Probability of Default (PD) Modeling under the IRB Approach

Melina Hafelt

July 9, 2025

Abstract

This report presents a complete end-to-end implementation of a Probability of Default (PD) model within the Internal Ratings-Based (IRB) framework. All data and parameters are synthetic and intended for demonstration only. The pipeline covers data cleaning, feature engineering, model training, validation, and performance evaluation, in alignment with regulatory and statistical best practices. Logistic regression is used as the core modeling technique, with attention to calibration, discrimination, and robustness under stress.

1 Theoretical Background

1.1 Definition of PD

The Probability of Default (PD) is defined as the likelihood that a counterparty will default on its obligations within a given time horizon, usually 12 months. Formally, let D_i be a binary indicator for default:

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

Then PD is defined as:

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

1.2 IRB Risk Components

Under Basel II/III, the regulatory capital requirement for credit risk is based on:

- **PD:** Probability of Default
- **LGD:** Loss Given Default
- **EAD:** Exposure at Default

Capital is computed via a risk-weight function:

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

where M is maturity and corr is asset correlation. For this report, only PD is modeled.

2 Data Description and Preparation

2.1 Synthetic Dataset Overview

The dataset includes features across customer segments, exposures, credit scores, and macro indicators. Key variables include:

- `annual_income`, `exposure_at_default`, `rating_grade`, `credit_score_internal`, `default_flag`

2.2 Feature Engineering

- **Log-Transformation:**

$$\texttt{log\_annual\_income} = \log(1 + \texttt{annual\_income})$$

- **Loan-to-Income Ratio:**

$$\texttt{loan\_to\_income} = \frac{\texttt{exposure\_at\_default}}{\texttt{annual\_income}}$$

- **Quantile Binning:** Credit scores are discretized into 5 quantiles.
- **Age Buckets:** Cut into 18–30, 31–45, etc.

2.3 Target Variable

`default_flag` is the binary response variable:

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

3 Modeling Methodology

3.1 Logistic Regression

We use logistic regression to model PD:

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

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

3.2 Model Training and Scaling

Features are standardized to mean 0 and standard deviation 1 using:

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

3.3 Performance Metrics

- **AUROC:** Area under the ROC curve
- **Gini Coefficient:** $2 \times \text{AUC} - 1$
- **Brier Score:** Mean squared error of probabilities

4 Model Validation

4.1 Calibration

Compare observed default rate vs predicted PD in deciles. Visualized using calibration plots with binomial confidence intervals.

4.2 Hosmer–Lemeshow Test

Test goodness-of-fit:

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

where O_g is observed defaults and E_g is expected in bin g.

4.3 Population Stability Index (PSI)

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

Tracks feature distribution drift.

4.4 Herfindahl Index

Measures concentration:

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

where s_i is share of exposure in segment i.

4.5 Jeffreys Confidence Test

Uses Bayesian interval with Jeffreys prior (Beta(0.5, 0.5)):

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

4.6 Stress Testing

Apply shocks (e.g., income $\times 0.8$, credit score -50 pts) and recompute PD distribution.

5 Simulation and Robustness

Monte Carlo simulation with 10,000 bootstrapped observations.

- Scaled using original scaler
- PDs computed from model
- Output distribution plotted with histogram

6 Conclusion

This project demonstrates a complete pipeline for PD modeling in line with IRB expectations. Calibration, validation, and robustness techniques are implemented to regulatory standard. Future extensions may include time-dependent models, macroeconomic overlays, or Bayesian updating of PD curves.

Disclaimer

This document and its contents are fully synthetic and created solely to demonstrate modeling competence in a job application context. No real-world financial data or customer information is included.