Probability of Default Model Development and Validation with Generalized Linear Models

@by Donald Kpatcha

INTRODUCTION

Credit risk pertains to the potential that a borrower may fail to fulfill their loan obligations, posing a substantial concern for lenders like banks and financial institutions due to the possibility of financial losses, impacting profitability and stability. Evaluating and managing this risk is pivotal in lending operations, involving the utilization of diverse techniques and models. Factors such as credit history, repayment capability, loan terms, and annual income are scrutinized to assess this risk.

Various forms of credit risk include default risk, where borrowers may not meet payment obligations, credit spread risk, which involves losses due to widening credit spreads, counterparty risk, concentration risk, and country risk linked to investing or lending in specific countries due to political or economic instability.

Credit risk assessment employs several methodologies, including credit scoring models utilizing historical data and predictive variables, credit rating agencies assigning credit ratings based on repayment capacity, financial statement analysis, collateral evaluation, and consideration of qualitative factors like management quality and industry outlook.

Many organizations, especially financial institutions, evaluate credit risk for existing and potential customers, often leveraging machine learning to analyze customer data for risk profiling. Machine learning aids in determining the probability of default and assessing the financial impact on lenders in such scenarios.

The probability of default (PD) is a fundamental metric in credit risk modeling, serving as a key indicator of the likelihood that a borrower will fail to meet their debt obligations. It plays a central role in assessing the risk associated with lending and investment decisions, influencing credit pricing, portfolio management strategies, and regulatory compliance. By quantifying the likelihood of default for individual borrowers or portfolios, PD enables financial institutions to make informed decisions, allocate resources effectively, and mitigate potential losses in their credit portfolios.

PROBABILITY OF DEFAULT

Probability of default (PD) refers to the ability of a borrower to repay debt obligations. The higher this ability, the less likely the borrower will default.

Two generic key steps characterize the PD model development process: scorecard estimation and model calibration. One important step before diving into modelling is the *default definition and data preparation*. This step allows to define criteria or conditions for default. From a modelling perspective, a default flag is defined as binary variable. It conventionally assumes the value 0 if no default occurs, and 1 in the case of default.

Following **Basel II principles** “A default is considered to have occurred when: the banking institution considers that an obligor is unlikely to repay in full its credit obligations to the banking group, without recourse by the banking institution to actions such as realising security; or the obligor has breached its contractual repayment schedule and is past due for more than 90 days on any material credit obligation to the banking group”.

From an **IFRS9 perspective**, it does not directly define default, but requires entities to align with internal credit risk management. A rebuttable presumption holds. Default does not occur later than a financial asset is 90 days past due.

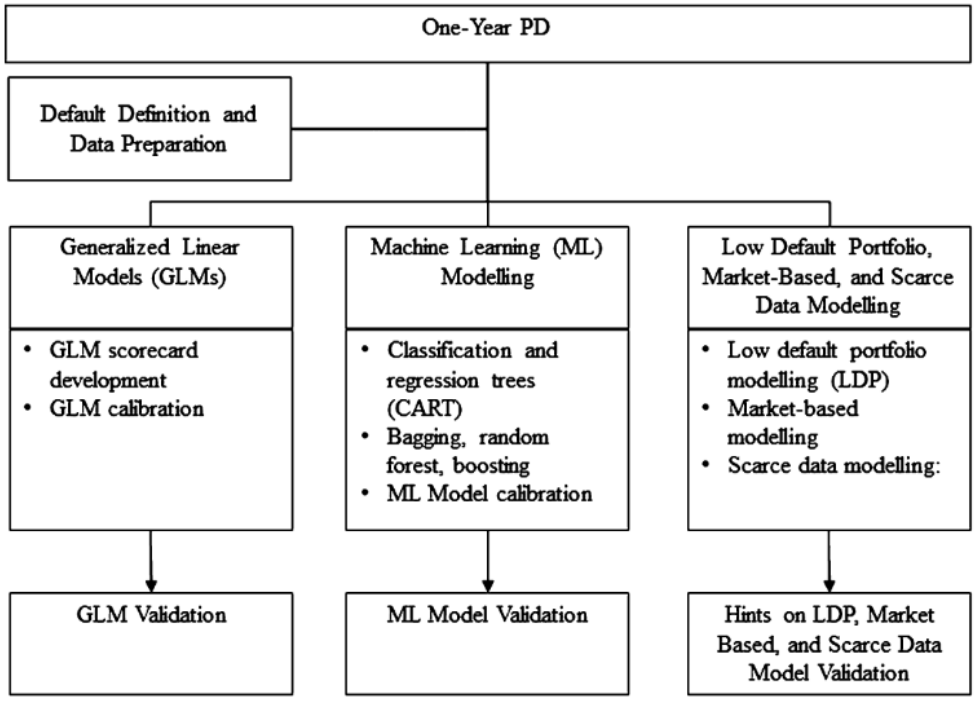

Figure 1: Summary of the main steps for modeling Probability of Default. Two main approaches, including Generalized Linear Models and Machine Leaning. Figure from Tiziano Bellini 2019.

Some factors can trigger the default.

- Quantitative indicators: 90 days past due or 3 months in arrears are common default triggers. A counter is put in place to account for delays or arrears balance.
- Qualitative indicators: The following elements are commonly used in practice as default triggers:
  - Bankruptcy: This event may be triggered with regards to a bank's specific exposure, or due to exposures towards other institutions.
  - Maturity or term expiry: Default is caused by an outstanding balance due after the maturity date.
  - Other indicators, such as forbearances, play a major role as a default trigger.

Next, there are several methods of modelling the PD.

1. **Generalised linear models (GLMs)**

GLMs can be used for PD model development, from scorecards development analysis, to its calibration on the most recent historical period under investigation (that is, point-in-time (PIT) estimation). This approach is the most commonly adopted in the banking industry.

2. **Machine learning (ML) modelling**

Big data play a major role in our economy. Techniques dealing with them are in high demand for credit risk purposes. Machine learning procedures by means of classification and regression trees (CARTs), bagging, random forest, and boosting can be used. These methods provide a consistent challenge to existing models (for example, GLMs), and allow us to explore alternative ways of analysing data and estimate one-year PDs.

3. Low default portfolio, market-based, and scarce data modelling

If wide data availability boosts ML utilisation, data scarcity remains a key challenge. This is due either to lending business characteristics, as in the case of low default portfolios (example, insurance or banking industries), or banks' difficulties to collect data. These issues can be tackled by means of bespoke approaches, such as Pluto and Tasche (2005), distance to default, and more qualitative methods based on expert judgements.

In this note, I will focus on summarizing what to know before modelling PD with GLMs. The other methods are left for dedicated articles or notes.

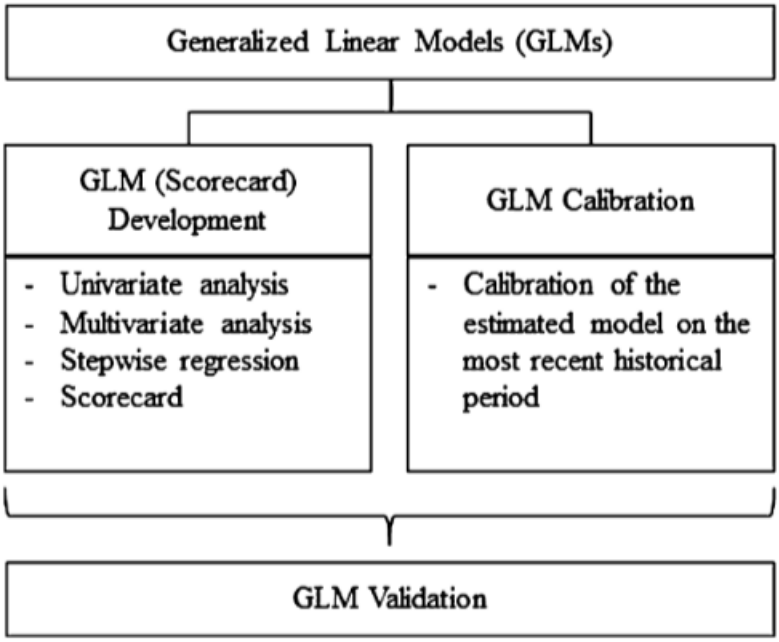

Figure 2: Generic steps for modelling PD with Generalized Linear Models. Figure from Tiziano Bellini (2019).

GLMs MODELLING

A two-step process:

- **GLM development (Scorecard**): A credit scorecard can be developed by means of GLMs (e.g. logit regression). The main goal of a scorecard is to rank accounts based on default likelihood.

  Default flag can be defined by using a set of explanatory variables to fit a binary response representing defaults,

$$y_i = \begin{cases} 1 & default, \\ 0 & non\text{-}default, \end{cases}$$

  where *i* indicates a account within the portfolio under investigation, or consider alternative functional forms, like the logistic relationship

$$SC_i = \frac{1}{1 + e^{-(\beta_0 + \beta_1 \chi_{i,1} + \ldots + \beta_k \chi_{i,k})}},$$

  where SC_i and χ_i are the score for account *i*, and a vector of explanatory behavioral variables. Note that the parameter $\ln(SC_i/(1-SC_i))$ is the commonly denominated logit.

- **GLM calibration and Validation:** This process permits to rescale a scorecard to derive PDs. As an additional step, one may aggregate accounts in rating classes. This process is particularly relevant for lifetime PD modelling.

Steps for GLM development (Scorecard)

Here are the sequential steps for developing scorecards with GLM.

1. Default flags definition and data preparation
2. Univariate analysis (Information Value (IV) assessment)
3. Multivariate analysis
4. Stepwise regression
    - Discard highly correlated variables
    - Stepwise model fitting

Calibration

1. *GLM Calibration (score normalization):* A score represents the relative creditworthiness of an account compared to others within a portfolio. To begin, it's often prudent to normalize scores within a defined interval as an initial step. This normalization process helps standardize the scoring system, making it easier to interpret and compare scores across different accounts or time periods.
2. *PD Calibration:* The goal of Probability of Default (PD) calibration at Point-in-Time (PIT) is to align the average portfolio PD with its most recent default rate. Calibration can be accomplished using various methods, often involving mathematical functions. When there's ample internal default data, logistic regression is a favored choice. However, if internal data is insufficient, external sources such as ratings may be used. In such cases, linear or log-linear regression is commonly employed for PD calibration. When neither internal nor external data is available, ad-hoc procedures must be considered. These procedures aim to adjust the model's PD estimates based on available information and expert judgment.
3. *Rating class assignment:* The final stage of calibration involves mapping Probability of Default (PD) values to rating classes, a necessary step for Internal Ratings-Based (IRB) purposes, as risk-weighted assets (RWAs) are dependent on ratings. However, for accounting standards such as IFRS 9 and CECL, this mapping is not strictly required, as PDs can be directly utilized to calculate expected credit losses. Once a rating grid is established, a straightforward mapping process assigns each account to the appropriate rating band, simplifying implementation and facilitating risk assessment.

Validation

1. *Data*

   The data validation process can be organised by means of:
   - Data representativeness: One of the first issues one faces in developing a one-year PD model is to use internal or external sources. Irrespective to the source of the data available, one have to demonstrate that data used for modelling truly represent the phenomenon under investigation.
   - Variable appropriateness: A check is required to show that all variables used for modelling are relevant and appropriate for the scope of one-year PD modelling. The variables representing facts and events not relevant for the analysis should be disregarded.
   - Data completeness: One needs to consider a database wide enough to include information over the most relevant past history to support PIT estimates.

2. *Methodology*

   The primary benefits of GLMs lie in their simplicity and the straightforward interpretation of their outcomes. Additionally, GLMs offer a rigorous and robust statistical framework for the development and calibration of PD models. However, it's crucial to reassess key assumptions, considering data availability,

and to compare against alternative methods to ensure the model's reliability and accuracy across different contexts.

3. *Statistical testing*
   - The disciminatory power calibration: example of metrics that can be used for assessing the model's discriminatory power:
     - Receiver operating characteristic (ROC) curve,
     - Area under the curve (AUC),
     - Gini index,
     - Information value (IV), which is specific to assessing the predictive power of individual variables in binary classification tasks like credit scoring.
     - Weight of evidence (WOE).
   - To verify the alignment of the model's predictions with actual defaults, a comparison between actual and fitted PDs (by score band) is essential. Control measures are necessary to validate the model's calibration effectively. Below is a brief overview of the key checks typically conducted to ensure calibration.

     Comparison of actual and fitted PDs by score band: This step evaluates how well the model's predicted probabilities align with observed default rates across different score bands.

     Calibration process: Ensuring the average portfolio PDs are effectively aligned is crucial. This involves adjusting the model to match the expected default rates in the portfolio.

     Definition of tolerance bands: Tolerance bands can be established to set acceptable limits for the disparity between predicted and observed default rates. These bands can be defined in terms of both absolute levels and relative thresholds to provide a comprehensive assessment of model performance.

4. *Out-of-sample and out-of-time stability*

   Stability checks aim to assess whether changes have rendered a model unsuitable for its intended purpose. Initially, it's crucial to compare the population currently using the model to the original development population. Any disparities require an examination of individual variable stability within the model. Commonly adopted metrics for assessing model stability include:
   - *Stability Index:* comparison is made between model development and actual current population. The difference between populations is measured by means of a Stability Index (*SI*). One may point out the similarity with information value.

$$SI = \sum_{j=1}^{J} (R_j - O_j) \ln \frac{R_j}{O_j},$$

     In this formula, j indicates the class under analysis, R_j represents the percentage of the reference population in class j, and O_j is the percentage of the observed population in class j. High values of *SI* indicate more substantial shifts in the population. As a rule of thumb, when $SI \leq 0.1$, no significant shift occurred. In case of $0.1 < SI \leq 0.25$, then, a minor shift occurred, whereas in the case of $SI > 0.25$, a major shift took place. Expert intuition is also important to mitigate the rigidity of this classification.
   - *Other statistics can also be used for the purpose of verifying PD stability.* Furthermore, qualitative checks are usually recommended to have a full picture of how a model behaves under different circumstances.

5. *Cross-validation*

   Cross-validation involves partitioning the dataset into g folds. A model is built using g-1 training folds and evaluated on the remaining validation fold. This process is iterated for all possible validation folds,

producing g performance estimates. Model performance metrics such as AUC or Gini index can then be averaged to assess overall performance.

6. *Reproducibility*
   One of the primary challenges in replicating a PD model lies in managing the data. Comprehensive data documentation and careful storage practices are integral to ensuring reproducibility. Additionally, it's essential to provide thorough descriptions of the entire process and the software utilized for both development and implementation. Finally, conducting a replication study using the original dataset employed for model creation is imperative.

REFERENCE:

Tiziano Bellini 2019. IFRS 9 and CECL Credit Risk Modelling and Validation. A Practical Guide with Examples Worked in R and SAS. ISBN 9780128149409, https://doi.org/10.1016/C2017-0-02756-8.