector (e.g., Energy) to the Training set. This leaves **zero defaults** in the Test set for that sector.
- **Literature Insight:** Standard sampling practices often recommend small samples that fail to capture the granularity of rare classes (Crone and Finlay, 2012). However, multi-way stratification allows for "increased precision of estimates of each of the variables" simultaneously (Winkler, 2001). By controlling the allocation, it ensures that the sample is distributed across cells defined by multiple criteria (Sector and Target), preventing empty cells in the validation set.
- **Advantage of Multi-Variate:** It guarantees that the Test set contains at least one default for every sector, making it possible to calculate performance metrics (Recall, AUC) for every industry segment individually.

1.1.3 Out-of-Time (OOT) Stability & Population Drift

- **Issue:** Credit risk models are subject to "population drift" where the relationship between features and default changes over time (e.g., an oil price shock specifically affects the Energy sector in Year 3).
- **Literature Insight:** Applicant populations and distributions evolve over time due to changes in the economic environment (Crone and Finlay, 2012). If a model is trained on a univariate sample that accidentally over-represents a specific sector during a benign economic period, it will fail to generalize when that sector's specific risk profile shifts in the future (OOT).
- **Advantage of Multi-Variate:** By ensuring that the sector distribution in the Training/Test split strictly mirrors the population structure (or is balanced via weights), the model is less likely to learn spurious correlations driven by a single sector's temporary performance, resulting in more robust OOT predictions.

1.2 Multivariate stratified sampling

1. **Create a Stratification Key**
   Create a new temporary column in the dataset that combines the two variables of interest: `Sector` and `Target` (Default Status).
   - *Example:* If a row has Sector = "Energy" and Target = 1, the key becomes `"Energy_1"`.
   - *Example:* If a row has Sector = "Retail" and Target = 0, the key becomes `"Retail_0"`.
2. **Filter "Singleton" Groups**
   Check the counts of each unique key. If any group has only 1 observation (e.g., only one "Energy" company that defaulted), it cannot be split between Train and Test. One must either:
   - Exclude these rows from the split (and force them into Training later); or
   - Accept that the stratifier will fail for these specific rows.
3. **Perform the Stratified Split**
   Run a standard stratified sampling algorithm, but set the *stratification target* to be the new Key column (instead of just the `Target`). This forces the algorithm to pick:
   - 80% of `"Energy_0"` and 20% of `"Energy_0"`
   - 80% of `"Energy_1"` and 20% of `"Energy_1"`
   - ...and so on for every sector.
4. **Clean Up**
   Remove the temporary key column from both the Training and Test sets to return the data to its original structure.

Chapter 2

Other sampling techniques

2.1 SMOTE + ENN

2.1.1 The Procedure

To address the challenges of credit risk prediction in extremely imbalanced datasets, Zhao et al. (2024) propose a novel framework called **SH-SENN** (Strategic Hybrid SMOTE with Double Edited Nearest Neighbors). This method departs from traditional sampling by treating the sampling strategy as a hyperparameter rather than a fixed preprocessing step. The procedure consists of the following steps:

1. **Determine Strategy Ratio (α)**
   The framework defines a strategy ratio $\alpha = N_{nmin}/N_{maj}$, representing the proportion of minority class samples to majority class samples after oversampling. This ratio typically ranges from 0.1 to 0.9 and is treated as a hyperparameter subject to cross-validation (Zhao et al., 2024).

2. **Strategic SMOTE Oversampling**
   Minority class samples are oversampled using SMOTE to reach the target proportion defined by α (e.g., 10% to 90% of the majority class size).

3. **First ENN (Undersampling)**
   A search for neighboring samples is conducted using the Edited Nearest Neighbors (ENN) rule, and misclassified samples are removed to clean the dataset.

4. **Second ENN (Double Cleaning)**
   Crucially, the framework applies ENN a *second time*. The dataset is re-evaluated, treating the resampled data as new data. The ENN algorithm is adjusted to identify and remove boundary noise and overlapping samples that may have been introduced or left behind by the initial SMOTE+ENN process (Zhao et al., 2024).

5. **Validation and Selection**
   The effectiveness of different strategy ratios (e.g., SH-SENN 0.5 vs. SH-SENN 0.9) is evaluated on a validation set to select the optimal strategy before application to the test set.

2.1.2 Solving Prior Issues

The SH-SENN framework specifically addresses limitations found in traditional resampling techniques when applied to extreme imbalance (e.g., 1% default rates):

- **Noise Reduction in Extreme Imbalance:** Standard SMOTE can blindly generate synthetic samples in noisy areas, obscuring the decision boundary. SH-SENN's primary advantage is the "Double ENN" mechanism, which further reduces the introduction of new noise and interference items that standard SENN might create. This effectively clarifies the decision boundary (Zhao et al., 2024).

- **Adaptability to Dataset Characteristics:** Existing methods often apply a "one-size-fits-all" approach to balancing. SH-SENN is designed as an extensible framework where the sampling intensity (α) is adjusted based on the dataset's specific Imbalance Ratio (IR). For instance, extreme imbalance might require a strategy ratio of 0.5, while normal imbalance might benefit from 0.9 (Zhao et al., 2024).

- **Enhancement of Ensemble Classifiers:** While modern classifiers like CatBoost handle categorical features well, they still struggle with extreme class imbalance on their own. SH-SENN significantly enhances the predictive capabilities of these ensemble classifiers compared to individual classifiers, providing robust improvements in metrics like AUC and G-mean (Zhao et al., 2024).

Bibliography

Sven F. Crone and Steven Finlay. Instance sampling in credit scoring: An empirical study of sample size and balancing. *International Journal of Forecasting*, 28(1):224–238, 2012.

Kurt Hornik, Rainer Jankowitsch, Manuel Lingo, Stefan Pichler, and Gerhard Winkler. Determinants of heterogeneity in european credit ratings. *Financial Markets and Portfolio Management*, 24(3):271–287, 2010.

Han Wei. Smote algorithm optimization and application in corporate credit risk prediction with diversification strategy consideration. *Scientific Reports*, 15(1):23598, 2025.

William E. Winkler. Multi-way survey stratification and sampling. Research Report Series Statistics #2001-01, Statistical Research Division, U.S. Bureau of the Census, 2001.

Zixue Zhao, Tianxiang Cui, Shusheng Ding, Jiawei Li, and Anthony Graham Bellotti. Resampling techniques study on class imbalance problem in credit risk prediction. *Mathematics*, 12(5):701, 2024. doi: 10.3390/math12050701.