**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 (Part 1 – Cleaning and Validation)**

Handle missing values and detect/treat outliers. Check for data leakage and correct anomalies that could distort analysis.

*Why it matters:* Handle missing values and detect/treat outliers. Check for data leakage and correct anomalies that could distort analysis.

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

With clean, validated data:

* Analyse feature correlations, multicollinearity, and interactions.
* Use dimensionality-reduction techniques (e.g., PCA) or clustering to explore structure.
* Assess relationships between features and the target variable (Default).

*Why it matters:* Post-cleaning EDA uncovers deeper insights that inform feature engineering and help avoid redundancy or misleading predictors.

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

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

Apply statistical and model-based methods to select informative variables and eliminate multicollinearity.

*Why it matters:* Constructing meaningful, non-redundant features is key to predictive power and model generalisability.

* **Step 6: Final Preprocessing for Modelling**

Encode categorical variables, scale numerical features, and address class imbalance through resampling if needed.

*Why it matters:* These transformations prepare the dataset for model training, ensuring compatibility with algorithms and preventing bias from class imbalance.

* **Step 7: 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 8: 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 9: 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 10: 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)**

Before diving into building a probability of default (PD) model, it's essential to understand the structure, quality, and distribution of the data. Preliminary Exploratory Data Analysis (EDA) helps uncover potential data quality issues such as missing values, duplicates, or outliers, while also revealing useful insights about variable types, ranges, and relationships. This process ensures the dataset is well-prepared for modelling and that key patterns or anomalies are not overlooked in the early stages.

In this section, I will walk through the key steps of the preliminary EDA conducted on the dataset, including an overview of its structure, an assessment of missing values, the distribution of each feature, and basic descriptive statistics.

**5.1 Dataset Overview**

The dataset used for this analysis contains 32,581 records and 12 features, capturing a mix of borrower attributes, loan characteristics, and credit-related information. Here's a brief summary of the data structure:

Data Types:

* Numerical (int64/float64): 7 features, including person\_age, person\_income, loan\_amnt, loan\_int\_rate, and cb\_person\_cred\_hist\_length.
* Categorical (object/bool): 5 features, such as person\_home\_ownership, loan\_intent, loan\_grade, and cb\_person\_default\_on\_file.

This combination of numerical and categorical variables offers a comprehensive view of both the applicants' demographic and financial profiles.

Additionally, 165 duplicate records were identified, which may indicate data collection or entry issues. These duplicates will be considered for removal to avoid potential bias or redundancy in the modelling phase.

Understanding the dataset's structure early on helps guide subsequent data cleaning, transformation, and feature engineering steps.

Identifying missing data is a critical part of exploratory analysis, as it directly influences the choice of imputation strategies or whether to exclude certain records or features altogether.

Here’s a summary of the missing data in the dataset:

|  |  |  |
| --- | --- | --- |
| Feature | Missing Values | Missing % |
| Employment Length | 895 | 2.75% |
| Interest Rate | 3116 | 9.56% |
| Debt to Income | 9 | 0.03% |

Table 1: Missing data in the dataset

The remaining features are complete with no missing values.

**Interpretation & Next Steps:**

* “Interest Rate” has the highest proportion of missing values (nearly 10%). Given its importance in financial modelling, we should consider whether the missingness is random or related to certain borrower traits. Depending on this, we may impute values using median/mean by loan grade, or flag them as a separate category.
* “Employment Length” has moderate missingness (~2.75%). This might be imputable based on similar income/age groups or handled with binning strategies.
* “Debt to Income” has minimal missing data and can be safely imputed with the median without much concern.

These decisions will be further refined during the data cleaning phase, but understanding where and how much data is missing sets the foundation for a robust preprocessing pipeline.

**5.2 Descriptive Statistics and Distribution Insights**

This step provides a snapshot of each numerical feature’s central tendencies, variability, and distributional shape. These insights help identify data skewness, outliers, or transformations that might be needed before modelling.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Feature | Mean | Std Dev | Skewness | Kurtosis | Remarks |
| Age | 27.73 | 6.35 | 2.58 | 18.56 | Highly skewed right with extreme outliers (max = 144). Consider capping. |
| Income | 66,075 | 61,983 | 32.87 | 2693.27 | Extreme right skew and kurtosis. Outliers likely present, needs log scaling. |
| Employment Length | 4.79 | 4.14 | 2.61 | 43.72 | Skewed right. Outliers may indicate data entry errors (max = 123). |
| Loan Amount | 9,589 | 6,322 | 1.19 | 1.42 | Mild right skew. Distribution looks typical for loan applications. |
| Interest Rate | 11.01 | 3.24 | 0.21 | -0.67 | Fairly symmetric. Missing values need attention but distribution is normal. |
| Debt to Income | 0.17 | 0.11 | 1.07 | 1.22 | Slight right skew. Might reflect higher burden on lower-income individuals. |
| Credit History Length | 5.80 | 4.06 | 1.66 | 3.72 | Moderate right skew, long tail of borrowers with lengthy histories. |

Table 2: Key Observations in Descriptive Statistics

|  |  |
| --- | --- |
|  |  |
|  |  |

Figure 2: Probability Distribution of selected features

**Additional Insights:**

* **Outliers** are evident in several features, especially person\_age, person\_income, and person\_emp\_length, where maximum values are far from the 95th percentile.
* **Skewness & Kurtosis** indicate that person\_income in particular has extreme values, suggesting a potential need for transformation (e.g., log scale) before feeding into ML models.
* The typical borrower appears to be a **26–30-year-old with ~£55,000 income**, a **loan of ~£8,000**, and a **credit history of ~4–6 years**.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Feature | 25% | 50% (Median) | 75% | Interpretation |
| Age | 23 | 26 | 30 | Typical borrower is **26–30 years old** |
| Income | 38,500 | 55,000 | 79,200 | Median is **£55,000** |
| Loan Amount | 5,000 | 8,000 | 12,200 | Median loan is **£8,000** |
| Credit history length | 3 | 4 | 8 | Median is 4, so a range of **~4–6 years** is fair |

Table 3: Quartiles data to interpret a typical borrower specification

**5.3 Categorical Feature Analysis & Interpretation**

Exploring key categorical variables and their potential impact:

|  |  |
| --- | --- |
|  |  |
|  |  |

Figure 3: Plot Histograms for selected features

**1. Home Ownership**

* Categories: RENT, OWN, MORTGAGE, OTHER
* May reflect financial stability; OWN/MORTGAGE could signal creditworthiness.

**2. Loan Intent**

* Categories include: EDUCATION, MEDICAL, VENTURE, etc.
* Purpose may influence risk; e.g., DEBTCONSOLIDATION vs. EDUCATION.

**3. Loan Grade**

* Grades: A–G
* Likely reflects internal credit scoring — strong feature for modeling.

**4. Past Defaults on Loans**

* Binary: Yes/No
* Past defaults could be a strong risk indicator.

**5. Loan Status (Target Variable)**

* Binary: 0 = non-default, 1 = default
* Check for class imbalance; key for model evaluation metrics.

**5.4 Scatter Plots for Relationship Analysis**

To explore potential relationships between key numerical variables and identify patterns that may influence loan outcomes, a series of scatter plots were generated and analysed below.

|  |  |
| --- | --- |
|  |  |
|  |  |

Figure 4: Scatter Plots for selected features

**1. Person Income vs Person Age**

* Most individuals earn below 1e6, with ages ranging from 20s to 80s.
* A very weak positive correlation exists between age and income (R² = 0.030).
* Both loan statuses appear across all ages, with no clear age-related pattern in defaults.
* Implication: Age alone is not a strong predictor of income or loan default.

**2. Person Income vs Loan Amount**

* Majority of loans are for amounts below 25,000 and incomes below 0.5e6.
* Slight positive trend (R² = 0.071), but the data is highly scattered.
* Defaults (loan\_status = 1) occur across income levels, but high-income borrowers default less.
* Implication: Income only weakly predicts loan size; other factors likely influence lending decisions more.

**3. Loan Amount vs Interest Rate**

* Loans of similar amounts span a wide range of interest rates.
* Slight indication that higher loan amounts occur less frequently at higher interest rates.
* Defaults are slightly more concentrated at higher rates, suggesting a link to credit risk.
* Implication: No strong correlation, but higher rates may reflect risk-based pricing.

**4. Interest Rate vs Person Income**

* As income decreases, interest rate tends to increase—indicative of credit risk–based pricing.
* Defaults (loan\_status = 1) are more common at higher interest rates and lower incomes.
* High-income individuals mostly receive lower rates and rarely default.
* Implication: Stronger indication that interest rate, income, and default risk are interrelated.

**5.5 Box Plots for Comparing Distributions Across Categories**

To explore how key numerical variables differ across categorical segments, box plots were utilised to visualise the distribution, central tendency, variability, and presence of outliers within each category.

|  |  |
| --- | --- |
|  |  |
|  |  |

Figure 5:Box Plots for selected features

**1. Income grouped by Loan Intent**

* **Consistent Median Incomes:** Median income levels are relatively similar across loan purposes, suggesting intent does not strongly dictate income.
* **Widespread Outliers:** High-income outliers are present in all categories, especially in PERSONAL and VENTURE.
* **Notable Categories:**
  + DEBTCONSOLIDATION & EDUCATION: Slightly higher boxes suggest somewhat higher income borrowers.
  + VENTURE: Shows greater income variability within the interquartile range.

**2. Loan Amount Grouped by Home Ownership**

* **Higher Loans for Mortgage Holders:** MORTGAGE and OTHER groups show higher median loan amounts.
* **Lower Loans for OWN and RENT:** These categories show lower median and tighter distributions.
* **Outliers Across All Categories:** Regardless of home ownership, some individuals borrow significantly more than others.

**3. Interest Rate Grouped by Default History**

* **Higher Rates for Defaulters:** Individuals with past defaults ('Y') face noticeably higher and more varied interest rates.
* **Wider Spread and Outliers:** The 'Y' group has a larger spread and more high-end outliers, suggesting inconsistent risk-based pricing.

**4. Age by Loan Intent**

* **Consistent Median Ages:** Most loan intents show similar median ages, mainly late 20s to early 30s.
* **Young Adults Dominate:** Majority of borrowers are in their 20s–30s across all categories.
* **Older Borrowers Present:** Outliers in every category show borrowing continues into older age.
* **Unrealistic Outliers:** Extreme values above 120 suggest data entry errors needing cleanup.

**5.6** **Crosstab Insights: Linking Categorical Features to Default Risk**

**1. Home Ownership vs Loan Status**

* **RENT** shows the highest default rate (73%).
* **MORTGAGE** holders have the lowest (24%).
* **OWN** is a small category; **OTHER** appears unused.
* **Interpretation**: Renters may lack financial stability; mortgage holders likely more reliable.
* **Implication**: Home ownership is a key risk indicator for lenders.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Home Ownership | MORTGAGE | OTHER | OWN | RENT |
| Loan Status |  |  |  |  |
| 0 | 46% | 0% | 9% | 44% |
| 1 | 24% | 0% | 3% | 73% |
| All | 41% | 0% | 8% | 50% |

Table 4: Crosstab Analysis of Home Ownership vs Loan Status

**2. Loan Intent vs Loan Status**

* MEDICAL loans have the highest default rate (23%).
* VENTURE loans show the lowest (12%).
* EDUCATION loans are relatively safe (16% default).
* **Most common purposes:** EDUCATION, MEDICAL, VENTURE.
* **Interpretation**: Medical expenses may be unpredictable; venture loans might attract stronger applicants.
* **Implication**: Loan purpose should factor into credit risk models and pricing.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Loan Intent | DEBTCONSOLIDATION | EDUCATION | HOMEIMPROVEMENT | MEDICAL | PERSONAL | VENTURE |
| Loan Status |  |  |  |  |  |  |
| 0 | 15% | 21% | 10% | 17% | 17% | 19% |
| 1 | 21% | 16% | 13% | 23% | 15% | 12% |
| All | 16% | 20% | 11% | 19% | 17% | 18% |

Table 5: Crosstab Analysis of Loan Intent vs Loan Status

**3. Past Defaults vs Loan Status**

* **'Y'** (prior default): 31% default rate.
* **'N'** (no prior default): 14% default rate.
* 82% of borrowers have a clean credit history.
* **Interpretation**: Past behavior is a strong predictor of future risk.
* **Implication**: Default history is a critical feature in loan approval and risk scoring.

|  |  |  |
| --- | --- | --- |
| Past Defaults | N | Y |
| Loan Status |  |  |
| 0 | 86% | 14% |
| 1 | 69% | 31% |
| All | 82% | 18% |

Table 6: Crosstab Analysis of Past Defaults vs Loan Status

**4. Loan Status vs Past Defaults vs Home Ownership**

* Highest default risk: Renters with prior defaults (55%).
* Lowest default risk: Mortgage holders with no prior defaults (10%).
* Prior defaults ('Y') significantly increase risk across all home ownership categories.
* 'OTHER' home ownership category shows inconsistent data.
* Interpretation: Past defaults and renting combine to amplify risk; mortgage holders with clean records are most stable.
* Implication: Lenders should heavily weigh these factors in risk assessment and tailor loan terms accordingly.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Past Defaults** | **N** | | | | **Y** | | | |
| **Home Ownership** | **MORTGAGE** | **OTHER** | **OWN** | **RENT** | **MORTGAGE** | **OTHER** | **OWN** | **RENT** |
| **Loan Status** |  |  |  |  |  |  |  |  |
| 0 | 90% | 78% | 93% | 72% | 71% | 45% | 89% | 53% |
| 1 | 10% | 22% | 7% | 28% | 29% | 55% | 11% | 47% |

Table 7: Crosstab Analysis of Loan Status vs Past Defaults vs Home Ownership

**5.7 Summary & Next Steps**

Our exploratory analysis revealed meaningful patterns and relationships across key features, especially those linked to loan default risk, such as home ownership, loan intent, and past default history. These insights will guide our data preparation process. In the next section, we will perform data preprocessing, where we’ll address missing values, handle outliers, and apply appropriate data imputation strategies to ensure the dataset is clean and ready for modelling.

**6. Data Preprocessing (Part 1 – Cleaning and Validation)**

Before advancing to deeper analysis and model preparation, it is essential to ensure the dataset is clean, consistent, and free from distortions. This stage focuses on handling missing values, identifying and treating outliers, and correcting anomalies or potential data leakage issues. These steps form the foundation for reliable feature analysis and modelling by ensuring that the input data accurately reflects underlying patterns rather than noise or error.

**6.1** **Handling Missing Values**

To reduce dimensionality and prevent bias from excessive imputation, I began by removing features with more than 30% missing values using a threshold-based approach. This aligns with common practice, where variables exceeding 30–40% missingness are often considered unreliable unless they hold critical value. Eliminating such features helps avoid distortion from imputation and improves model generalisability and training efficiency (Little & Rubin, 2002; Hastie, Tibshirani, & Friedman, 2009).

After removing features with high missingness (threshold = 30%), no features met the removal criteria. I then addressed the remaining missing values using tailored, feature-level strategies. For numerical variables prone to outliers—such as income, age, and credit history length—I used median imputation to ensure robustness. Categorical variables like home ownership were imputed using the mode. In contrast, for critical variables like interest rate, loan amount, and the target variable, I dropped rows with missing values to preserve data integrity. In total, 3,116 records were removed, and 827 records were imputed. These strategies were applied modularly using a configurable pipeline that supports various imputation methods. This approach ensures transparency, minimizes bias, and maintains the reliability of subsequent analyses (Little and Rubin, 2002).

|  |  |
| --- | --- |
| Strategy | Description |
| drop | Removes rows where the feature value is missing or identified as an outlier. |
| fill\_mean | Replaces missing/outlier values with the mean of the feature. |
| fill\_median | Replaces missing/outlier values with the median of the feature. |
| fill\_mode | Replaces missing/outlier values with the most frequent (mode) value. |
| fill\_value | Replaces with a custom user-defined value. Requires fill\_value input. |
| ffill | Applies forward fill: uses the previous valid value. |
| bfill | Applies backward fill: uses the next valid value. |
| fill\_interpolate | Uses linear interpolation between valid surrounding values. |
| keep / none | Takes no action—the missing/outlier value is retained as-is. |

Table 8: Imputation Strategies for Handling Missing Values and Outliers

**6.2 Outlier Detection and Handling**

Before delving into outlier treatment at the feature level, a *row-wise threshold for removal* was defined to ensure the overall data integrity per record. Similar to handling missing values, where rows with excessive missingness can distort model training or reduce generalizability, records with a high proportion of outlier features are often too anomalous or corrupted to retain. For this reason, I applied a configurable threshold (set at 50%)—meaning if more than half of a row’s numerical features were identified as outliers, the record would be excluded entirely. This approach balances data quality and retention, removing only those observations likely to introduce noise or instability into the model (Osborne and Overbay, 2004).

After confirming that no rows exceeded this threshold:

* **0 records were removed** due to excessive outlier features.

Following this check, *feature-level outlier detection* was performed using various statistical and machine learning methods, including:

|  |  |
| --- | --- |
| Method | Description |
| Interquartile Range (IQR) | Detects values that lie outside 1.5× IQR from Q1 or Q3. |
| Z-score Filtering | Flags values with standard scores (z-scores) beyond a threshold (e.g., ±3). |
| Isolation Forest | Anomaly detection using tree structures to isolate rare points. |
| Local Outlier Factor (LOF) | Compares local density to detect points that are sparser than neighbours. |
| Distribution Fitting | Fits statistical distributions (e.g., lognormal) to identify poor-fitting values. |
| Custom Bounds | Applies domain-specific upper and/or lower limits defined by the user. |

Table 9: Summary of statistical and machine learning techniques used to detect outliers

Each feature was examined independently, starting with descriptive statistics to flag skewed or unrealistic distributions. Two variables in particular—*Age* and *Employment Length*—showed characteristics of data entry or logical errors. For instance:

* Age ranged from **20 to 144**, with high skewness (**2.58**) and extreme kurtosis (**18.56**).
* Employment Length went up to **123**, with kurtosis exceeding **43**—a clear indicator of extreme outliers.

Using custom bounds:

* Age was restricted to the range **18–80**, with outliers replaced by the **median**.
* Employment Length values exceeding **50** were replaced with **50**, a sensible domain-specific cap.

A unified pipeline was executed to apply these strategies based on configuration settings. Features such as Income, Loan Amount, and Loan Interest Rate were inspected but retained as-is, given their business relevance and plausible distributions.

**Results:**

* Total custom-based outliers found: **7**
* **0** records removed due to row-wise outlier thresholds
* **7** records handled through replacement strategies

This approach ensured rigorous cleaning while preserving the informative content necessary for downstream modelling.

**6.3 Summary & Next Steps**

In this section, we cleaned and validated the dataset by handling missing values through targeted imputation strategies and detecting/removing outliers using both statistical and machine learning-based methods. These steps enhanced data reliability and ensured consistency across features. With a clean and trustworthy dataset now in place, the next step is to perform a full exploratory data analysis (EDA). This will involve examining feature correlations, checking for multicollinearity and interactions, applying dimensionality reduction (e.g., PCA), and exploring relationships with the target variable (Default) to inform robust feature engineering and model design.

**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].

Hastie, T., Tibshirani, R. and Friedman, J., 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. New York: Springer.

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.

Little, R.J.A. and Rubin, D.B., 2002. *Statistical Analysis with Missing Data*. 2nd ed. Hoboken, NJ: Wiley.

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].

Osborne, J.W. and Overbay, A., 2004. The power of outliers (and why researchers should always check for them). *Practical Assessment, Research & Evaluation*, 9(6), pp.1–12.

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].