PD Modeling under the IRB Approach: Technical Documentation and Model Report

Melina Hafelt

July 10, 2025

Contents

1 Abstract

This report outlines the methodology, validation, and performance evaluation of a **Probability of Default (PD)** model developed in accordance with the Internal Ratings-Based (IRB) approach. The dataset and modeling choices are entirely synthetic and constructed for demonstration purposes. The modeling pipeline includes preprocessing, feature engineering, logistic regression modeling, and robustness testing, aligned with Basel and CRR requirements.

Highlights:

- Logistic regression-based default probability estimation
- Comprehensive feature engineering and scaling
- Validation via AUROC, Brier Score, and calibration plots
- Stability and concentration risk analysis (PSI, Herfindahl Index)
- Stress testing with scenario-based parameter shocks

Note: All data are simulated for illustrative purposes only.

2 Theoretical Background

2.1 Definition of PD

The **Probability of Default (PD)** represents the likelihood that a borrower defaults within a given time horizon, typically 12 months. Formally, define a binary indicator:

$$D_i = \begin{cases} 1 & \text{if obligor } i \text{ defaults within 12 months} \\ 0 & \text{otherwise} \end{cases}$$

Then, the individual PD is given by:

$$\text{PD}_i = P(D_i = 1)$$

2.2 IRB Risk Components

Under Basel II/III, the IRB approach estimates capital requirements using the following parameters:

- **PD:** Probability of Default
- **LGD:** Loss Given Default

- **EAD:** Exposure at Default

Capital requirement is a function of:

$$\text{Capital} = f(\text{PD}, \text{LGD}, \text{EAD}, M, \text{corr})$$

where M is maturity and corr denotes correlation. This report focuses exclusively on PD.

3 Data Preparation and Quality Review

3.1 Synthetic Dataset Overview

The dataset includes the following key variables:

- `annual_income`
- `exposure_at_default`
- `credit_score_internal`
- `rating_grade`
- `default_flag` (binary target)

3.2 Feature Engineering

- **Log Income:** $\log(1 + \texttt{annual\_income})$
- **Loan-to-Income:** $\frac{\texttt{exposure\_at\_default}}{\texttt{annual\_income}}$
- **Credit Score Quantiles:** Discretized into 5 groups
- **Age Buckets:** Defined as 18–30, 31–45, etc.

3.3 Target Variable

The binary response is defined as:

$$Y_i = \begin{cases} 1 & \text{if defaulted} \\ 0 & \text{otherwise} \end{cases}$$

4 Model Development

4.1 Model Type: Logistic Regression

The PD is modeled using logistic regression:

$$\text{logit}(\pi_i) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}$$

$$\text{where } \pi_i = P(Y_i = 1 \mid \mathbf{x}_i)$$

4.2 Feature Scaling

Standardization of inputs is performed:

$$z_j = \frac{x_j - \mu_j}{\sigma_j}$$

4.3 Performance Metrics

- **AUROC:** Area Under ROC Curve
- **Gini Coefficient:** $2 \times \text{AUROC} - 1$
- **Brier Score:** Mean squared error of predicted probabilities

5 Model Validation

5.1 Calibration by Decile

Observed vs predicted default rates evaluated across PD deciles. Visualized using calibration plots with binomial confidence bands.

5.2 Hosmer–Lemeshow Test

A goodness-of-fit test based on observed and expected defaults:

$$\chi^2 = \sum_{g=1}^{G} \frac{(O_g - E_g)^2}{E_g(1 - \hat{p}_g)}$$

5.3 Population Stability Index (PSI)

Tracks changes in feature distribution over time:

$$\text{PSI} = \sum_{i=1}^{n} (A_i - E_i) \log\left(\frac{A_i}{E_i}\right)$$

5.4 Herfindahl Index

Measures portfolio concentration:

$$\text{HI} = \sum_{i=1}^{n} s_i^2$$

where s_i is the share of exposure in segment i.

5.5 Jeffreys Confidence Interval

Bayesian confidence interval using Jeffreys prior (Beta(0.5, 0.5)):

$$\text{CI} = \text{BetaInv}(\alpha/2, d + 0.5, n - d + 0.5)$$

6 Stress Testing

6.1 Scenario-Based Shocks

- Income reduced by 20%
- Credit score reduced by 50 points
- Increased loan-to-income ratios

Predicted PDs are recomputed under stress conditions to assess model sensitivity.

7 Simulation and Robustness

- Monte Carlo simulation with 10,000 resamples
- Scaling aligned with training data
- Resulting PDs visualized as distribution histogram

8 Conclusion

- The logistic PD model demonstrates strong calibration and discrimination
- Validation metrics meet IRB expectations
- Stress testing and stability metrics are included

Recommendation: PD model is suitable for IRB-aligned internal estimation and monitoring. Enhancements could include time-dependent or macro-driven modeling extensions.

Disclaimer

This document is fully synthetic and produced solely for demonstration purposes in a professional or academic context. No real data or sensitive information is used.