OeNB Industry Lab - Documentation

Tristan Leiter

November 7, 2025

Contents

1 Einleitung **2**
1.1 Einleitung . . . 2

2 Exploratory Data Analysis **3**
2.1 Overview . . . 3
2.1.1 Description of the dataset . . . 3
2.1.2 Description of the dataset . . . 3
2.2 Data Splitting with Imbalanced Data . . . 3
2.2.1 Class Imbalance . . . 3
2.2.2 The Solution: Stratified Sampling . . . 3

3 Generalized linear models (GLM) **4**
3.1 Overview . . . 4

A Overview **5**

Chapter 1

Einleitung

1.1 Einleitung

Chapter 2

Exploratory Data Analysis

2.1 Overview

2.1.1 Description of the dataset

2.1.2 Description of the dataset

2.2 Data Splitting with Imbalanced Data

2.2.1 Class Imbalance

In this credit risk dataset, the target variable, y, is inherently imbalanced. The number of non-defaults (0) significantly outnumbers the number of defaults (1). This is a common and expected characteristic of credit risk data.

This imbalance poses a significant challenge for model development. If we were to use a simple random split to create our training, validation, and test sets, we would face a high risk of creating unrepresentative samples. For example, a small test set could, purely by chance, end up with a much higher or lower percentage of defaults than the original dataset—or, in the worst case, zero defaults.

2.2.2 The Solution: Stratified Sampling

To prevent this, we employ stratified sampling. This is a technique that ensures the original class distribution of the target variable is preserved in each of the new data splits.

Here is the reasoning for its use:

Guarantees Representation: Stratification forces the splits to maintain the original ratio of defaults to non-defaults. If 0.0865% of the original dataset are defaults, the training set, validation set, and test set will all contain approximately 0.0865% defaults.

Enables Reliable Evaluation: When the test set is representative, the performance metrics we calculate (like accuracy, precision, recall, and F1-score) are meaningful. Evaluating a model on a test set with a skewed default rate would give us a misleading and over-optimistic (or pessimistic) score.

Promotes Model Generalization: By training the model on a set that accurately reflects the real-world data distribution, we help it learn the patterns of both the majority (non-default) and minority (default) classes, leading to a more robust and generalizable model.

In summary, using stratified sampling on the y variable is a critical step to ensure our model is trained and evaluated on a reliable, representative foundation.

Chapter 3

Generalized linear models (GLM)

3.1 Overview

Appendix A

Overview