**AI-Powered Probability of Default (PD) Modelling: A Case Study in Credit Risk Analytics**

**Abstract**

This study focuses on modelling the *Probability of Default (PD)*—a central concept in credit risk management that underpins critical banking activities such as regulatory capital calculations, credit decisioning, and portfolio risk monitoring. The goal is to explore PD not only from a statistical modelling perspective but also through business and analytical lenses, where understanding the purpose behind prediction is as essential as achieving model accuracy.

To this end, I adopt a tri-fold role:

* As a business analyst, I emphasize the importance of aligning model development with its intended use case—whether it’s complying with Basel III and IFRS 9 frameworks, improving credit scoring decisions, or enhancing risk-based pricing strategies. This ensures that the model delivers actionable insights and contributes meaningfully to risk-aware decision-making.
* As a data analyst, I bridge the gap between raw data and business context by interpreting patterns, spotting anomalies, and identifying key risk drivers. This role is crucial in transforming data into relevant features and metrics that improve model clarity and usability.
* As a data scientist, I implement a modular, open-source data science solution—Daanish (meaning *knowledge* in Persian)—to operationalize the full modelling pipeline. Built to be scalable, reusable, and interpretable

This integrated approach enables the creation of a full-stack PD modelling solution that is analytically rigorous, business-aware, and data-informed. Each step is rooted in real-world decision-making scenarios, empowering stakeholders not just to build predictive models, but to understand, trust, and act on them.

The entire implementation is based on a dataset from *DataCamp’s Credit Risk Modeling in Python* course, and all components are developed using the Daanish platform to ensure transparency, repeatability, and extensibility.

**1. Introduction: Understanding Probability of Default in Modern Banking**

In the realm of financial risk management, Probability of Default (PD) is a cornerstone concept. It refers to the likelihood that a borrower will fail to meet their debt obligations, either fully or partially, within a specified time frame—usually one year. This probability plays a critical role in credit risk assessment, capital adequacy calculations, and financial reporting under both regulatory frameworks and accounting standards.

The importance of PD modelling has only intensified following financial crises such as the 2008 global meltdown, where underestimated credit risk and overvalued securities led to widespread losses. Since then, global banking regulators and accounting bodies have introduced frameworks such as the Basel Accords and IFRS 9, which place heavy emphasis on the accurate estimation of expected credit losses.

Given the centrality of PD in both compliance and performance-based credit strategy, this study aims to model PD with a dual focus: ensuring alignment with business decision-making needs while maintaining analytical robustness. To this end, the study presents a modular solution, *Daanish*—a Python-based, open-source framework that supports every stage of PD modelling, from data understanding to model validation. By combining business insight with advanced analytics, this research addresses both the *why* and *how* of PD prediction.

**1.1 The Role of PD in Risk and Regulation**

Probability of Default (PD) plays a foundational role in both regulatory compliance and internal risk-based decision-making.

Under Basel II and III, PD is one of the three key inputs — alongside Exposure at Default (EAD) and Loss Given Default (LGD) — used to estimate credit risk and determine how much capital a bank should hold. These frameworks help ensure that banks remain solvent during stress events by aligning capital buffers with underlying credit risks.

In IFRS 9, PD is part of a forward-looking expected credit loss (ECL) model, which aims to provide a more timely and realistic recognition of losses in financial statements — a major shift from the incurred-loss model that contributed to delayed loss recognition during the 2008 crisis.

Beyond compliance, PD is also critical in internal risk management and performance measurement:

* **Value at Risk (VaR)** models incorporate PD when estimating the potential for unexpected losses in credit portfolios. PD influences the tail-risk behaviour of loss distributions, shaping stress-testing and capital adequacy assessments.
* **RAROC (Risk-Adjusted Return on Capital)** uses PD to estimate expected losses and adjust returns accordingly. By embedding PD, RAROC enables banks to compare the profitability of different customers or business lines relative to their risk, promoting more informed lending and pricing decisions.

In essence, PD is not just a regulatory checkbox — it underpins capital planning, credit strategy, loan pricing, and portfolio optimization, making it one of the most impactful metrics in the banking world. These diverse applications underscore the need for PD models that are not only statistically valid but also interpretable and purpose-driven. In this study, these considerations inform the selection of variables, the structuring of the model pipeline, and the evaluation criteria — all of which are embedded into the Daanish framework to ensure practical business value and compliance-readiness.

**1.2 Methods for Modelling PD**

Over the years, a range of techniques has been employed to estimate PD, including:

* **Statistical methods** (e.g., logistic regression, discriminant analysis),
* **Machine learning models** (e.g., decision trees, random forests, neural networks),
* **Expert systems and rule-based scoring**
* **Hybrid approaches** that combine financial knowledge with AI-based predictions.

Each method has its merits and limitations depending on data availability, regulatory constraints, and business requirements. However, modelling PD is not a one-size-fits-all task—the assumptions, inputs, and even the interpretation of default can vary across jurisdictions and institutions.

In this study, we begin with logistic regression for its interpretability and widespread acceptance in regulated environments, and then explore tree-based models such as Random Forests and Gradient Boosting to improve predictive power. The Daanish platform enables seamless integration of these methods while supporting essential preprocessing, feature selection, and evaluation techniques that respect both analytical soundness and business requirements.

**2. Objective of This Study**

The objective of this study is not to advocate for a single best model, but to provide a comprehensive, real-world guide for financial modelers, data analysts, and data scientists on building Probability of Default (PD) models that are both effective and meaningful in practical banking contexts.

This study adopts a holistic perspective by bridging three interconnected roles:

* As a *business analyst*, it emphasizes understanding the *strategic purpose* of PD estimation—be it regulatory compliance, credit approval, or pricing.
* As a *data analyst*, it explores how to extract, clean, and engineer meaningful features from complex loan datasets.
* And as a *data scientist*, it addresses how to build, evaluate, and interpret models with transparency and scalability in mind.

Rather than focusing narrowly on algorithmic details, the study highlights the key steps, challenges, and considerations that shape the end-to-end modelling process—from data ingestion and EDA, to model training and post-modelling business impact.

This is achieved through the development of **Daanish**, an open-source, modular data science solution specifically designed to operationalize every stage of the PD modelling lifecycle. With Daanish, the study demonstrates how to implement robust analytics pipelines that are reusable, interpretable, and adaptable to different credit risk scenarios.

Ultimately, the goal is to empower practitioners with a structured framework that ensures PD models are not only statistically sound but also aligned with the broader objectives of credit risk management.

**3. Roadmap for PD Modelling**

This section outlines the end-to-end modelling pipeline adopted in this study using the Daanish platform. Each step is designed to ensure that both technical rigor and business relevance are maintained throughout the process—from defining objectives and wrangling data to deploying interpretable models that support real-world decision-making.

* **Step 1: Data Collection and Business Understanding**

Define the business objective clearly — understand whether the PD model is for regulatory compliance, credit decisioning, risk-based pricing, or portfolio monitoring.

Gather data from internal (e.g., loan records, repayment history) and external sources (e.g., credit scores, macroeconomic indicators).

*Why it matters:* Aligning the data collection with the business goal ensures relevance, focus, and practical value of the model.

* **Step 2:** **Preliminary Exploratory Data Analysis (EDA)**

Perform basic descriptive statistics and visualizations (e.g., histograms, scatter plots, boxplots) to understand feature distributions and relationships. Early crosstab analyses to reveal interactions between categorical variables and default status.

*Why it matters:* This step lays the groundwork for more rigorous preprocessing by helping us ask better questions — which features appear relevant, where potential data quality issues may arise, and whether any early patterns are already emerging.

* **Step 3: Data Preprocessing**

Encode categorical variables, scale features, impute missing values, and address class imbalance.  
Check for data leakage and correct data anomalies.

*Why it matters:* This prepares the data for modelling, improving both accuracy and generalizability.

In Daanish, these preprocessing steps are modularized and traceable, supporting transparency and audit-readiness for regulatory contexts.

* **Step 4: Full Exploratory Data Analysis (EDA)**

With clean, pre-processed data:

* Analyse feature correlations, multicollinearity, and interactions.
* Use PCA or clustering to explore structure.
* Assess relationships with the target variable (Default).

*Why it matters:* Post-preprocessing EDA uncovers insights that inform feature engineering and model design.

* **Step 5: Feature Engineering and Selection**

Generate new features (e.g., default history ratios, credit utilization) and borrower aggregates (e.g., delinquency rates, recent defaults).

Use statistical techniques and model-based methods to reduce dimensionality and eliminate multicollinearity.

*Why it matters:* Meaningful features drive predictive power and generalisability.

* **Step 6: Model Building**

Split the dataset into training and test sets.

Select appropriate algorithms (e.g., logistic regression, tree-based models, ensemble methods) and train using cross-validation.

Tune hyperparameters for optimal performance.

*Why it matters:* A well-trained model is both accurate and robust when faced with new data.

* **Step 7: Model Evaluation**

Evaluate using appropriate metrics such as AUC-ROC, Precision, Recall, F1-score, and KS-statistic.  
Use confusion matrices and lift charts to interpret performance, especially in imbalanced datasets.

*Why it matters:* Evaluation metrics ensure the model effectively identifies potential defaulters without overpredicting risk.

* **Step 8: Model Validation and Testing**

Test the final model on a hold-out set or through time-based validation to assess real-world performance.  
Check for overfitting, generalization, and stability across different segments.

*Why it matters:* Validation builds confidence in the model’s reliability and fairness when deployed.

* **Step 9: Model Deployment and Prediction**

Deploy the model into production for real-time or batch scoring.

Integrate with business systems and set up ongoing monitoring for model drift, data drift, and performance degradation.

*Why it matters:* Deployment turns insights into action, allowing the business to make informed, data-driven credit decisions.

This structured roadmap provides a blueprint not only for building accurate PD models but also for ensuring their relevance and resilience in high-stakes banking environments. In the next section, we demonstrate this process step-by-step using a synthetic but realistic dataset from DataCamp’s *Credit Risk Modelling in Python* course. While not derived from actual financial institutions, the dataset reflects key characteristics and challenges of real-world credit risk scenarios, making it well-suited for educational and prototyping purposes.

**4. Data Collection and Business Understanding**

Every effective predictive modelling project begins with a deep understanding of the business problem it aims to solve. From a business analyst's perspective, this step is not simply about gathering data — it's about aligning data collection and model design with strategic business objectives.

In the case of Probability of Default (PD) modelling, it’s critical to first establish the specific business purpose the model will serve. For example:

* **Regulatory Compliance**: Meeting Basel III or IFRS 9 standards requires adherence to strict model validation and governance procedures. This often involves using conservative assumptions and maintaining full audit trails of model decisions.

*Banco Bilbao Vizcaya Argentaria (BBVA)* has implemented comprehensive PD models to comply with regulatory standards such as Basel III. These models assess the creditworthiness of borrowers by estimating the likelihood of default within a year. BBVA employs various scoring tools—reactive, behavioural, proactive, and bureau scoring—to evaluate retail credit products like consumer loans, mortgages, and credit cards.

* **Credit Decisioning**: Models influence real-time lending approvals and thus require high interpretability (e.g., decision trees or logistic regression over black-box models).  
  An Indian private sector bank developed a PD model using logistic regression to enhance credit decisioning. By analysing employment status, debt-to-income ratio, and monthly expenses, the bank improved the accuracy and transparency of lending decisions.
* **Risk-Based Pricing**: Aligning loan interest rates with borrower risk levels helps optimise profitability while remaining competitive. PD scores may be combined with LGD and EAD to support a risk-adjusted pricing strategy.

*G-Square Solutions* collaborated with a major Indian bank to build a PD model identifying creditworthy customers for temporary lending facilities. The model helped adjust interest rates based on risk, enhancing profitability.

* **Portfolio Monitoring and Stress Testing**: Supports ongoing assessment of credit quality for capital planning, provisioning, and early warning systems — especially under adverse economic conditions.  
  A study of *Shinkin banks in Japan* used PD models to evaluate financial stability, enabling proactive credit risk management and robust stress testing frameworks.

Each use case affects:

* The choice of features (e.g., repayment patterns for monitoring vs. application details for credit decisioning),
* The acceptable model complexity (simple models for operations, more complex ones for stress testing),
* And the interpretation of results by business stakeholders.

A clear business use case ensures that the model’s insights are relevant, explainable, and directly tied to actionable decision-making — a core principle in effective business analysis.

**4.1 Data Collection Strategy**

Once the business goal is defined, the next step is identifying and sourcing relevant data, guided by domain knowledge and stakeholder input. A robust PD model integrates a variety of loan-specific, borrower-specific, and credit history-related features, including:

* **Internal Data**: Loan application records, repayment behaviour, default status, and collateral information. These data points provide direct insight into borrower behaviour and are foundational for credit risk modelling (Qi and Zhao, 2011).
* **External Data**: Credit bureau scores, regional economic indicators, and macroeconomic factors (e.g., unemployment rates, interest rate levels) are used to capture systemic risks (Miyamoto and Takeda, 2020).
* **Derived Metrics**: Aggregated risk indicators across historical loans (e.g., Weighted Average Loan-to-Value (LTV), Recent Default Indicator), payment behaviour ratios (e.g., Early Payment Rate, Delinquency Rate), and risk-based segmentation (e.g., high-risk region categorization) help uncover hidden patterns in borrower behaviour (Anderson, 2007).

At this point, no transformations or filtering are applied. The aim is to preserve all potentially useful information for downstream analysis. Even features that seem noisy or redundant may carry predictive value in combinations or under specific conditions.

*Note*: While minimal formatting (e.g., date parsing, type checks) may occur here, full preprocessing — including missing value handling, feature encoding, and normalisation — is deliberately deferred to ensure more thoughtful, context-aware treatment.

**4.2 Dataset Used in This Study**

To illustrate the end-to-end process of building a Probability of Default (PD) model, we use a synthetic dataset sourced from the *Credit Risk Modeling in Python course on DataCamp*. While the data is fictional, it captures many of the fundamental characteristics of real-world lending portfolios, making it a practical resource for prototyping and educational purposes.

The dataset contains the following features:

|  |  |
| --- | --- |
| Column | Description |
| Age | Applicant’s age (integer) |
| Income | Total yearly income (integer) |
| Home ownership | Current home ownership type (categorical: rent, own, mortgage) |
| Employment length | Years employed (decimal) |
| Loan intent | Purpose of the loan (e.g., debt consolidation, personal, medical) |
| Loan grade | Credit grade of the loan (categorical: A, B, C, ...) |
| Loan amount | Amount of the loan (integer) |
| Interest rate | Annual interest rate (decimal) |
| Loan status | Target variable (0 = non-default, 1 = default) |
| Debt to income | Share of income used for debt repayment (decimal) |
| Defaulted before | Whether the applicant has previously defaulted (boolean) |
| Credit history length | Number of years since credit history started (integer) |

While the dataset lacks advanced features such as collateral types, LTV ratios, or borrower-level aggregates, the modelling structure we apply is designed to scale easily. In a production environment, these additional business-driven features would significantly enhance model robustness, regulatory compliance, and actionable insights.

**4.2.1 Justification of Key Features**

To build an effective and interpretable PD model, it is essential to include features that align with credit risk theory and empirical findings in lending analytics. Below are justifications for some of the most influential features used in this study:

* **Loan Intent**

The purpose for seeking credit reflects underlying borrower motivations and risk profiles. For instance, debt consolidation loans might carry lower risk than speculative investments or medical loans. Including this feature helps the model capture patterns tied to financial intent, which has a known impact on creditworthiness (Thomas, Crook & Edelman, 2017).

* **Loan Grade**

This feature provides a credit assessment score assigned by the lender based on a variety of underwriting factors. It is a composite indicator of perceived risk and has historically shown strong predictive power in determining default probability (Anderson, 2007).

* **Loan Amount**

Larger loan amounts tend to increase financial burden and may elevate repayment stress, especially for lower-income borrowers. Thus, the loan amount is a direct contributor to exposure at default and is central to credit risk assessment (Basel Committee, 2006).

* **Interest Rate**

The interest rate reflects the lender’s pricing of risk. Higher rates often indicate weaker borrower profiles or higher default probabilities, making this a key predictive feature for supervised learning models.

* **Loan Status**

This is the target outcome for PD modelling—whether or not the borrower defaulted. Accurate labelling of this variable is critical for supervised learning tasks. Any records with missing loan status are unsuitable for training and must be excluded.

* **Defaulted Before**

Past behaviour is one of the strongest indicators of future outcomes. A prior default strongly signals elevated risk, and its inclusion allows the model to capture borrower-level credit history without relying solely on credit bureau scores (Altman & Saunders, 1998).

* **Debt-to-Income Ratio**

This ratio measures the borrower’s capacity to service additional debt. High DTI ratios are often linked to financial distress and are commonly used in industry underwriting practices to evaluate repayment ability (Federal Reserve, 2020).

* **Employment Length & Income**

These features serve as proxies for financial stability and earning power. Longer employment and higher income typically correspond to lower default risk, reinforcing the borrower’s capacity to meet future obligations.

It is important to note that the above justifications represent initial hypotheses grounded in domain knowledge and credit risk theory. Their ultimate inclusion and relative importance in the final model will be determined through data exploration, correlation analysis, and performance evaluation techniques during later stages of the modelling pipeline.

**4.3 Preliminary Exploratory Data Analysis (EDA): Getting to Know the Data**

Before diving into data cleaning or model building, it is valuable to conduct a preliminary exploratory analysis to develop an initial sense of the dataset. At this stage, no assumptions are made, and no transformations are applied — the data remains raw but structured. The goal is simply to observe, not to judge.

Using simple descriptive statistics, distribution plots, and basic visualizations such as histograms, scatter plots, and boxplots, we begin forming a mental model of how different features behave, how they relate to one another, and how they might influence the target variable. For example, we examine how loan amount varies across loan grades, or how loan status differs by intent or home ownership. Additionally, early crosstab analyses offer insights into interactions between categorical variables and default status.

This step lays the groundwork for more rigorous preprocessing by helping us ask better questions: Which features seem relevant? Where might data quality issues arise? Are some patterns emerging already? Although no cleaning or feature engineering is performed at this point, this foundational understanding ensures that subsequent steps are guided by data-driven intuition.

**5. Preliminary Exploratory Data Analysis (EDA)**

**References**

Altman, E.I. and Saunders, A., 1998. *Credit risk measurement: Developments over the last 20 years*. Journal of Banking & Finance, 21(11-12), pp.1721–1742.

Anderson, R., 2007. *The Credit Scoring Toolkit: Theory and Practice for Retail Credit Risk Management and Decision Automation*. Oxford: Oxford University Press.

BBVA, 2012. *Probability of Default (PD)*. [online] BBVA Shareholders and Investors. Available at: <https://shareholdersandinvestors.bbva.com/microsites/bbva2012/en/Riskmanagement/ProbabilityofdefaultPD.html> [Accessed 4 Apr. 2025].

Basel Committee on Banking Supervision (2006) *International Convergence of Capital Measurement and Capital Standards: A Revised Framework – Comprehensive Version*. Bank for International Settlements. Available at: <https://www.bis.org/publ/bcbs128.htm> (Accessed: 4 April 2025).

Basel Committee on Banking Supervision (2011) *Basel III: A global regulatory framework for more resilient banks and banking systems*. Bank for International Settlements. Available at: <https://www.bis.org/publ/bcbs189.htm> (Accessed: 4 April 2025).

Crouhy, M., Galai, D. and Mark, R. (2006) *The Essentials of Risk Management*. New York: McGraw-Hill.

Federal Reserve Board, 2020. *Consumer Credit Reports: A Study of Medical and Non-Medical Collections*. Washington, D.C.: Board of Governors of the Federal Reserve System.

G-Square Solutions, n.d. *PD and Credit Risk Model for a Large Bank*. [online] Available at: <https://g-square.in/pd-and-credit-risk-model-for-a-large-bank/> [Accessed 4 Apr. 2025].

Hori, M., 2012. *Predicting Bank Failure Using a Hazard Model: The Case of Shinkin Banks*. Economic and Social Research Institute, Cabinet Office, Government of Japan. [online] International Accounting Standards Board (2014) *IFRS 9 Financial Instruments*. IFRS Foundation. Available at: <https://www.ifrs.org/issued-standards/list-of-standards/ifrs-9-financial-instruments/> (Accessed: 4 April 2025).

Jorion, P. (2007) *Value at Risk: The New Benchmark for Managing Financial Risk*. 3rd edn. New York: McGraw-Hill.

Miyamoto, S. and Takeda, Y., 2020. *SME default prediction framework with the effective use of external public credit data*. [online] ResearchGate. Available at: [https://www.researchgate.net/publication/344974850\_SME\_Default\_Prediction\_Framework\_with\_the\_Effective\_Use\_of\_External\_Public\_Credit\_Data](https://www.researchgate.net/publication/344974850_SME_Default_Prediction_Framework_with_the_Effective_Use_of_External_Public_Credit_Data%20) [Accessed 4 Apr. 2025].

Qi, M. and Zhao, X., 2011. *Comparison of modeling methods for consumer credit risk*. Journal of Forecasting, 30(6), pp.547–562. Available at: [https://www.sciencedirect.com/science/article/abs/pii/S1544612316300605](https://www.sciencedirect.com/science/article/abs/pii/S1544612316300605%20) [Accessed 4 Apr. 2025].

Resti, A. and Sironi, A. (2007) *Risk Management and Shareholders’ Value in Banking: From Risk Measurement Models to Capital Allocation Policies*. Chichester: Wiley.

Saunders, A. and Allen, L. (2010) *Credit Risk Management In and Out of the Financial Crisis: New Approaches to Value at Risk and Other Paradigms*. 3rd edn. Hoboken, NJ: Wiley Finance.

TransOrg Analytics, n.d. *Credit Risk Modelling – A Comprehensive Case Study*. [online] Available at: <https://www.transorg.com/case-study/credit-risk-modelling-a-comprehensive-case-study/> [Accessed 4 Apr. 2025].

Available at: <https://www.esri.cao.go.jp/en/esri/archive/e_dis/abstract/e_dis349-e.html> [Accessed 4 Apr. 2025].